Word Sense Selection in Texts: An Integrated Model

Oi Yee Kwong

Downing College
University of Cambridge

A Dissertation Submitted for the Degree of
Doctor of Philosophy

May 2000

To my parents

Preface

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration.

This dissertation is not substantially the same as any that I have submitted for a degree or diploma or other qualification at any other University. Further, no part of my dissertation has already been or is being concurrently submitted for any such degree, diploma or other qualification.

This dissertation does not exceed 60,000 words, including tables, footnotes and bibliography, but excluding appendices and diagrams.

Oi Yee Kwong
Downing College
University of Cambridge
May 2000

Abstract

Early systems for word sense disambiguation (WSD) often depended on individual tailor-made lexical resources, hand-coded with as much lexical information as needed, but of severely limited vocabulary size. Recent studies tend to extract lexical information from a variety of existing resources (e.g. machine-readable dictionaries, corpora) for broad coverage. However, this raises the issue of how to combine the information from different resources.

Thus while different types of resource could make different contribution to WSD, studies to date have not shown what contribution they make, how they should be combined, and whether they are equally relevant to all words to be disambiguated. This thesis proposes an *Integrated Model* as a framework to study the inter-relatedness of three major parameters in WSD: *Lexical Resource*, *Contextual Information*, and *Nature of Target Words*. We argue that it is their interaction which shapes the effectiveness of any WSD system.

A generalised, structurally-based sense-mapping algorithm was designed to combine various types of lexical resource. This enables information from these resources to be used simultaneously and compatibly, while respecting their distinctive structures. In studying the effect of context on WSD, different semantic relations available from the combined resources were used, and a recursive filtering algorithm was designed to overcome combinatorial explosion. We then investigated, from two directions, how the target words themselves could affect the usefulness of different types of knowledge. In particular, we modelled WSD with the cloze test format, i.e. as texts with blanks and all senses for one specific word as alternative choices for filling the blank.

A full-scale combination of WordNet and Roget's Thesaurus was done, linking more than 30,000 senses. Using these two resources in combination, a range of disambiguation tests was done on more than 60,000 noun instances from corpus texts of different types, and 60 blanks from real cloze texts. Results show that combining resources is useful for enriching lexical information, and hence making WSD more effective though not completely. Also, different target words make different demand on contextual information, and this interaction is closely related to text types. Future work is suggested for expanding the analysis on target nature and making the combination of disambiguation evidence sensitive to the requirements of the word being disambiguated.

Acknowledgements

I wish to thank my supervisor, Prof. Karen Sparck Jones, for her guidance, advice and inspiration throughout this project, as well as for the insightful and helpful comments on the earlier drafts of this thesis.

I am grateful to Dr. Ted Briscoe and Dr. Stephen Pulman for useful pointers and the many opportunities for broadening my knowledge of the field. I would also like to say thank you to all my colleagues, past and present, in the Natural Language Processing group of the Cambridge University Computer Laboratory, for their peer support and for sharing the ups and downs of my research. My appreciation extends to all staff members in the Computer Laboratory and Downing College who have helped in various ways during my study; and also to all my friends for their warm friendship which has brightened the days in Cambridge.

My greatest gratitude to my parents for their unconditional love, support and encouragement all along can never be adequately put into words.

This work was financially supported at various stages by the Committee of Vice-Chancellors and Principals of the Universities of the United Kingdom (Overseas Research Students Award), the Cambridge Commonwealth Trust (Scholarship), Downing College (Downing-Müller Cambridge Scholarship), and the Croucher Foundation (Scholarship). Travel to conferences was generously funded by the Cambridge University Computer Laboratory, Downing College, and the Croucher Foundation.

Contents

Acronyms and Abbreviations

CACM	Communications of the ACM
CC	Conventionalised Context
CIDE	Cambridge International Dictionary of English
COBUILD	Collins COBUILD English Dictionary
COLLINS	Collins English Dictionary
IF	Information Filter
ILR	Integrated Lexical Resource
IR	Information Retrieval
ITA	Inter-Tagger Agreement
LDOCE	Longman Dictionary of Contemporary English
LLOCE	Longman Lexicon of Contemporary English
LOB	Lancaster Oslo-Bergen (Corpus)
M-C	Multiple-Choice
MOR	Measurement of Relatedness
MRD	Machine-Readable Dictionary
MT	Machine Translation
NLP	Natural Language Processing
OALDCE	Oxford Advanced Learner's Dictionary of Current English
POS	Part of Speech
RIT	Roget's International Thesaurus
ROG11	1911 Crowell Roget's Thesaurus
ROG87	1987 Penguin Roget's Thesaurus
ROGET	Roget's Thesaurus
SEMCOR	Semantic Concordances (from Brown Corpus)
SP	Selectional Preference
SR	Selectional Restriction
STC	Sense Type Classifier
TC	Triggering Context
VLNN	Very Large Neural Network
W7	Webster's Seventh Collegiate Dictionary
WN	WordNet
WN15	WordNet 1.5
WN16	WordNet 1.6
WSD	Word Sense Disambiguation
WSJ	Wall Street Journal

1

Introduction

"We have different kinds of word ... But how we group words into kinds will depend on the aim of the classification, – and on our own inclination." (Wittgenstein, 1958: §17)

1.1 Objectives

Word sense disambiguation (WSD) is the process of identifying word meanings in a discourse. Although word meanings are the building blocks for the meaning of discourse, they often have to be understood by way of the discourse as a whole. The automation of this process thus largely depends on the ability of a system to uncover various semantic relations among words in a text.

The conviction that some sort of WSD is needed for nearly every natural language processing (NLP) application might have encouraged WSD research all along. In the last few decades, progress has permeated nearly every aspect of the subject. For instance, more statistical methods have been used in addition to symbolic ones; machine-readable dictionaries and large corpora have almost replaced individual hand-coded, tailor-made lexical resources; disambiguation information has extended from mainly pragmatic and semantic types to the inclusion of more syntactic types, and so on. We will review all such development thoroughly in Chapter 2.

The fact to date is, however, that WSD is still not *all* solved. Although encouraging results are reported in many studies, we believe that what can be achieved for WSD for the present is limited by the following problems. First, well-defined definitions of sense are lacking. Second, the parameters involved in WSD and their interaction are not thoroughly understood. Third, the behaviour of the words themselves is overlooked.

The objective of this thesis is therefore to take a step back to conduct a more systematic study of some determining factors in WSD. Our major hypothesis is that the success or failure of WSD is dependent on many *interacting* parameters. Some of them have been studied individually in the past, but some have not. We believe that only by finding out their systematic relationship can we gain a substantial breakthrough and produce more genuine and reliable disambiguation results. For this purpose, we propose an *Integrated Model* as a framework in this study. The model will be described in Chapter 3.

In this introductory chapter, we will first briefly discuss the notion of word sense in Section 1.2 and the various models for the lexicalisation of word meaning in Section 1.3. We will then explore in Section 1.4 the validity of using some given set of institutionalised word senses as a starting point in WSD, a usual practice which we will also follow in this work. An example of parameter interaction in WSD will be shown in Section 1.5. We will study three major parameters, and our approach will be explained in Section 1.6. An overview of the organisation of this thesis will be given in Section 1.7.

A remark on terminology before we go on: in our discussion, we will assume the straightforward use of *word* as referring to any *headword* as found in dictionaries, regardless of the various forms in which they can appear in texts. Also, WSD in our discussion mostly refers to automatic WSD, and *sense* and *meaning* will be used interchangeably.

1.2 The Notion of Sense

It sounds unimaginable for us to talk about WSD without being able to say what exactly a word sense is, but that is not far from the truth. Sense is an abstraction. We cannot possibly point to something and say that is the sense of a word. We can only grasp the sense of a word via our understanding of the message in which the word is used. According to Hurford and Heasley (1983), "the sense of an expression is its place in a system of semantic relationships with other expressions in the language". Thus sense is a linguistic concept, and it only covers the linguistic meaning, in contrast to other types such as speaker meaning. For example, while we are walking together, my friend complains about someone's bad conduct. I see a hound coming towards us and I say to my friend: "Dog!" He might ponder in his mind:

1. Am I referring to a 4-legged animal or a morally reprehensible person?

2. Am I asking him to protect me or am I on his side about that person's behaviour?

As far as the *sense* of "dog" is concerned, only the first question is relevant. To disambiguate word senses is therefore to find out the *use*, among all possible *uses*, that a word adopts at a particular occasion. Although the distinction is a linguistic one, the process does not preclude the use of extralinguistic information. For instance, my friend might need to know if I am usually afraid of dogs to get my message.

Ambiguity arises when the same word can potentially be used to mean different things. We will not trace the historical reason why language evolves this way, but will simply accept the presence of ambiguity as a linguistic fact. Two levels of lexical ambiguity are of particular interest, namely *homonymy* and *polysemy*. Homonymy refers to a word having meanings which lack obvious semantic relations among themselves, normally because they have distinct etymological sources (e.g. "pen" as "a writing instrument" or "an enclosure for livestock"). Polysemy, on the contrary, refers to a word having different but closely related senses (e.g. "pen" as "an enclosure for livestock" or "an enclosure in which babies play"). It can be systematic or idiosyncratic, and sense shifts are not uncommon (Clark and Clark, 1979; Copestake and Briscoe, 1995).[1] Polysemy is thus rendered more problematic for WSD.

[1]The problem is coupled by what is known as the *vagueness* of sense, as opposed to genuine polysemy, see e.g. Lytinen (1988).

In this study, we deal with homonymy and polysemy with respect to one syntactic category only, namely nouns. We believe that restricting ourselves to noun ambiguities will make a better focus. Although we realise that ambiguities for other parts of speech (POSs) and ambiguities across different POSs (also known as categorial ambiguities) can raise unique processing issues, we are not concerned with them here.

The reality of word senses is supported not only psychologically (Jorgensen, 1990), but also practically, for we always need some representation of lexical ambiguity to model the handling of "ambiguous words" computationally (Wilks et al., 1989). In the next section, we will look at some models for representing word senses in terms of other words in the language.

1.3 Lexicalisation of Word Meaning

Despite the abstractness of word senses, it is possible to encode the meaning of a word with other words in the language. Such *lexicalisation* of senses may be imprecise, but it is useful for the users, particularly new learners, of the language to "make sense" of a word when it is heard or used. In the following, we summarise three main models for defining and representing word senses, namely the *descriptive model*, the *relational model*, and the *generative model*. We will compare how they view the stability of senses and how explicitly they define senses. By *explicitness*, we refer to the preciseness of the relationship between the word being defined and the word(s) used to define it. This should not be confused with *concreteness*, which we refer to as the perceptibility of any image that a definition carries.

1.3.1 Descriptive Model

The descriptive model is the most intuitive and common model for defining senses. It is seen in most conventional monolingual dictionaries. Senses are usually defined via some necessary and sufficient conditions, similar to constructing a prototype of a concept. For example, in the Longman Dictionary of Contemporary English (LDOCE), we find a sense of "dog" as follows:

> *a common 4-legged flesh-eating animal, esp. any of the many varieties used by man as a companion or for hunting, working, guarding, etc. It is often called "man's best friend".*

Compare this with a sense of "cat":

> *a small animal with soft fur and sharp teeth and claws (nails), often kept as a pet or in buildings to catch mice and rats.*

It can be seen that definitions like these can be prone to inconsistency and inadequacy. For instance, why does a "dog" have to be 4-legged (normally), but a "cat" does not? Nonetheless, there is rich information in dictionary definitions, often more than just linguistic information. Although the relations between the headword and the words in its definition are implicit, the patterns of definitions are often exploited to make the relations explicit for computational use (e.g. Amsler, 1981). More discussion on the transformation of dictionary definitions to taxonomies will be found in Chapter 2.

The style of definitions can vary from dictionary to dictionary, depending on their individual editorial philosophy as well as their target audience. For example, LDOCE restricts the words in its definitions to a defining vocabulary of about 2,000 words and uses a rather regular syntax for the definitions (Procter, 1978). Following a different philosophy, the Collins COBUILD English Dictionary (COBUILD) defines words by illustrating their uses via corpus linguistics. It uses authentic examples from naturally occurring English texts (in contrast to artificial examples as used in some other dictionaries) to "explain" word uses, and such "explanations" often include pragmatic meanings like the positive/negative connotation in the use of a word (Sinclair, 1987). It is also well known that dictionaries differ from one another in terms of sense granularity. For instance, although the Cambridge International Dictionary of English (CIDE) resembles LDOCE in almost every way, it is claimed to have more fine-grained senses than LDOCE (Harley and Glennon, 1997).

The descriptive model is biased towards a static view of senses, which suggests that senses are discrete and enumerable. This appears reasonable on the assumption that each word has a set of core senses and they are listed in the dictionary. However, the model is often criticised for failing to cover word senses which are not core, and reducing word senses to decontextualised definitions. Despite this, the descriptive model remains the most common model for sense representation to date.

1.3.2 Relational Model

Relational models define word senses directly in terms of their semantic relations with other words in the language. Semantic relations are synonymy, antonymy, hyponymy, and so on. We describe two specific models of this kind below (see e.g. Evens (1988) for a fuller account of relational models).

Thesaurus Model with Synonymy (Sparck Jones, 1986)

To make thesaurus construction more systematic, Sparck Jones (1986) argues that senses can only be most appropriately defined in terms of synonymy, among other semantic relations. She defines a sense in the form of a "row", which consists of words substitutable for one another at a particular position in a sentence without changing the ploy of the sentence. A row may contain only one word, but more often several, and will usually be justified by many context sentences. Rows can be combined into more general classes, forming a semantic structure for the words in the language. This constitutes a bottom-up model of constructing thesaural classes. With such classes, words in a sentence might then be able to mutually disambiguate one another through some "group-intersection" procedure.

To get a reasonable number of rows, one would need a huge amount of sample sentences to test for substitutability. Lacking large corpora, Sparck Jones bootstrapped rows from definitions with synonyms in a dictionary and tested substitutability against the usage illustrations in the dictionary. The classification of words in WordNet (Miller et al., 1990) is also grounded on synonymy. Similarly, Plate constructed thesaural classes with co-occurrence statistics obtained from LDOCE definitions (Wilks et al., 1989). This makes the process less laborious, but the resulting clusters of words are not necessarily substantiated solely by synonymy. They are more like what one finds in Roget's Thesaurus (Roget, 1852; Dutch, 1975; Kirkpatrick, 1987).

Word Space Model (Schütze, 1992)

Schütze's (1992; 1997; 1998) Word Space Model defines senses as context groups, derived from corpus co-occurrence data and represented in a multi-dimensional vector space. All words co-occurring with a word w (as a dimension in the vector space, which usually suggests some topic or subject matter) in a fixed-sized context have an entry in the word vector of w. It indicates the number of times the word co-occurs with w in the corpus. These entries together form the "Word Space". With this representation, the semantic relatedness between two words can be crudely measured by a cosine function. A context vector is given by the centroid of the word vectors occurring in the context. The component of a context vector with respect to a certain dimension thus reflects the topical strength of the context for that dimension. The context vectors of an ambiguous word are clustered into a predetermined number of groups, and the senses of the word are represented by the centroids of these context groups. WSD is subsequently reduced to the discrimination among context groups, that is to find the sense vector of the word which is closest to the context vector of its new occurrence. In Schütze (1998), data were gathered from a 17-month record of the New York Time News Service (about 60.5M words).

Kilgarriff (1997a) proposed a corpus-based sense clustering method for defining senses (as opposed to dictionary definitions) in the same spirit. Not only does the method require a lot of corpus data, but it also exploits the bias in the data to get the most representative set of word senses for the data. It is therefore possible that some intuitively core senses are excluded.

Relational models treat word senses in a more dynamic way. The actual data determine the set of senses. Sense boundaries are soft. Although definitions are generally more free from extralinguistic information, relational models do differ in their explicitness. For instance, Sparck Jones' model is very explicit because rows are always governed by synonymy, and higher level classes are derived from them. Other clustering methods may be less explicit since we do not know what relations tie up a cluster. It is only obvious that words in the same cluster tend to be found in the context for a sense.

1.3.3 Generative Model

Pustejovsky (1991) proposed an even more dynamic model for lexicalising word senses, where the encoding is actually left open-ended. In his generative model, compositionality is assumed, and lexical entries contain a range of representative aspects of lexical meaning (i.e. features) at different levels. They are "argument structure", "event structure", "qualia structure", and "lexical inheritance structure" (see Pustejovsky (1991) for details). These levels of representation can be connected via type coercion. Ambiguity resolution therefore becomes a process of highlighting mutually compatible and relevant lexical components of words and phrases. This is achieved by "co-composition", i.e. performing specialised inference in predefined ways which control the composition of knowledge structures of words in context. Consider the example "a fast car". Given the following qualia structure for "car":

$$\begin{bmatrix} \mathbf{car(x)} \\ \text{CONST} = \{\mathbf{body,engine,...}\} \\ \text{FORMAL} = \mathbf{physobj(x)} \\ \text{TELIC} = \mathbf{drive(P,y,x)} \\ \text{AGENTIVE} = \mathbf{artifact(x)} \end{bmatrix}$$

the adjective "fast" can be resolved as modifying the event "drive", which means the process but not the car itself is fast. Similarly, given the telic role of "motorway":

$$[Telic : travel(P, cars) \wedge on(P, x)]$$

"fast" in "a fast motorway" should describe "the process of travelling on a motorway", rather than the motorway itself (Pustejovsky and Boguraev, 1993). Therefore, by predefining the general behaviour of a word like "fast" as always predicating of an event type (which matches the telic role of a nominal), subtly different uses of the word can be encoded. It is otherwise impossible to list such slight sense variations exhaustively.

Relations between words are explicitly encoded in the generative model, although the relations are more logical than linguistic. In addition, the model has been shown to be able to represent systematically some regular polysemy and sense extensions[2] (Copestake and Briscoe, 1995). Despite its advantages, the model nevertheless makes a constrained form of dictionary, which is tricky to be grounded properly in the first place. This difficulty thus limits the utility and popularity of the model in WSD.

To sum up from this whole section, although word senses are abstract and ever evolving, some core ones can usually be represented in some way. The descriptive model and relational model are more common, mostly because corresponding lexical resources, i.e. dictionaries and thesauri especially, are readily available. The senses listed in these resources may be inadequate, but they can give some leverage to WSD studies. Using the predetermined senses in existing resources as reference has therefore become the mainstream in WSD. In the next section, we will discuss the validity and limitation of this practice.

1.4 Two Basic Assumptions

As said, most WSD studies nowadays start with the predetermined senses in existing lexical resources as a reference for sense assignment. The plausibility of this practice rests on two fundamental assumptions:

[2] e.g. logical metonymy (a fast car, a fast typist, a fast motorway, etc.), broadening (a cloud of something), nominal metonymy (e.g. the ham sandwich), portioning (three beers = three portions of beer), grinding (e.g. lamb as the animal or the meat), and so on.

A1 *Senses exist and can always be represented in some way.*

A2 *Senses are discrete, and in principle they can be unambiguously assigned to novel occurrences of words in a discourse.*

We will take A1 for granted since the reality of word senses is supported in the literature, and we have seen above how senses can be represented in various ways. More disputable is A2. This assumption carries two questions: whether senses will ever be the same so that a sense previously seen can be used to name a later one, and how definitively senses are bounded so that we can say it is one but not the other.

Since a "sense" of a word means a "use" of it, it follows that no two senses will be the same because no two occasions (including linguistic and extralinguistic contexts) where the word is used will be identical in the strictest sense. But people do not use words randomly or senselessly. They use a word because they believe the word will convey a meaning which has been conveyed by it before (except when people deliberately use words innovatively). Thus, although there are two occasions of use, we can make a compromise by saying the two occurrences carry the same sense if they appear in contexts which are similar enough and it is likely that they convey the same message.

We can think of the situation symbolically as in Figure 1.1. A lexical resource only lists a fixed number of senses (e.g. S1, S2), but there is a range of contexts (the span covered by each curve) within which we will perceive a novel use as the listed sense. For instance, one may not care whether to have chicken nuggets with barbecue sauce or with honey mustard sauce, as long as one is not served pork chop when one asks for chicken. Therefore, we reckon that the discreteness of sense assumed in A2 should more appropriately be of a perceptual kind, over a continuum of contexts and uses. This has no contradiction to other viewpoints on the discreteness of sense, such as Lyons' (1981):

> "Discreteness in language is a property of form, not meaning." (p.148)

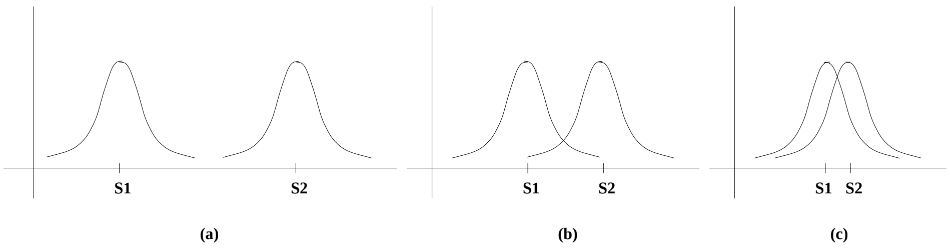


Figure 1.1 *Perceptual Discreteness and Boundaries of Word Senses*

As far as sense discreteness is concerned, the classic "Bank Model" must be mentioned. It suggests that most words are like the word "bank" which has perfectly distinct senses, and predicts that each occurrence of "bank" would refer to one sense (a money bank) or the other (a river bank). More importantly, such sense recognition should be instant and effortless for English speakers.

However, Kilgarriff (1992) argued that "the Bank Model is fatally flawed". He started by analysing a sample of words and their definitions in LDOCE for the type of sense distinction (e.g. if one sense is a generalising metaphor of another, or a domain shift, etc.). It turned out that most sense distinctions could not fit nicely into any of his four sense distinction types. He then tried to match each occurrence of his test words in the Lancaster Oslo-Bergen (LOB) corpus, with ± 7 words as context, to one of their senses listed in LDOCE. He found that 87% (60 out of 69) of the test words had at least one usage which could not be confidently assigned only one sense. He therefore concluded that word senses were by no means discrete, and the Bank Model should not hold.

All the same, we find that Kilgarriff's results have not served their purpose for arguing against the use of predetermined senses in existing lexical resources for WSD. On the one hand, Wilks (1998) pointed out that the 87%-claim was misleading. There was absolutely no contradiction to saying that the majority of the text usage could be associated with only one dictionary sense (note the contrast of "word" and "usage"), as supported by the promising results in studies like Cowie et al. (1992). On the other hand, the Bank Model would not be completely invalidated by Kilgarriff's results. If we have accepted the point on perceptual discreteness, then whether one can assign confidently only one sense to a novel use would correspond exactly to the second question posed by A2, that is how distinct the senses are.

We can thus say that the Bank Model in inadequate (but not fatally flawed) since it only covers one extreme of sense distinction, that is two senses with little semantic relation or homonymy, as represented by Figure 1.1(a). For the rest of the continuum, there are different degrees of polysemy where senses are more closely related, as in Figures 1.1(b) and (c). Therefore senses can in principle be unambiguously assigned, but the chance depends on how closely related the senses are. For example, one may find it easier to distinguish "colour" from "shape" than to distinguish "breadth" from "height". The distinctness between senses, of course, partly depends on the sense inventory of individual lexical resources.

Thus according to our interpretation of the discreteness assumption, the predetermined senses in lexical resources function as a tool for construing senses. This is just like using a ruler as an instrument for measuring length. One reads off a discrete number (e.g. 1.3 cm) from the ruler, but length itself is still a continuous function. The sense inventories we have may be incomplete and the representation may be inexact, but as suggested by Wittgenstein (1958), "inexact" does not mean "unusable".

Consequently, the real limitation of grounding WSD on a set of predetermined senses does not lie in the discreteness arguments, but in the adequacy of the set of senses and the associated lexical information available. Since we often obtain the senses from some lexical resource, the conditions for adequacy thus comprise: (i) the number and the distinctness of senses listed in the resource, and (ii) the usefulness of the lexical information in it for matching an occurrence of a word to one (or more) of its senses listed.

The number of senses listed in a lexical resource affects the predictability of the sense for a novel occurrence of a word. Although how many senses a resource contains may depend on its size, the target readers, and so on, there should be some standard which one can expect. As Cruse (1986) suggested:

> "The number of fully established senses is presumably finite at any one time (though it may differ for different members of the language community, and at different times for the same speaker)." (p.70)

The distinctness between senses, as mentioned, determines how likely senses can be unambiguously assigned to novel occurrences. However, the best level of sense granularity seems to be application-dependent.

The usefulness of the lexical information for assisting sense assignment is a less straightforward issue. According to Wittgenstein (1958), for example, words of different kinds are explained and understood by different mechanisms. The resources used in WSD should therefore characterise senses via various types of lexical information, so that different types of evidence are available for hypothesising a sense under a variety of circumstances. Although a single resource only has limited types of knowledge (even if it has the perfect sense set), it is possible to complement resources with one another by combining them.

In fact, the information adequacy may vary not only from resource to resource in general, but also from word to word individually. However, this latter variation is often overlooked, as we pointed out in the beginning of this chapter. The factors which determine the effectiveness of a WSD system are intertwined. The interaction takes its form as we start with a set of predetermined senses, represented in some type of lexical information. First, words have different numbers of senses of varied granularity, and how these are reflected depends on the lexical resource from which we get the set of senses. Second, different resources may use a different model to represent senses. Various types of lexical information for the words, which are potentially useful in WSD, may thus spread across resources. Third, senses may be best characterised by different types of lexical information, so the most effective disambiguation evidence could be quite dynamic because a different type of information may serve the purpose best in each ambiguous case. Hence, it is important to understand this interaction thoroughly in order to improve WSD systems. Yet this is what we are lacking – another problem we mentioned in the beginning. We will give an illustration of such parameter interaction in WSD in the next section, and explain how we intend to investigate the problem in the one that follows.

Before starting the next section, we need to make a note on the scope of our work. First, we are not going to handle peculiar uses of words as in metaphor and metonymy. They deserve special treatment of their own (see e.g. Fass (1991)). Second, we will study only part of the linguistic context for which there is corresponding lexical information in the resources we use. So we will not talk about the role of discourse focus, inference, world knowledge, and so on, although such information might also be useful for WSD.

1.5 Parameter Interaction in Word Sense Disambiguation

In this section, we will introduce three important parameters in WSD which we are going to study, and illustrate their interaction with an example. We use Covington's (1994) textbook WSD algorithm to show three major parameters involved in disambiguating the word "pen" in the sentence "The pen is full of pigs":

- **Lexical Resource**

  So, first of all, we have a lexical resource which provides the predetermined senses of "pen", including "a writing instrument" and "an enclosure for livestock", labelled as *pen_for_animals* and *pen_for_writing* respectively. In Covington's particular resource, senses are characterised in the form of a group of cue words. For example, the cue words for *pen_for_animals* may include "farm", "pig", and so on; and those for *pen_for_writing* may include "ink", "paper", and so forth. We call such knowledge about the senses the *Conventionalised Context* (CC) of the senses.

- **Contextual Information**

  Should anything provide the hint for the sense of an ambiguous word, it must be the actual context embedding the occurrence of the word. We call this the *Triggering Context* (TC). Many types of contextual information can be extracted from the TC, but not all are useful for disambiguation. The usefulness has to do partly with the form of the CC in the lexical resource, and partly with the word to be disambiguated, i.e. the *Target Word*. Hence WSD is the search for the best match between TC and CC from all the candidate senses. Since the CC in this case is in the form of cue words, we try to find matching cue words in the TC. Consequently, the match from "pigs" renders the sense *pen_for_animals* more likely.

- **Nature of Target Words**

  It happens that a word like "pen", with the two senses considered, can be effectively distinguished via the presence of particular cue words. However, not all words have strongly associated cue words, and such words may require other types of information for effective disambiguation. Therefore, it is obvious that any one type of information will not work for all, and we have to study the behaviour of the target words so as to make effective use of various types of information accordingly.

Distinguished by their respective origins, we refer to *Lexical Resource* and *(Triggering) Contextual Information* as *extrinsic* parameters, and *Nature of Target Words* as an *intrinsic* parameter.

1.6 Our Approach: An Integrated Model

We mentioned three obstacles in WSD research in Section 1.1. It seems that the one regarding sense definitions is unlikely to be removed easily. Owing to the peculiar nature of word senses, no particular model can fully represent them. However, we have argued for the validity of grounding WSD on a set of institutionalised senses as a realistic solution. The limitation for WSD then depends mainly on the adequacy of the set of senses we use (in terms of number and granularity) and the usefulness of the lexical information characterising the senses. We have suggested that such limitation is a result of the interaction among three major parameters. Two extrinsic parameters (*Lexical Resource* and *Contextual Information*) and one intrinsic parameter (*Nature of Target Words*) together shape the effectiveness of any WSD system. They account for the other two obstacles regarding parameter interaction and word behaviour, and the investigation of these problems constitutes the core of our work.

Past studies, however, mostly worked on the two extrinsic parameters, often in isolation. For instance, there has been much work on the acquisition of lexical knowledge from various sources, to provide useful information for WSD, but the resultant resource is seldom used in actual WSD. Also, research on algorithms for matching TC to CC with respect to different types of information is abundant. On the contrary, the role of target words (especially their individual requirements on WSD) has only rarely and also inadequately been addressed. The interaction among these determining factors in WSD is far from being well understood.

We therefore propose an *Integrated Model* as a framework so that these parameters can be studied from a *holistic* perspective. The model will be described in Chapter 3. In particular, we will innovatively study the nature of target words in the form of *Cloze Tests*. We believe this method has modelling and processing advantages, and allows us better control in the disambiguation experiments. Although we will separately focus on each parameter in the upcoming chapters, the discussion should not be taken as unconnected units. Each parameter would be studied in a controlled environment so that a systematic account of the interwoven relationship could be given.

1.7 Overview of Subsequent Chapters

The remaining chapters in this thesis are organised as follows:

In **Chapter 2** *(Literature Review)*, we will give an account of the past studies in WSD. To sketch a more systematic picture of the state of the art, the studies will be organised along four axes: *Resource*, *Algorithm*, *Application*, and *Evaluation*. We will also relate our work closely to these axes.

In **Chapter 3** *(An Integrated Model of WSD)*, we will present the Integrated Model which we propose as a framework for studying the interaction among the parameters in WSD. The materials and methodology will be described.

In **Chapter 4** *(Extrinsic Parameter I: Lexical Resource)*, a structurally-based sense-mapping algorithm is developed for linking up various kinds of lexical resource, and a full-scale linkage of two existing resources is performed, forming an integrated resource. Lexical knowledge is enriched, and more importantly, the different types of information from various resources can be used simultaneously and compatibly.

In **Chapter 5** *(Extrinsic Parameter II: Contextual Information)*, we demonstrate the usefulness of our integrated resource by applying various types of semantic relation obtainable from it in disambiguation. Also, a recursive filtering algorithm is designed to enable full-text context to be tested while avoiding combinatorial explosion.

In **Chapter 6** *(Intrinsic Parameter: Nature of Target Words)*, we investigate the role of target words from different directions. Although it is difficult to come up with a definite classification of words with respect to their responses to different types of disambiguation information, the varied effectiveness of disambiguation information on different test instances is more than evident.

In **Chapter 7** *(Conclusion)*, we close the loop by summarising the achievements of our work and consolidating the Integrated Model with our experimental results. We will also suggest directions for future work.

2

Literature Review

A most recent and comprehensive survey on word sense disambiguation (WSD) is that by Ide and Veronis (1998), who discuss the major approaches from mainly a historical perspective and suggest future research directions. In this chapter, we will review the WSD-related studies from a different perspective, namely a motivational one. We set up four axes, described in Section 2.1, as a skeleton to organise our discussion. Each axis will be further explored in the subsequent sections (2.2 to 2.5). Finally in Section 2.6, we will explain how our work relates to the state of the art, and its orientation with respect to the four axes.

2.1 Four Axes of WSD Studies

The number of studies bearing on WSD[1], directly or indirectly, is enormous. In addition, they differ in various respects. Consequently, it is difficult to understand them systematically without concentrating on particular dimensions. Here we suggest four axes for this purpose based on the motivation of the studies:

- **Resource Axis**

  Semantic lexical resources, where senses are defined and their conventionalised contexts characterised by different types of lexical information, are clearly the prime requirement of WSD. Three ways of forming comprehensive resources for WSD will be discussed: (1) from scratch, (2) semi-automatically by acquiring information from machine-readable dictionaries (MRDs) or large corpora, and (3) combining multiple resources.

- **Algorithm Axis**

  Many of the relevant studies are devoted to the design of algorithms for associating the triggering context with the conventionalised context to get the best matched sense. We will survey the different algorithms with respect to the type of contextual information on which they depend.

[1] The subject is also known by other names, e.g. word sense discrimination (McRoy, 1992; Schütze, 1998), lexical ambiguity resolution (Hirst, 1988), automatic sense disambiguation (Lesk, 1986), sense tagging (Wilks and Stevenson, 1997), etc.

- **Application Axis**

  Since WSD is an intermediate task within larger scale natural language processing[2] (NLP), although recently it has been studied as an end in itself, some WSD studies are conducted in the context of certain applications to give a test bed for the effectiveness and usefulness of this subtask.

- **Evaluation Axis**

  The experimental conditions of individual WSD studies used to differ so much that comparison between studies is often difficult and fruitless. Efforts have been made recently to bring the whole enterprise to a common ground so that results from individual studies can be more objectively and meaningfully evaluated in an application-independent setting.

The present survey will be organised along these four axes, which are not totally orthogonal. They interact, as we will see shortly. An overview will be given in this chapter, and the particularly relevant aspects of individual studies will be discussed more thoroughly in the corresponding chapters to follow.

2.2 The Resource Axis

A WSD system needs to look up a lexical database containing the information required to disambiguate word senses. The resource axis thus concerns where to obtain such information.

2.2.1 Semantic Knowledge Sources

Humans make use of language tools like dictionaries and thesauri to help understand word meanings. As discussed in Chapter 1, such different types of lexical resource arise from different models for defining word senses. For instance, words are ordered alphabetically in dictionaries[3], and consequently there is no mandatory correlation between the physical positions and the meanings of words in the way they are listed. Each word entry has a rich set of information, which typically includes pronunciation, morphology, parts of speech (POSs), sense definitions, example usages, and so on. Senses may be listed in different orders (e.g. etymological, frequency, etc.), depending on individual dictionaries. On the other hand, words in thesauri[4] are grouped into broad conceptual categories, and the relations among the words are somehow reflected by the physical positions of the words. However, the amount of lexical semantic information for individual words can vary from resource to resource. At the same

[2] WSD systems used to be studied in the context of some general language understanding system. The features and operations in the former might therefore be constrained by the architecture of the latter (e.g. Wilks, 1975a; Small and Rieger, 1982; Hirst, 1987). Recent studies tend to treat WSD in a modularised way, putting more emphasis on its own requirements. In doing so, however, these studies are apparently making quite strong assumptions about the point where WSD will be done in a larger system, or that it can be invoked anywhere.

[3] e.g. LDOCE, COBUILD, CIDE, as mentioned in Chapter 1; other examples are the Webster's Seventh Collegiate Dictionary (W7), Collins English Dictionary (COLLINS), Oxford Advanced Learner's Dictionary of Current English (OALDCE), etc.

[4] e.g. Roget's Thesaurus (ROGET), etc.

time, the basic relations among words as in thesauri are often implicitly encoded in the sense definitions in most dictionaries and can possibly be extracted with proper procedures. This offers an alternative way of constructing a thesaurus to that solely by the lexicographer's intuition (Michiels and Noël, 1982).

Some conventional lexical resources also exist in machine-readable forms. We use machine-readable dictionaries (MRDs) to refer to electronic versions of both dictionaries and thesauri. However, such machine-readable versions are usually just the typesetting tapes or plain texts of the resources. Since these lexical resources, in their printed version, are originally intended for human readers, a lot of information which appears obvious to the readers may not at all be obvious to a computer. So although such electronic resources may contain a rich set of information, they need considerable manipulation, customisation and supplementation before they are ready for NLP use (Calzolari, 1982; Byrd et al., 1987; Calzolari, 1988; Chodorow et al., 1988; Boguraev and Briscoe, 1989).

Consequently there are various approaches for obtaining the lexical information needed for WSD. For instance, lexical knowledge can be manually encoded from scratch with much lexicographic input as in early WSD studies like Wilks (1975b) and Small and Rieger (1982). However, although the information is well-structured and tailor-made for the purpose, intensive efforts have to be invested and such work can often be done only in limited scale, as discussed in Section 2.2.2.

Alternatively, semantic information can be acquired (semi-)automatically from resources like MRDs and corpora which contain rich semantic information but are not primarily structured for computational use. As will be discussed in Sections 2.2.3 and 2.2.4, a wide range of lexical information can be systematically acquired from such resources.

Nevertheless, although this approach for obtaining lexical information is more efficient than hand-coding, it may not be entirely sufficient. A single resource is unlikely to contain all types of information which WSD typically requires. For instance, Wilks (1978) proposed using more types of information, linking his Preference Semantics (see Section 2.2.2) to thesaurus rows, although this idea was not extensively tested. Also, Calzolari (1988) suggested that lexical resources should allow information access by logical relations, in addition to traditional alphabetical orders.

Thus a better approach for obtaining semantic information for WSD is to combine multiple (types of) resources. One way is to link various resources via the corresponding senses they share. In that case, one in effect combines the resources (and the information in them) into an integrated repository, but with their original information structures preserved. By combining existing resources, we minimise the manual work but maximise the automatic work for obtaining different types of information for WSD, which would otherwise be too overwhelming if each type of information is encoded from scratch. Nevertheless, since disagreement in sense distinction and granularity among resources is a notorious problem, additional work is required to align senses before combining the resources. More will be discussed in Section 2.2.5.

2.2.2 Hand-coding Semantic Information

In early studies, semantic information was often hand-coded for particular language understanding systems. Such lexical resources usually contain very rich information (often including world knowledge), but that also makes them time-consuming to form and thus severely limited in size. At the same time, the information is usually represented in very specific data structures and accessed with a pre-defined set of operations. This makes the lexical resources unportable.

Wilks' (1975a; 1975b) Preference Semantics System, for instance, represents lexical knowledge with constructs at three levels. "Semantic elements" consist of 70 primitive semantic units expressing entities, states, actions, and so on. These are composed together with brackets and other syntactic constraints to give "semantic formulas", representing the senses of English words. "Semantic templates" define allowable patterns for combining semantic formulas into triples representing message patterns. The vocabulary size was about 500 (Wilks, 1975b) and was claimed to be the largest of any operating deep-structure semantic analyser at that time.

In Small's Word Expert Parser (Small and Rieger, 1982; Adriaens and Small, 1988), each word is treated as a complex procedural knowledge source (i.e. "word expert"), in the form of a discrimination network with context probing questions at the nodes and the contextual meanings at the arcs. The potential contextual involvement of a word with other words (e.g. words expected in the immediate context) is defined by the Sense Discrimination Language and the interaction among word experts (i.e. the flow of the process) is controlled by the Lexical Interaction Language. Each expert is of considerable size, like the one for "throw" was reported to be six pages long already but still expected to be ten times bigger, so the limited coverage of the lexicon can be imagined.

Both Hayes (1977) and Hirst (1987) represented knowledge in semantic networks. Hayes (1977) represented senses as nodes and semantic relations as links, and superimposed some frame-like structures (called "depictions") on them in the "Choosing the Senses of Ambiguous Words in context" system. Hirst (1987) encoded the lexicon into a semantic network of frames, using Frail as the knowledge representation and inference component in a bigger language understanding system. Case slot restrictions and semantic relations like IS-A were contained in each frame.

All of the above studies made serious attack on the lexical resource itself and its application in WSD. However, since hand-coding semantic information is heavy duty, the resultant resources were often limited in scale and therefore could not be tested effectively.

In contrast, WordNet (Miller et al., 1990), developed at Princeton University, is probably the first broad coverage general computational lexical database. Its large vocabulary size may be at the expense of fewer types of semantic information included, but the resource is nevertheless widely available to the NLP community. The database and interface software were designed to be easily portable (Tengi, 1998). The project started as a psycholinguistic one for testing the scalability of relational lexical semantics, but the resulting database turned out to be of great interest to computational linguists. Words are grouped into concepts based on synonymy, forming "synsets" as the building blocks, which essentially represent the senses of words. In the latest version (1.6), there are approximately 90,000 synsets in the database. Synsets are linked to one another by relational pointers (e.g. hypernym, antonym, etc.), forming semantic nets, one for each of nouns, verbs, adjectives, and adverbs. For the noun database, according to Miller (1998), synsets are primarily linked via hypernymy/hyponymy (i.e. IS-A relations) to form hierarchies with unique beginners which correspond to distinct semantic fields (e.g. "entity", "abstraction", "psychological feature", and so on). Other relational pointers for noun synsets include holonym, meronym, antonym, and attribute. Paradigmatic associations between words are well captured through these pointers, but not syntagmatic relations, though the deficiency can be remedied to a certain extent by the work on semantic tagging (see Landes et al. (1998) for the associated project on building semantic concordances). Also, the database is continuously expanding and on-going efforts are made to identify other relations among words. We will make substantial use of the IS-A hierarchy of nouns in WordNet in our work. Information about other parts of the resource in addition to nouns is documented in Miller (1990) (commonly known as the *Five Papers on WordNet*) and considerably updated in Fellbaum (1998).

2.2.3 Acquiring Semantic Information from MRDs

In view of the intensive time and effort needed for hand-coding semantic information into a lexical resource of a reasonable scale, most studies (e.g. Boguraev and Briscoe, 1987; Calzolari and Picchi, 1988; Vossen et al., 1989; Klavans et al., 1990; Vossen and Copestake, 1993) have actually taken the alternative route. They extract information from existing MRDs semi-automatically. This is feasible because MRDs already contain a fair amount of different types of lexical information, often expressed in identifiable patterns. As a result, the apparent "free text" in the sense definitions of dictionaries can in principle be reorganised into structured information, making it "reusable" in NLP systems (Fontenelle, 1990), though it may not be so simple in practice.

(a) From Definitions to Taxonomy

A widely researched aspect of acquiring semantic information from dictionary definitions is the construction of some taxonomy from these definitions. Amsler (1981), analysing the Merriam-Webster Pocket Dictionary, found that "each textual definition in the dictionary is syntactically a noun phrase or verb phrase with one or more kernel terms". Identifying these kernel terms makes it possible to assemble all headword-kernel pairs into a taxonomic semi-lattice, and Amsler's analysis resulted in lattices (linked by the IS-A relation) of 24,000 noun nodes and 11,000 verb nodes respectively.

The success of this kind of analysis depends heavily on the definition patterns, so not all dictionaries are suitable for the purpose. Most studies on the subject thus use either W7 (Chodorow et al., 1985; Markowitz et al., 1986; Klavans et al., 1990) or LDOCE (Nakamura and Nagao, 1988; Vossen et al., 1989; Guthrie et al., 1990; Bruce and Guthrie, 1992; Vossen and Copestake, 1993). The restricted defining vocabulary of the latter has an additional advantage as it imposes further constraints on the syntax of the definitions, making them easier to formalise (Fontenelle and Vanandroye, 1989). Calzolari (1984) did similar research with the Italian language.

However, when done in a large scale, there can be considerable variations for a general *genus-differentiae* definition pattern as often observed in LDOCE, e.g. a "car" is a vehicle (genus) moving on wheels (differentiae). So identifying different defining formulas is important for accurate analysis and extensive coverage. For instance, there can be "empty heads" like "a *type* of ..." and "a *member* of ..." (Chodorow et al., 1985; Guthrie et al., 1990), or genus-headword relations other than simple IS-A (Markowitz et al., 1986; Nakamura and Nagao, 1988; Vossen et al., 1989; Klavans et al., 1990). After identifying the various defining formulas, different parts of the formulas are located in the definitions usually by simple pattern matching (Chodorow et al., 1985; Calzolari and Picchi, 1988; Nakamura and Nagao, 1988), or less often, by serious definition parsing (Guthrie et al., 1990; Klavans et al., 1990).

The genus terms, apart from being identified, also need to be disambiguated to make the resultant taxonomy useful for WSD subsequently. Amsler (1981) probably did genus disambiguation manually and Nakamura and Nagao (1988) did not do it at all, but others have attempted to disambiguate genus terms automatically. For example, Chodorow et al. (1985) disambiguated genus terms via "sprouting", where a tree was grown from a user-selected node, adding hyponyms recursively, though pruning was by human decision. The final tree would be a disambiguated one. Klavans et al. (1990) unified the full parse trees from the definitions with the morphological, syntactic, and lexical patterns defined as criteria for different senses of N1 in <N1 of ...> and NP in <group of plural-NP>. With N1 as "unit" and NP having the feature "human", 80% and 100% disambiguation accuracy was obtained for 222 and 26 samples from W7 respectively. Guthrie et al. (1990) disambiguated genus terms by comparing the LDOCE box codes (for semantics) and subject codes (for pragmatics) between the genus terms and the corresponding headwords, and going for frequency if no match was found. Bruce and Guthrie (1992) tested different weighted combinations of these sources of information on 520 random LDOCE definitions. Accuracy reached about 80% at the best combination, and even 90% if very abstract words (e.g. part, piece, etc.) were left out.

Distinguishing different kinds of defining formula is not only important for extracting genus terms. As far as the formal structure of the taxonomy is concerned, such distinctions are important where property inheritance is to be enforced. The inheritance mechanism can be violated, for example, with disjunctive concepts and compound words, or some genus-headword relation which is not one of strict IS-A. In these cases, modification must be made to the representation in the lexical database to accommodate different kinds of relations and only allow property inheritance where appropriate (Nakamura and Nagao, 1988; Guthrie et al., 1990; Vossen and Copestake, 1993).

All in all, even though a lot of information can be extracted from analysing dictionary definitions, the results may not be tidy enough for ready use. There may be inconsistencies among definitions. Also, the resultant taxonomy may not only be a tangled hierarchy but also contain some ring-like parts since kernels may be cyclically used to define one another (Amsler, 1981; Nakamura and Nagao, 1988; Vossen et al., 1989). Therefore, human intervention is probably needed to clean up the results.

(b) From Definitions to Semantic Network

Veronis and Ide (1990) demonstrated the construction of very large neural networks (VLNNs) from the definitions in COLLINS. VLNNs, unlike the kind of taxonomy seen above, contain less restrictive semantic relations between words. A headword forms a node itself, and so does each of its senses. The word node is connected to its sense nodes by excitatory links, while each sense node is in turn connected by excitatory links to all word nodes of the words appearing in the definition of that sense. At the same time, senses of the same word are connected by inhibitory links. The connection process is repeated a number of times (for words in the definitions in the second round, and so forth). Veronis and Ide's final network consisted of a few thousand nodes with 10-20 thousand transitions. See Section 2.3.3 for the use of VLNNs in WSD.

(c) Other Lexical Semantic Information from MRDs

Apart from untangling definitions for various types of semantic relation between words, there are other types of semantic information which can possibly be extracted from MRDs and can bear on WSD, e.g.:

- **co-occurrence information** given by sense definitions and subject codes in LDOCE (Wilks et al., 1989; Guthrie et al., 1990);
- **lexical properties** like verb classes inferred from definitions and grammatical codes (Fontenelle and Vanandroye, 1989; Fontenelle, 1990; Fontenelle, 1991);
- **zero-affix morphology** identified from definitions and POSs (Krovetz, 1992);
- **selectional restrictions** directly indicated in some MRDs by semantic codes required for arguments.[5]

2.2.4 Acquiring Semantic Information from Corpora

Apart from MRDs, large corpora provide another potential gold mine for semantic information. The amount of information contained could be enormous, but the information is not lexicalised, and has to be inferred from the actual uses of words. Also, what one can learn from corpora depends on how much one needs to do to them. There are various types of semantic information which can in principle be obtained efficiently in large volume, as follows.

(a) Selectional Restrictions

Resnik (1993) took an influential step in the learning of selectional restrictions (SRs) from parsed corpora by formalising selectional constraints as "an information-theoretic relationship between predicates and the taxonomic classes of arguments". He used WordNet as a computational model of the noun taxonomy. The fitness of a class of argument to a given predicate depends on the contrast between the posterior and prior distributions of the class. He thus defined the "selectional preference strength":

$$\eta_i = \sum_c \mathrm{p}(c \mid p_i) \log \frac{\mathrm{p}(c \mid p_i)}{\mathrm{p}(c)}$$

[5] However, Klavans (1988) found that only 59% of the 2,323 verbs in LDOCE marked as requiring a human subject in all senses actually do so, suggesting the selectional restriction information in LDOCE should be exploited with care.

and the "selectional association":

$$\mathrm{A}(p_i, c) = \frac{1}{\eta_i} \mathrm{p}(c \mid p_i) \log \frac{\mathrm{p}(c \mid p_i)}{\mathrm{p}(c)}$$

where p_i is a particular predicate and c is a class of argument. In that case, ranking the "selectional associations" of a particular predicate with different classes of argument would reflect in order how strongly the predicate selects a class relative to others. SRs learned by this method were applied to WSD (see Section 2.3.5).

Ribas (1994; 1995) learned SRs for four common verbs in the Wall Street Journal (WSJ) corpus ("rise", "report", "seek", and "present"), following Resnik's model. The best classes with association scores crossing a threshold were selected, and only disjoint classes were kept. Different ways of score computation and thresholds were tested, and the best results were about 80% precision and 78% recall.

Although syntactically annotated, the corpora used were semantically unannotated. Resnik (1993) believed that unlikely classes would show little association, but Ribas (1994) reported considerable noise from the wrong senses of the words which often led to spurious results. This suggests that some sense disambiguation for the corpora followed by supervised learning may be beneficial for acquiring SRs in this way. Such mutual relationship is supported in McCarthy (1997). She found that if the training corpus was sense-tagged beforehand, the association scores turned out to be slightly stronger for the right classes of selection. When these learned SRs were fed back to the WSD system, it helped verbs with sparse data and those which do not select strongly. Sanfilippo (1997) also tried to acquire disambiguated restriction pairs. Disambiguated clusters of words from all arguments of a particular predicate and from all predicates of a particular argument were first obtained. These "half-disambiguated" pairs were then intersected to disambiguate the undisambiguated halves of one another.

(b) Co-occurrence Information

Broad semantic relations in the form of general co-occurrence or association can be estimated by simple conditional probabilities from large corpora. More specifically, Church and Hanks (1990) identified certain significant syntactic and semantic co-occurrence patterns by means of mutual information I between two words x and y which have probabilities $P(x)$ and $P(y)$ respectively:

$$I(x, y) \equiv \log_2 \frac{P(x, y)}{P(x)P(y)}$$

The value of $I(x, y)$ reflects the significance of the association between x and y. $I(x, y) \gg 0$ if they are genuinely associated, otherwise $I(x, y) \approx 0$ if they just co-occur by chance, or $I(x, y) \ll 0$ if they are complementary to each other. Knowing the relative association between words could help decide the sense of a word from the presence of other words.

Generalising from such pairwise associations, co-occurring words of different topic categories can be learned especially from domain-specific corpora (Riloff and Shepherd, 1997; Roark and Charniak, 1998). Generally, the process starts with a few seed words for a given category (e.g. "vehicles", "weapons") and conditional probabilities of words co-occurring with them in particular syntactic structures (e.g. conjunctions, lists, etc.) with respect to the category are computed. Words at the top of the probability list will be chosen as seed words for the next round, and the process is repeated until enough words are collected for the category. Although the method is efficient and captures many domain-specific words, robustness is questionable. Results seemed to depend on the particular categories (Riloff and Shepherd, 1999). Also, the noise ratio could be unacceptably high, though this is partly remediable by multi-level bootstrapping (Riloff and Jones, 1999).

Extending on the above work, Caraballo (1999) attempted to develop a labelled, WordNet-like, hierarchy of nouns from co-occurring nouns within conjunctive and appositive structures in the WSJ corpus. However, the noise in the corpus was again so high that there were many spurious clusters. Also, words with multiple senses were not properly accounted for, and no particular semantic relation was assumed for the links between the nodes. Therefore, although semantic knowledge can in principle be learned from corpora, in practice it is rather difficult to obtain reliable lexical information solely from them.

2.2.5 Combining Multiple Resources

As seen from the previous sections, encoding all necessary information for WSD from scratch is heavy work. While different types of lexical information might be acquired from some resource, this may not be entirely satisfactory because a single resource may not contain or may not be good for all types of information, and the acquired information is not always complete and reliable. But WSD typically requires a wide range of lexical semantic information. So a better solution is to join various (types of) existing resources, which are exemplars for representing different types of information, into an integrated repository. By *integration*, we mean that the various resources are linked in some way but the different types of information are preserved in their original structures in individual resources. Such a linkage thus enables us to access the different types of information simultaneously, compatibly, and flexibly.

There may be other advantages of combining resources. For instance, different NLP applications may prefer different sense distinctions and granularity as exemplified in particular lexical resources, but all applications may similarly need an extensive set of lexical information. When resources are combined, their information can be enriched yet their original features can be preserved at the same time. Also, in the combination process, the strengths and weaknesses of individual resources would be uncovered, and such findings might help subsequent lexicographic work.

An obvious interface between different resources is the senses in common to them. Hence combining multiple resources can be treated as a task of sense mapping among the resources (Klavans et al., 1990; Sanfilippo and Poznański, 1992; Knight and Luk, 1994; McHale and Crowter, 1994; Klavans and Tzoukermann, 1995; McHale, 1995; Chen and Chang, 1998). However, this may not be straightforward. On the one hand, resources notoriously differ in sense distinction and granularity. On the other hand, the type(s) of information associated with the senses in different resources can be very different (but that is what we want to preserve) which makes the linkage more tricky. For instance, McHale and Crowter (1994) and McHale (1995) tried to map LDOCE senses to the 1,024 categories in the Roget's International Thesaurus (RIT). Only moderate results were obtained for the mappings. We suspect that the knowledge gap between a conventional dictionary and a thesaurus is too big to be crossed in one step, and may be better done with an intermediate step. Such a step could be built on work like Knight and Luk (1994). Further discussion of these studies will be found in Chapter 4, where we will attempt to generalise the sense mapping process across different types of resource, by making the method structurally based. In addition, most studies to date which claim to be using multiple knowledge sources for WSD do not actually make use of multiple types of resource. They just maximise the exploitation of a single (type of) resource. Although various studies set out to study the sense mapping task between different resources and combine them, few results (i.e. the combined resource) have actually been applied and tested in WSD. On the contrary, *we will actually make use of our integrated resource (i.e. multiple resources in combination) to study WSD in this thesis.*

2.3 The Algorithm Axis

We now turn to the actual process of WSD which uses the lexical resources and the information in them. In this section, we first introduce various aspects from which studies on WSD algorithms can be analysed and compared, and then concentrate on the use of different types of information to do WSD.

2.3.1 Analysis of WSD Algorithms

All WSD algorithms share a common objective: to find effective ways of relating the triggering context (TC) embedding the target word to the conventionalised context (CC) specifying the meanings of the word by some type of information, so as to identify the sense of the target word. This point is very important because it threads through all reported algorithms despite their apparent differences. In this subsection, we first introduce various aspects from which we can analyse and compare WSD studies.

(a) Sense distinction

The following kinds of sense distinction are adopted in different studies:

- as listed in existing lexical resources such as LDOCE (e.g. Cowie et al., 1992; Stevenson and Wilks, 1999), Roget's Thesaurus (e.g. Yarowsky, 1992), WordNet (e.g. Leacock et al., 1993; Agirre and Rigau, 1996; Ng and Lee, 1996), etc.

- as in a customised sense inventory (e.g. Small and Rieger, 1982; Hirst, 1987; McRoy, 1992)

- as word clusters derived from texts (e.g. Schütze, 1998)

- as translation difference in a second language (e.g. Brown et al., 1991; Gale et al., 1992a; Dagan and Itai, 1994)

- as simulated ambiguities like pseudo-words (e.g. Yarowsky, 1993; Sanderson, 1994; Dagan et al., 1997)

In addition, while many studies take all possible senses into account, some only consider a reduced set (e.g. Leacock et al., 1993; Bruce and Wiebe, 1994a; Leacock et al., 1998; Towell and Voorhees, 1998), or even only two distinct senses (e.g. Brown et al., 1991; Gale et al., 1992a; Yarowsky, 1995), usually because of data availability and the respective methods used.

(b) Resource for lexical knowledge

Lexical resources provide the repository of CCs for senses. As discussed in Section 2.2, there are different ways to gather semantic information for WSD. Nevertheless, in practice the trend regarding lexical resources may reflect the actual availability of some types of resource at particular periods of time. Thus recent studies mostly found their way to WordNet, MRDs, and large corpora (or a combination of them). For the use of existing resources, particular versions may also give different effects. For instance, studies using Roget's Thesaurus refer to a great variety of different versions of the thesaurus, while those using WordNet often more homogeneously refer to the same few versions.

(c) Contextual information used

The lexical resource determines in what type of lexical information CCs are represented. It therefore also determines what information from TC can be effectively compared against CC, which may make a difference in the actual algorithms. For instance, Ide and Veronis (1998) pointed out three major classes of WSD algorithm: AI-based (e.g. Hayes, 1977; Hirst, 1987; Cottrell, 1988; Kawamoto, 1988), knowledge-based (e.g. Lesk, 1986; Agirre and Rigau, 1996), and corpus-based (e.g. Leacock et al., 1993; Yarowsky, 1995). But in general, TC and CC are usually compared via one or more of the following types of information:

- Pragmatic Information / World Knowledge

  Often richly hand-coded into lexical resources in early studies (as contextual frames, inference rules, etc.), this type of knowledge was mostly used in early AI-based systems (e.g. Wilks, 1975a; Hirst, 1987). Although included in some MRDs in the form of pragmatic/subject codes (e.g. in LDOCE), it is often insufficiently represented and therefore only used together with other types of knowledge (e.g. Stevenson and Wilks, 1999).

- Semantic Information

  Semantic relations are central to WSD, and various types are seen in different lexical resources (see also Section 2.2). Words within some distance in the TC are compared with respect to such relations. The context size and the type of relation are thus two major concerns. We have seen the following designs in past studies: a wide window of words as an unordered set of simple co-occurrence (e.g. Gale et al., 1992a; Yarowsky, 1992), sentential context with words evaluated for similarity with respect to some knowledge organisation (e.g. Lesk, 1986; Cowie et al., 1992; Veronis and Ide, 1990), short context for local information and other positional constraints (e.g. Yarowsky, 1993; Resnik, 1997), among other combinations. *In this thesis, we will study the effect of different types of semantic relation on WSD.* We review their use in past studies more thoroughly in the next few subsections.

- Other information

  This includes the use of syntactic structures (e.g. Dagan and Itai, 1994; Lin, 1997) and other basic information like POS and morphology, often as supplementary information. POS of the target word is shown to be useful if the word has categorial ambiguities (Wilks and Stevenson, 1996). However, we only work with noun ambiguities in this study.

(d) Test materials

Apart from the number of candidate senses considered, the actual test data are also important. Some studies disambiguate all content words (e.g. Cowie et al., 1992; Harley and Glennon, 1997; Stevenson and Wilks, 1999) while others just handle words of a particular POS (e.g. Agirre and Rigau, 1996; Resnik, 1997). The type-token ratio may differ considerably among studies. Some studies test many instances (tokens) of several selected words (types) (e.g. Yarowsky, 1992; Bruce and Wiebe, 1994b; Leacock et al., 1998), while others tackle all word instances in a text (e.g. Agirre and Rigau, 1996; Stevenson and Wilks, 1999). The language of the test materials may also be noteworthy.

(e) Evaluation criteria

Definitions of performance measures vary from study to study. Some WSD algorithms are evaluated in the context of particular applications (see Section 2.4). Some are assessed by commonly used measures like precision and recall. Also, whether results reported cover only polysemous words or include monosemous words overall, and whether they are for the homograph level or more fine-grained sense level, may also be important concerns (see Section 2.5 for further discussion on evaluation).

Hence, WSD studies can be divergent in so many ways, that one must be careful when comparing them. Nevertheless, the common theme that we mentioned at the beginning of this section (i.e. matching TC to CC) will be a good vantage point to survey various algorithms. We will focus on how they relate TC and CC via various types of semantic relation. To introduce such semantic relations, we begin with semantic memory and word associations in general.

2.3.2 Contextual Information and Semantic Memory

The resolution of word sense ambiguities has primarily depended on some measure of word association which reflects the closeness between words in context with respect to some semantic relation, though the process is sometimes supplemented with some non-semantic information. Studies on human semantic processing have largely inspired such a practice in WSD systems. Quillian (1968) proposed a computational model of human memory for storing the "meanings" of words. Apparently, our memory stores not isolated but connected information. So a model of semantic memory should have, in Quillian's words, "the ability to use information input in one frame of reference to answer questions in another". Thus his model has a network structure with a mass of nodes interconnected by associative links of different kinds, though in practice the actual substance in the network is as important as the formal structure of it.

Quillian's semantic network model allows two word concepts to be compared and contrasted via the association between them. This was demonstrated with a simulation program, which was able to say which meaning of each word was involved in each semantic similarity or contrast found – essentially what is expected of WSD. The association between concepts was found by expanding the search for related concepts from the two given concepts through the associative links and applying an activation tag to every concept reached, until an intersection (i.e. when two activation tags meet) was found. This method is generally known as "spreading activation with marker passing", and underlies many later WSD programs.[6]

For example, Hayes (1977) represented word senses as nodes in a semantic network, and superimposed some frame-like structures (called "depictions") on them. Depictions were connected by binders, and there were certain representational constructs to capture special cases of the subset-superset and part-whole relations. These structures functioned with corresponding association rules, so that only the relevant depictions and thus the appropriate senses for the words were activated.

Hirst (1987) also used marker passing with a semantic network of frames, but controlled the process with additional constraints on the number of propagations and on the nodes from which activation is to be spread. Moreover, since parsing and disambiguation were synchronised with text input, where no connection path was readily found for an ambiguous word, producing no definite analysis, Hirst introduced the use of Polaroid words. These were faked semantic objects with packets of knowledge used to substitute the actual analysis for a word temporarily until more information became available for eliminating the inappropriate senses.

The associative links in semantic networks such as those described above may indicate different kinds of semantic relation without being distinguished. We can roughly identify three kinds of association: *broad* semantic relations, *narrow* semantic relations, and *positional* semantic relations. At the same time, hand-coded semantic networks like those above are hard to form and severely limited in scale. Nevertheless, as proposed and done by Quillian himself, the contents in his semantic network could actually be obtained by encoding dictionary definitions accordingly. We have seen in Section 2.2 how semantic information could be obtained from different resources at considerably broad coverage. This suggests that we can in effect simulate a hand-coded semantic network by assembling different types of information extracted from various places, and hence suggests the idea of combining resources. In the following sections, we will review the variant realisations (i.e. the actual use) of different types of semantic relation in WSD outside a hand-coded network.

---

[6]Subsequent findings on semantic priming, e.g. Collins and Loftus (1975), are in accord with Quillian's network representation of semantic memory. Although various modes of lexical access have been proposed and none is inarguably supported by empirical evidence (see Hirst (1987) for a thorough review), the spreading activation model is generally accepted as a plausible mechanism for human semantic processing.

2.3.3 Broad Semantic Relations for WSD

(a) Association as word overlaps

When a word sense is expressed by a group of other words, word overlapping is a common way to measure closeness between two word senses. Sparck Jones' (1986) row intersection method is an example. Lesk (1986) was perhaps the first to use intact dictionary definitions for similar purposes. WSD was done by finding the combination of senses (from all content words in a window) which had most overlaps of words in their definitions. Only informal results of 50-70% accuracy were reported, using a default window of 10 words and definitions in OALDCE.

Despite its simplicity and effectiveness, Lesk's method nevertheless suffers from some disadvantages: (1) The method is closely affected by the size of the dictionary used and the bulk of information in it, as well as the exact wording of the definitions. (2) The method is easily trapped by combinatorial explosion.[7] (3) The method falls short at uncovering more indirect relationships between words, because the "spreading activation" it does is limited to "first order". Later studies addressed these problems as described below.

Wilks et al. (1989) and Guthrie et al. (1990) used co-occurrence data extracted from dictionary definitions to expand the definitions before counting overlaps, to increase the chance of finding some.

Cowie et al. (1992) applied Lesk's method with the algorithm of simulated annealing to optimise the process. An energy function E for a configuration C (i.e. a combination of senses) is computed from the overlaps of words in the definitions and subject codes in LDOCE. More overlaps result in a smaller E, and the aim is to reach a configuration which minimises E. The process starts with the combination of all first senses, and is repeated with new configurations a few thousand times, ending in the final stable configuration which suggests the disambiguated senses. Tests on 50 example sentences from LDOCE gave 47% correct disambiguation at the sense level, and 72% at the homograph level.

Veronis and Ide (1990) extended Lesk's method by building VLNNs from dictionary definitions (see Section 2.2.3). This enables more distant and thus even broader relationships to be captured. Traversing the network with the traditional spreading activation method for a few dozen cycles and with feedback and inhibition sent between nodes accordingly, the strongest activated sense node for each word at the final stable state would suggest the disambiguated sense. VLNN was claimed to be superior to Lesk's method, because it even succeeded at the failed cases of the latter. However, the scalability of the method was still under experimentation.

(b) Association as co-occurrence

Broad semantic relations can also be modelled by simple co-occurrence data from large corpora. Ideally, every instance of a target word in a training corpus is tagged with the appropriate sense s, and statistical classifiers are trained on these tagged examples. Thus in such *supervised learning*, parameters are estimated from all words w

[7] Wilks and Stevenson (1996) remarked that a 12-word sentence could give rise to more than 10^9 sense combinations to evaluate.

co-occurring with the target sense within a given context size. Then for a Bayesian classifier, disambiguation is typically done by finding

$$\arg\max_{s \in S} \prod_{w\ in\ context} \frac{p(w|s)p(s)}{p(w)}$$

where S is the set of all candidate senses of a target word.[8] Variants of this formula may be seen in different studies depending on their individual experimental conditions (e.g. the number of training samples for each candidate sense of a target word).

Gale et al. (1992a), for instance, tested this method on six polysemous nouns (two senses each in terms of French translations), using a 100-word context, and obtained accuracies from 82% to 100%. They also observed that multiple occurrences of the same word in the same discourse always had the same sense, and this "one sense per discourse" property was exploited in some later studies, e.g. Yarowsky (1995).

Supervised learning requires large amount of sense-tagged data to be reliable. Leacock et al. (1993) compared three statistical methods (Bayesian, content vectors, and neural network) for disambiguating the noun "line", with examples from WSJ for 6 out of 25 of its senses in WordNet. They found that the performance for all methods improved with the number of training examples. All reached about 70% accuracy at 200 training samples. However, hand tagging is not only labour intensive but also error-prone. For example, Yarowsky (1995) took three years to hand-tag 37,232 examples, and Ng and Lee (1996) reported only 57% of inter-tagger agreement. Ng (1997) even suggested that a 3.2M sense-tagged corpus[9] would be needed for any broad coverage, high accuracy WSD program, and the tagging was estimated to need 16 man-years of work!

Thus alternatives to hand-tagging are sought. For example, using bilingual corpora, hand-tagging can be replaced automatically by the different translations in a second language (e.g. Brown et al., 1991; Gale et al., 1992a). Class-based sense distinction is another way. Yarowsky (1992) took all instances for words in the same Roget category as training examples for a same sense.[10] Using the Bayesian method and ± 50 words as context to disambiguate 12 polysemous nouns found in the WSD literature yielded an average accuracy of 92%. However, the method cannot be used for finer sense distinctions. Resnik (1995b; 1997) also used class-based sense distinctions but with respect to WordNet concepts.

Yarowsky (1995) proposed *unsupervised learning* from untagged corpora. The training (of decision lists) started with a small number of seed collocations (i.e. representative examples) for each sense of the ambiguous word. More examples were incrementally acquired from the residual set until the residual set became stable. Testing 12 words (two senses each from French translations) gave an average accuracy of about 95%, which was claimed to be comparable to that achieved with supervised methods. Using a similar approach with neural networks, however, Towell and Voorhees (1998)

---

[8] Bayesian classifiers are often used (e.g. Yarowsky, 1992; Leacock et al., 1998). Other types of classifiers like neural networks (e.g. Towell and Voorhees, 1998) are possible, but they may be more expensive to train.

[9] 3,200 most frequent words in Brown Corpus covers 90% of all word occurrences, and each should have 1,000 instances tagged.

[10] It was assumed that noise due to the presence of other polysemous members in the category was relatively small.

found considerable distance between their results and the upper bound (i.e. all examples labelled). Leacock et al. (1998), on the other hand, did unsupervised training from examples of the monosemous relatives of the target words as defined in WordNet. Four of the six test words attained similar results as training from manually tagged examples, but the other two got substantially lower results. It was suggested that some monosemous relatives might be too specific.

Karov and Edelman (1998) avoided manual tagging altogether by augmenting the training set (i.e. sentences of a word W to be disambiguated) with a feedback set (i.e. sentences of the nouns found in the entry of different senses of W in some MRD). Each original sentence in the training set was finally tagged with the same sense as its most similar sentence in the feedback set, after an iterative learning phase. Sentences attracted to the same feedback set were clustered to give the typical use of that sense, and new examples were disambiguated by finding the most similar cluster, i.e. a similarity-based method. Test instances were from the Treebank-2 WSJ corpus; and for the MRD, the online versions of the Webster's and Oxford dictionaries as well as WordNet were used in combination. Tests on four polysemous words (two senses each, altogether 500 samples) gave 92% accuracy.

Mihalcea and Moldovan (1999b) even suggested acquiring sense tagged corpora by searching the Internet using queries with synonyms and definitions of the various senses of a word, and replacing the search phrases with the original word (sense-labelled correspondingly) in the retrieved examples. Manually checking 1,080 examples (for 120 different senses) retrieved from the top ranked documents gave 91% accuracy.

Another major problem for statistical WSD is that of data sparseness, and there are two sources of it. The first is that not all senses of a word have sufficient examples appearing in a corpus. Apart from smoothing as is conventionally done, one can attempt to either maximise the number of samples gathered or minimise the number of samples needed. Some of the methods above for solving the tagging problem also address the data sparseness problem simultaneously by maximising the examples one can get, e.g. Yarowsky (1992). From the minimisation end, there is the example-based (or exemplar-based) approach. A test instance is compared to each example in a representative set for the closest match (see Section 2.3.6 for Ng and Lee's (1996) method). To overcome the supervision and search overhead, Fujii et al. (1998) proposed even more effective sample selection by maximising the training utility. Their method was tested on 11 Japanese verbs, and the amount of samples needed to reach 80% accuracy was reduced to 1/3 of that in random sampling. The second source of data sparseness is that not all novel context will be covered by the training data. For this case, similarity-based methods are often used (e.g. Dagan et al., 1997; Lin, 1997; Karov and Edelman, 1998). Unobserved words are assumed to behave like their similar, observed counterparts.

We can thus see that the tagging and data sparseness problems limit the number of words and their senses tested in most statistical WSD methods. Scalability remains uncertain, but more importantly, only considering two and often very different senses of the test words may have underestimated the difficulty of WSD in general. Leacock et al. (1993) observed non-linear degradation of performance as the number of candidate senses increased (although they only looked at one word). Therefore, there may be more subtle effect from an intra-word level which a distinct, dichotomous split of senses can hardly disclose.

2.3.4 Narrow Semantic Relations for WSD

We refer to narrow semantic relations as some single, explicitly stated relationship between words, such as X IS-A Y, or A is PART-OF B, and so on. They are therefore part of the broad semantic relations discussed above, but are nevertheless handled quite differently when used separately. Since words are more specifically related under such relationships, it allows better formalised metrics for measuring the *similarity* between words.[11] For instance, Sparck Jones (1986) suggested that different degrees of synonymy could be measured in her semantic classification. WordNet is the most widely available, broad coverage lexical resource that organises concepts (as synsets) under narrow semantic relations. Its IS-A hierarchy is frequently utilised.

The simplest way for measuring the similarity between two nodes in such a hierarchy is edge counting (Rada et al., 1989): the more intervening edges, the less similar the two concepts. However, Richardson et al. (1994) found this method less accurate than expected with WordNet, and suggested that domain specificity of the hierarchy could make a difference. Resnik (1995a) found a correlation of only 0.6645 between edge counting with WordNet and similarity judgement by humans.[12] In addition, the links in the hierarchy are not necessarily uniform, and the deeper the hierarchy, the more variable the effect. More elaborated measures of similarity which takes the peculiarity of WordNet into account are thus proposed: we consider here those which have been applied to WSD.

Resnik (1995a) measured similarity between concepts in terms of "information content". The similarity of two concepts was given by the most informative concept in the WordNet hierarchy which subsumed both, while the probabilities of concepts were estimated from corpora. This method resulted in a correlation of 0.7911 with human judgement, which was better than simple edge counting. The measure was adapted to disambiguate noun groupings which were already known to be related (Resnik, 1995b). The algorithm found the most informative subsuming concept for each pair of words in the group, and assigned credits to the senses of the rest which were also subsumed by that concept. Evaluation on 125 words chosen from one Roget category gave an accuracy of about 60%, which was only moderate because the method considered only one of the many relations in a Roget category.

Agirre and Rigau (1996) used "conceptual density" to quantify the distance between concepts. Their measure was made sensitive to the depth and the density of concepts in the hierarchy, such that concepts in a deeper part or a denser part were ranked closer, according to their assumption. The algorithm looked for the concept subsuming the target noun (in the middle of a window) which gave the highest con-

[11] Note that *similarity* is more specialised than *association*, and it is possible that two words are "similar" in some respect without appearing intuitively as such, and vice versa.

[12] The human judgement scores were from Miller and Charles (1991). On the other hand, McHale (1998) found edge counting surprisingly good with Roget's Thesaurus, giving a correlation of 0.8862 with the same set of human judgement. The reason for the difference (of edge counting on WordNet and Roget) is not clear, but the total depth of the hierarchy may matter. Also, we believe that although humans may organise words relationally in the semantic memory, they might not isolate any particular relationship.

ceptual density with the other nouns in the window. Disambiguating all nouns in four SEMCOR (sense-tagged Brown Corpus) files, with window size of 30, gave 43% precision and 34.2% recall at the sense level, and 53.9% precision and 42.8% recall at the WordNet lexicographer file level (comparable to homographs).

Mihalcea and Moldovan (1999a) extended the "conceptual density" measure to enable disambiguation of verb-noun (v-n) pairs. Senses were first ranked by co-occurrence data obtained by conventional Internet searching. For the top two noun and verb senses, conceptual density of the members in common between (i) the group of nouns appearing in the glosses of all verbs in the subhierarchy of a verb sense and (ii) the group of nouns in the subhierarchy of a noun sense in the pair was computed for each combination. For 200 pairs of v-n samples (extracted from SEMCOR files br-a01 and br-a02), accuracies of 86.5% for nouns and 67% for verbs were found.

2.3.5 Positional Semantic Relations for WSD

The third kind of semantic relation holds between specific positions in a sentence. Examples are case slot restrictions, selectional restrictions, and local collocations. Case slot restrictions (such as that "patient" should be flagged by the syntactic object) were often used in AI systems, but seldom in later studies, probably because such information is lacking in other types of lexical resource.

(a) Selectional Restriction

Contrarily, selectional restrictions (SRs), which bind some predicate and argument by the properties of the latter, have been used in many WSD studies. Wilks (1975a; 1975b) was the earliest to exemplify the use of SR for ambiguity resolution in the form of selectional preference (SP). The basic mode of his Preference Semantics was essentially a template matching algorithm. As mentioned in Section 2.2, semantic templates were used to match the heads of the semantic formulas into triples. A basic triple linked up an agent, an action, and an object in a text fragment. Template matching alone might resolve some ambiguities. Other formulas in the fragment were then attached to the basic template if they satisfied the preferences expressed in the subparts of the formulas in the template, and the corresponding dependency was marked. The final network which contained most dependencies would be the preferred interpretation of the fragment, indicating the preferred word senses.

Resnik (1997) also made use of SR for WSD, combining statistical and knowledge-based methods (see Section 2.2.4 for a summary of his treatment of SR). Each candidate sense was assigned a score from the maximum "selectional association" between a predicate and the ancestors of the sense. The sense with the highest score was chosen. Test instances containing the 100 most strongly selecting predicates (identified from the untagged Brown Corpus) for various syntactic relationships (e.g. verb-object, verb-subject, etc.) were extracted from the tagged part of the corpus. Disambiguation of the noun arguments in these instances yielded accuracies from 35.3% (for adjective-noun) to 44.3% (for verb-object). All were better than the baseline of random choices. However, since many verbs and modifiers in English do not select strongly, it was suggested that SR might better be used as a complement to other types of knowledge.

Some dictionaries (e.g. LDOCE, CIDE) have selectional restrictions of verbs in the form of semantic codes, but these are usually inadequate to be used alone.

(b) Local Collocation

Local collocation can be defined generally as use of words in constant patterns, not necessarily idioms but more as habitual usage. Unlike SRs, local collocations may or may not require explicit syntactic constraints. Yarowsky (1993) treated collocation as "co-occurrence of two words in some defined relationship" while such relationships were defined by positions. He tested several types of collocation like "content word to immediate right" and "first content word to left". Binary ambiguities in various forms (e.g. multiple translations, pseudo-words, homophones, etc.) and POSs (nouns, verbs, and adjectives) were tested. Mean entropies of $P_r(Sense \mid Collocation)$ were computed from the training samples for every sense pair and each observed collocation. If a specific word in a given collocation for a test sample was among what was learned, the most probable sense was assigned, otherwise the most frequent sense was returned. An average precision of 95% (though at partial recall of about 50%) was found, and it was concluded that the same sense tended to hold for more than 90% of the time with a given specific collocation – the "one sense per collocation" property. While local collocation could capture some infrequent but useful evidence which was otherwise unavailable, it was observed that different POSs favoured different collocation types and widths of context, although such preference may be language dependent. Despite the apparently good results, the actual effectiveness of local collocation for genuine WSD was masked by the diversity of ambiguity types in the test. Nevertheless, local collocation is used in many studies in combination with other types of contextual information (see below).

2.3.6 Using Multiple Knowledge for WSD

As mentioned, early AI programs had different types of lexical information encoded in their lexical resources. With the Word Expert Parser, for example, WSD was achieved by traversing the sense discrimination network at each word expert, which contained extensive knowledge of the word itself (Adriaens and Small, 1988). However, the importance of such a variety of knowledge was not emphasised since all types of knowledge were already available in a single resource. It is not until different types of knowledge are distributed across (and needed to be obtained from) various resources that we become more fully aware of the importance of using multiple types of knowledge for WSD.

This idea was first seriously investigated by McRoy (1992). Consequently she used several types of knowledge including syntactic tag, morphology, collocation, word association, and selectional preference for WSD within the TRUMP language understanding system. All these different types of lexical information were collected and coded by hand and were distributed across a core lexicon (for general, coarse sense distinctions), a dynamic lexicon (for context-specific, finer sense distinctions), a concept hierarchy, a set of collocation patterns, and a set of cluster definitions. Language understanding was done in three phases. First, all core senses were retrieved based on morphology. Next, the obviously inappropriate core senses were eliminated and specialised senses were triggered where appropriate through a series of pre-processing steps, including tagging, identification of collocations and identification of clusters. Finally, semantic tags were given to the text after parsing and semantic interpretation (where different types of evidence were combined, see below).

Subsequent studies (Bruce and Wiebe, 1994a; Bruce and Wiebe, 1994b; Ng and Lee, 1996; Harley and Glennon, 1997; Rigau et al., 1997; Leacock et al., 1998; Towell and Voorhees, 1998; Wilks and Stevenson, 1998; Stevenson and Wilks, 1999) used different combinations of more or less similar types of knowledge as those used by McRoy, though the information was derived from different resources.[13] No quantitative assessment was available from McRoy's study, while varied results were reported from others.[14] Nevertheless, all these tests provided some support for the use of multiple knowledge for WSD.

Apart from the types of information used, another important consideration is how they should be combined and how to handle conflicting preferences. The most straightforward way is to simply add up their effects on the candidate senses to some score and choose the sense with the best total score. McRoy (1992), for instance, assigned each piece of evidence a positive strength for a match and a negative one for a failure, and chose the sense with the highest sum from all evidence. Similar treatment of different types of information was also seen in many other studies (e.g. Harley and Glennon, 1997; Rigau et al., 1997; Leacock et al., 1998; Towell and Voorhees, 1998).

Ng and Lee (1996), using an exemplar-based method, compared a test sample to each of the training samples with respect to a feature set (i.e. the values for various types of information) for the closest match. The closest match means the shortest distance between two examples, defined as the sum of the distances between the values of all the features of the two examples. A similar approach was used by Stevenson and Wilks (1999) (they called it memory-based learning).

While the above studies treated the various types of information as independent, Bruce and Wiebe (1994a; 1994b) attempted to systematically identify the interdependencies among them from decomposable models, retaining only the most important interdependencies among the features and therefore making parameter estimation easier. Initial tests on the word "interest" (for six of its senses in LDOCE) achieved about 78% accuracy. Although comparable results were obtained for several other words (including other POSs), the method seemed to work best with nouns. However, it is not sure how much importance one can attach to such numerical results, and thus how far one can generalise from just a few test words. Despite its advantage on parameter estimation, the biggest limitation of the method is the need for large amount of sense-tagged samples. Also, the values of the features need to be simple, often dichotomous, and as the number of features increases, it may become impractical when there are too many models to choose from.

[13] Slightly different is Pedersen (2000), who used one type of information, namely co-occurring words, but combined evidence from various window sizes.

[14] We believe the variation is largely due to the difference in sense distinction and test materials. For example, testing a few words with a reduced set of senses often gives 80% or more accuracy (e.g. Bruce and Wiebe, 1994b; Leacock et al., 1998). Stevenson and Wilks (1999) obtained similar results from testing on all content words with LDOCE senses converted from WordNet senses. Harley and Glennon (1997) reported 73% using CIDE senses. Ng and Lee (1996) reported 54%, testing on the Brown Corpus with WordNet senses, although much noise might have been introduced from their human tagging process.

Decision lists provide another way to get an optimal combination of multiple knowledge. For example, Wilks and Stevenson (1998) marked each sense of the target word as either appropriate or inappropriate in the context of each training example which contained a feature vector of several types of knowledge. A decision list was learned from these examples and applied to disambiguate words in novel contexts. General rules were also produced alongside the decision list so that test words not covered by the training data could also be handled.

Although it is widely believed that multiple types of information are needed for WSD, their individual and combined effects are not clearly understood. *We will address this problem with a systematic study on various types of contextual information and the demand of individual target words for such information in WSD.*

2.4 The Application Axis

WSD is often considered a subtask for NLP systems, but is not always researched in the context of a particular application. The major reason is that the actual impact of WSD on an application is at best unproven if the WSD program itself is not perfect. Thus if doing WSD gives less satisfactory results than not doing it, that makes it difficult to say whether WSD is not important for the application or just the WSD program is not good enough. However, the relationship between WSD and an application can be symbiotic, where the application provides a test bed to find out what is essential for that application and what is insufficient about the WSD program. We will look at how WSD relates to two major applications below, namely machine translation and information retrieval, which differ in their requirement for WSD.[15] Apparently the former has much stronger demand than the latter on language processing and hence on WSD.

2.4.1 WSD and Machine Translation

Word sense ambiguity is a major difficulty for machine translation (MT) systems. On the one hand, a word in the source language may have several senses which may or may not be translated differently. On the other hand, there may be more than one way in the target language to express a single source sense. Therefore, WSD is essential for natural and intelligible translations.

Translation patterns as elucidated in a bilingual corpus are thus very useful. Brown et al. (1991) assigned senses to a word w in the source language by asking a single binary question about one word of the context in which w appeared, which was identified by a flip-flop algorithm as the most informative site in the sentence in supplying disambiguation clues. Such questions were added for 500 English words and 200 French words in their translation system. Tests on 100 short sentences in the Hansard corpus resulted in 45 acceptable translations, compared to 37 without the WSD component. However, the method could only handle two-sense ambiguities.

[15]With varied importance, WSD is also believed to play a role in other areas like lexicography and general natural language understanding tasks (Kilgarriff, 1997b), as well as other subtasks in NLP, e.g. structural ambiguity resolution (Binot and Jensen, 1987), accent restoration (Yarowsky, 1994), acquisition of selectional preferences (McCarthy, 1997), and so on.

Bilingual corpora are usually expensive to obtain, so Dagan and Itai (1994) used a bilingual lexicon with a monolingual corpus of the target language. Target word selection was governed by a linguistic model and a statistical model. Word senses in the source language were defined as all possible translations of the word according to the lexicon. Syntactic relation tuples (e.g. (subj-verb: man walk)) were found correspondingly between the two languages, and the plausibility of alternative target tuples was estimated from the corpus. The highest ranked alternative had to be significantly more probable than the rest. When there were multiple ambiguous points in a source sentence, a constraint propagation algorithm was used to find mutually compatible solutions. Tests on Hebrew-English and German-English translations resulted in applicability of 68% and 50%, and precision of 91% and 78%, respectively.[16] It was suggested that the method could also be applied to monolingual WSD.

The evaluation of the effect of WSD on real MT systems is more tricky. However, it is likely that when MT systems are evaluated via a classification of error types, errors under "wrong equivalent" could be due to failure of WSD in the process (independent of a defective dictionary).

2.4.2 WSD and Information Retrieval

Another application which has been claimed to require WSD is information retrieval (IR). On the one hand, a wrong sense of a query term will cause non-relevant documents to be retrieved. On the other hand, relevant documents may not contain exactly the same words as in the query. Query expansion is often needed to increase the chance of retrieving those documents of similar concepts to the query. It is important to get the right sense for the expanded query terms, otherwise much noise would be introduced. Guthrie et al. (1991) suggested using subject-dependent co-occurrence neighbourhoods (i.e. word sets constructed from LDOCE definitions and subject codes) to expand queries. For the same purpose, Mandala et al. (1999) combined Roget's Thesaurus and other corpus statistics with WordNet. The weighting of the expansion terms was computed from their similarity to the whole query, so as to suppress terms expanded from a different sense.

An extensive investigation of the relation between sense ambiguity and IR was done by Krovetz and Croft (1992). In that study, they did IR on a set of titles and abstracts from Communications of the ACM (CACM) and a set of short articles from TIME magazine. They observed that word senses did provide a clear distinction between relevant and non-relevant documents, and there was considerable ambiguity even in a specialised database like CACM. However, performing sense weighting did not give better retrieval results, though such outcomes did not necessarily imply that retrieval performance was indifferent to WSD, as the queries contained common expressions which might give rise to spurious senses.

Other work also suggested, quite contrary to general speculation, that explicit WSD did not appear to boost IR performance.[17] Sanderson (1994), for instance, studied the effect of ambiguity on IR via pseudo-words, with the Reuters text cate-

[16] Baselines for precision were 63% and 56% for Hebrew-English and German-English translations respectively. Note that only the target word selection module was simulated, without any fully functioning MT systems.

[17] In conventional IR, WSD may be implicitly done via joint word matching.

gorisation collection. Pseudo-words of sizes 2, 3, 4, 5, and 10 were tested, and IR was found to be quite resistant to ambiguity, except with very short queries (1 or 2 words). Meanwhile, IR performance at various levels of WSD accuracy showed that it was very sensitive to erroneous disambiguation. This possibly suggests that no disambiguation is better than poor disambiguation for IR. Similarly, Krovetz (1997) observed that zero-affix morphological variants (e.g. "flat" as a noun or an adjective) often carried related senses. However, further experiments on IR using such variants for query expansion mostly produced degraded results. It was suggested that the poor performance of the intermediate processes for identifying the variants (e.g. idiosyncrasies with stemmer, errors from tagger, etc.) could be responsible.

In this thesis, we will study WSD as a self-contained problem in an application-independent setting.

2.5 The Evaluation Axis

Comparative evaluation has long been an underdeveloped axis of WSD, but is starting to get more serious attention. As Resnik and Yarowsky (1997) observed, "there are nearly as many test suites as there are researchers in this field." Also, as WSD research becomes more modularised, standardised evaluation procedures are required for more objective assessment of and fairer comparison among studies.

2.5.1 Assessing and Comparing WSD Systems

Assessment of WSD systems can be qualitative or quantitative. Early studies, testing only few samples, avoided quantitative evaluation altogether (e.g. Wilks, 1975b; Hirst, 1987; Adriaens and Small, 1988; McRoy, 1992). Recent studies tend to test much larger data sets and present results numerically (e.g. Agirre and Rigau, 1996; Ng and Lee, 1996; Harley and Glennon, 1997; Stevenson and Wilks, 1999). Numerical results are commonly reported by a simple accuracy figure (e.g. Lesk, 1986; Veronis and Ide, 1990; Yarowsky, 1995), or in terms of precision and recall (e.g. Agirre and Rigau, 1996; Rigau etal., 1997; Leacock et al., 1998). It may be more compact to compare numbers, but sometimes even individual numbers may not be directly comparable. Also, we believe that qualitative analysis should not be less important, especially if the task is "highly lexically sensitive" as Resnik and Yarowsky (1997) suggested.

The lack of consensus about evaluation is a notorious problem for inter-study comparison. In fact, given the many necessary considerations for a fair comparison (e.g. whether to force every system to use the same resources, should the evaluation be task-dependent, how coarse/fine a kind of WSD is needed, whether to assume WSD as a one-place process in a larger system, etc.), a consensus is difficult to reach. Gale et al. (1992b) proposed a relative scale for measuring WSD performance by estimating a lower bound and an upper bound. The lower bound was established by assigning each word its most likely sense. This, for instance, ranged from 48% to 96% for the dozen test words in Yarowsky (1992). The upper bound was derived from agreement among human judges. In view of the difficulty of a classification task like Jorgensen's (1990) which only resulted in 68% agreement, they replaced the task with one of discrimination. Judges were asked to say whether the target word, with their definitions provided, in each pair of concordance lines (altogether 82 pairs of 9 polysemous words) had the same sense. On average 96.8% agreement was obtained among the five judges. However, Wilks (1998) said that methodologically it was pointless to include the definitions because the judges did not need them to do the task, and the results only showed whether the judges were sense "lumpers" or "splitters". We also believe that a valid upper bound can be estimated only when the judges are doing the same thing as the system is. In that case a recognition task, in which judges actually need to say which definition applies, would be more suitable.

Given proper estimations, the upper and lower bounds offer a relative scale for systems to compare with one another indirectly. Direct comparisons are difficult, since studies often differ in many ways, as already seen in Section 2.3.1. Alternatively, Yarowsky (1992) attempted more direct comparison by selecting test words which were used in other studies. Agirre and Rigau (1996), with a similar intention, additionally implemented some methods from other studies and tested those methods on their own data set. However, this kind of "translation" between data or methods rarely preserves all the inherent characteristics of individual systems. Most notably, the set of candidate senses is not always the same across studies. Therefore, what we need is a framework for common evaluation and test set generation, an important remark in Resnik and Yarowsky (1997).[18] Their proposal catalysed the materialisation of SENSEVAL for the first time in 1998, where their suggestions regarding the size and sampling of the data set, the task difficulty for the annotators, and the accommodation of both supervised and unsupervised systems, were realised.

2.5.2 The SENSEVAL Project

As said, standardising the test set is one of the motivations for SENSEVAL, a project on WSD evaluation. A most remarkable achievement of the project is thus the production of a "gold standard" corpus tagged with correct answers as the data set (Kilgarriff, 1998a; Kilgarriff, 1998b). The corpus and the senses in its associated dictionary developed in the HECTOR project were chosen, mainly for its being unseen in the WSD circle, so that no system would have any advantage. For the pilot SENSEVAL in 1998, 35 words in various POSs were selected, based on word class, frequency, and degree of polysemy. All instances of them (about 200 for each) in the corpus were sampled for the data set. The corpus was annotated via lexical sense tagging (as opposed to textual tagging), where annotators tagged all instances of the same word in a row, then worked on another word.[19] The validity of the tagging was checked by inter-tagger agreement (ITA). Unlike what would have been expected, ITA turned out to be quite high, reaching 95%. Kilgarriff (1999) believed that the crux was in a clear task definition, qualified annotators, a good sense inventory, and double-tagging with an arbitration phase.

The SENSEVAL project was run in this order: defining the task; selecting, tagging, and distributing data; participants running their systems; scoring results; and meeting

[18] Other remarks which they made on WSD evaluation include using a cross-entropy kind of evaluation criterion to take the eliminative ability of the systems into account, adjusting penalty from confusability of subsenses, and tackling the long-standing sense inventory problem by defining senses based on multi-lingual data. The last suggestion was further explored in Resnik and Yarowsky (1999).

[19] Such tagging with actual senses is what we call a recognition task, which is presumably more demanding than discrimination but less so than classification.

for discussion. Participating systems varied a lot in different respects. Unsupervised systems produced an average of 75-80% accuracy, but were generally outperformed by supervised systems (Kilgarriff and Rosenzweig, 1999). The exercise was agreed to be feasible and worthwhile. More importantly, at least the first step has been taken, and the seemingly biggest gap overcome.

In this thesis, we will not contribute directly on the Evaluation Axis with wide-community comparisons. However, we will suggest an alternative way for estimating the upper bound of WSD performance.

2.6 Orientation of our Work

In this chapter, we have reviewed WSD studies along four axes: *Resource*, *Algorithm*, *Application*, and *Evaluation*. It is clear that the use of multiple sources of knowledge is becoming the state of the art for WSD. Starting from this general paradigm, we orientate our work in this thesis via the following remarks.

First, while various types of information useful for WSD can be extracted from different resources, we need an integrated lexical resource which allows us to use such information in combination, flexibly and compatibly. For instance, if knowledge A suggests sense A.1 and knowledge B suggests sense B.3 for a particular test sample, we require some means to say if A.1 and B.3 are compatible or conflicting. This obviously calls for some sense mapping mechanism.

Second, many studies which use multiple types of knowledge for WSD are actually maximising the knowledge they can get from a single resource. This is insufficient since a single resource may not have all types of information required, or some types of information in it may not be rich enough. Few studies have tried using information from multiple resources simultaneously. This is perhaps a consequence of the lack of reliable sense mappings. However, since each resource has its own strengths and weaknesses, we believe that combining resources is an effective way to enrich the range of lexical knowledge for WSD. *Therefore, we will first study the plausibility of combining various types of lexical resource into an integrated, potentially useful, resource* (Chapter 4). *This is the part we play on the Resource Axis.*

Third, although it is generally accepted that multiple types of knowledge are needed, it is nevertheless not well understood how they function individually and in combination. They are often uniformly combined as if they would produce the same effect on every sense of every word. This may not be the case, as the following suggest:

> "Further research is therefore needed to compare the relative importance of different information types and to find optimal ways of combining them." (Dagan and Itai, 1994:p.587)

> "... to date there has been little systematic study of the contribution of different information types for different types of target words. It is likely that this is a next necessary step in WSD work." (Ide and Veronis, 1998:p.21)

In this thesis, we will improve our understanding of the WSD problem by starting a systematic study of contextual information extracted from the triggering context

(TC). We will study this for different types of semantic relation which represent the conventionalised context (CC) of words in our integrated lexical resource (Chapter 5).

Fourth, the assumption that different information types contribute variably to different types of target word depends on, as pointed out in Chapter 1, the belief that different senses are best characterised in different ways. However, all algorithms to date concentrate on "positive evidence". They compare TC to CC for "what is present" to determine the likelihood of a particular sense, but the subtle properties of individual senses, such as how they are best distinguished from others, are mostly overlooked. That is why class-based methods (like those using WordNet concepts) often takes a wrong sense in giving a generalisation for some class. Optimising the combination of information by decision lists and the like addresses this inadequacy in some way, but is still inadequate as data sparseness often causes the system to regress to unsupervised information. *We will therefore propose a sense-based approach to capture the relationship between a target word and the relative importance of different types of evidence for its senses* (Chapter 6), *hence our intended contribution on the Algorithm Axis.*

We will have the least to say on the Application Axis, as we study WSD as a self-contained problem. However, the format of the test materials we use to test the effectiveness of the WSD module will resemble some kind of neutral application (that of general language understanding). We will not contribute directly to the Evaluation Axis either, but will offer an alternative for estimating the upper bound of WSD performance using our test materials, as will be explained shortly in the next chapter.

In other words, our work will be centered on the study of three major parameters in WSD, namely *Lexical Resource*, *(Triggering) Contextual Information*, and *Nature of Target Words*. More importantly, we argue that these three parameters must not be studied in isolation. As we may have suggested above, it is the interaction of all these parameters which shapes the effectiveness of a WSD system. We propose an *Integrated Model* as a framework to view their inter-relatedness, which we will immediately present.

3

An Integrated Model of WSD

In this chapter, we introduce our *Integrated Model* for WSD in Section 3.1 and the three major parameters involved, namely *Lexical Resource*, *(Triggering) Contextual Information*, and *Nature of Target Words* in Section 3.2. We will also, in Section 3.3, describe the materials used in this study and in Section 3.4, explain our motivation for studying WSD in the form of *Cloze Tests*.

3.1 Model Overview

The Integrated Model we propose for studying the inter-relatedness among various parameters in WSD is represented in Figure 3.1.

The three major parts in the model are marked by dashed boxes and labelled **I**, **II** and **III** in the diagram. The possible interaction among them which we will study is described below:

I As mentioned in the last chapter, WSD requires a wide range of lexical information distributed across various types of resource. In the model, several individual resources are combined as an *Integrated Lexical Resource* (ILR). The different types of information (*c1* ... *c5*) from the ILR represent the conventionalised context (CC) of word senses.

II The triggering context (TC) is usually based on a certain amount of text surrounding the target word. Different types of information are extracted from the TC (*t1* ... *t5*). They are compared to the CC from the lexical resource, thus forming the links between **I** and **II** marked by the horizontal dotted lines. The various pieces of evidence are combined in the *Information Filter* (IF) to find the most likely sense.

III The meanings of the target words are represented in the ILR (thus forming a hidden link between **I** and **III**). As suggested earlier, we believe that different senses are best characterised by different types of lexical knowledge. Consequently, the combination of disambiguation evidence should be made sensitive to the word being disambiguated. For this purpose, we introduce a *Sense Type Classifier* (STC), which gives an input to the IF. The IF, receiving inputs from both **II** and **III**, thus links the two parts in the model. Another interaction between **II** and **III** is that the results from the IF would give feedback to the STC, to refine the latter.

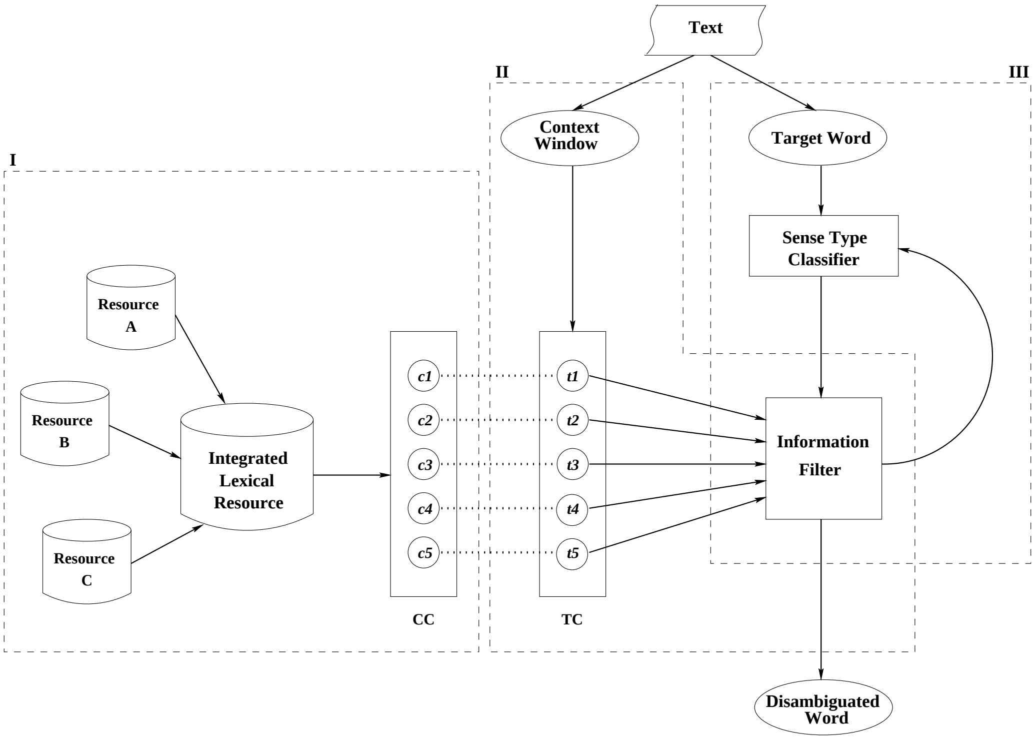


Figure 3.1 *An Integrated Model for Word Sense Disambiguation*

3.2 The Three Parameters

In this section, we introduce the three parameters corresponding to the three major parts in our Integrated Model in more detail, and explain how we are going to study them.

3.2.1 Lexical Resource

We have seen in the literature review that lexical resources providing the CC of words are indispensable in nearly every NLP task. WSD, in particular, requires a wide range of lexical information. These different types of information are nevertheless distributed across different (types of) resources. As a result, an obvious way to acquire the necessary information for the task is to combine two or more different (types of) resources to give an integrated resource.

In addition, as mentioned earlier on, most studies to date which use multiple types of information for WSD are often only maximising their exploitation of a single lexical resource. This is in some sense due to the lack of a good combined resource. However, although a single resource may also contain a rich set of information, it never contains all types of information. Therefore, it is likely that lexical information can be substantially enriched by using more resources together. Lexical resources of different types can supply different types of information, and resources of the same type may be able to complement and fill the gaps in each other.

There is another motivation for our investigation into the combination of lexical resources. As mentioned in Chapter 1, although dictionary senses may be an inadequate representation of senses, it should not prevent us from actually using them in WSD studies. In fact we are more interested in whether the lexical information contained in dictionaries (and other lexical resources) is sufficient for us to relate a novel occurrence of a word to one of its senses listed there. We therefore try to gather different types of information from various resources and use them in disambiguation experiments, to see how effectively such information represents different word senses.

Since a lexical resource is a prerequisite for doing an NLP task, we discuss this parameter first in Chapter 4. We start by designing a structurally-based sense-mapping algorithm to link up a conventional dictionary and a thesaurus, using some mediator resource which shares properties of both. This is followed by a series of pilot experiments. We then proceed with a full-scale linkage between WordNet (a typical mediator) and Roget's Thesaurus (a typical thesaurus) to give the ILR which is used extensively in subsequent chapters.[1]

3.2.2 Contextual Information

By contextual information, we refer to the information drawn from the TC. As discussed in the last chapter, although most people agree that contextual information is essential for WSD, few systematic studies on the effect of different types of contextual information were conducted. Context itself is also only vaguely defined, usually as linguistic context in the form of a window of text surrounding a target word. In this study, we consider such linguistic context in two related dimensions, the *context size* and the *context realisation* (which we define as the computational extraction and application of particular types of information).

On size, Choueka and Lusignan (1985) reported that short contexts (one- or two-word) were sufficient for disambiguation. However, considering the ambiguity types in their test data, we question the generality of the claim. Veronis and Ide (1990) suggested that "additional context can only contribute more to the disambiguation process". We believe that this is not necessarily true, unless one knows reasonably well which part of the context is useful for disambiguating which word.

On realisation, both Leacock et al. (1998) and Towell and Voorhees (1998) found that topical context (like word associations) worked better for nouns while local context (like collocation patterns) worked better for verbs and adjectives. Combining both types of context was beneficial to nouns and verbs, but not adjectives. Nevertheless, the biggest weakness of their studies is that they only tested on one word each for nouns, verbs, and adjectives respectively. Meanwhile, psycholinguistic studies often distinguish between priming/non-priming (or biased/unbiased) contexts for lexical access and disambiguation. As Tabossi (1989) remarked in her conclusion:

> "... one crucial task for future research is to specify the nature and strength of different types of sentential contexts and to determine how they operate in order to yield various effects on the comprehension of words in discourse, ambiguous and unambiguous." (p.39)

[1] Considering resource availability and the mapping effectiveness, no dictionary was used in the full-scale linkage.

This also applies to WSD research, but we need a more precise definition of context which allows us to exert more control on the variable. A computational setting may give us an advantage in this regard.

In Chapter 5, we will report our study on the context parameter in WSD, in terms of size and realisation. In the first part, we study the effect of context size for one type of realisation, namely a narrow semantic relation as defined by the WordNet IS-A hierarchy. We design a recursive filtering algorithm to approximate full-text context, avoiding combinatorial explosion. In the second part, we fix the context size at full-text and test the effect of different types of semantic relation obtainable from the ILR.

3.2.3 Nature of Target Words

The different requirements from words of various POSs in WSD have been addressed from time to time. For example, as Guo said in Wilks et al. (1989):

> "... different kinds of information are essential to the word sense disambiguation process of different categories of words. For example, case roles and frames are important to verb senses whereas for noun senses, specifications of the component parts and the properties of the referents are indispensable." (p.213)

Leacock et al. (1998) also remarked on the inter-POS difference in WSD. On the contrary, the intra-POS variation (i.e. between different words of the same POS) is rarely studied.

As we pointed out in the last chapter, past studies tend to focus on "positive evidence", but overlook "neutral" or "negative" evidence. For example, assume that we have a Bayesian classifier using some broad semantic relation as disambiguation information. It may work fine for a word with one or more senses which have strong associations with some other words. However, not all senses of a word will significantly associate with other words. In that case, either the scores for all candidate senses are indistinguishable, or even the highest score may not be high enough for a decision to be made with confidence. Under such circumstances, the prior probabilities will become dominant and the method is inevitably reduced to one which chooses the most frequent sense. Given that it is the most frequent sense, the chance for a hit and thus the accuracy is likely to be reasonably high. But such apparent effectiveness of the method could be illusory, except when uniform prior probabilities are assumed.

Although the optimal combination of knowledge in certain studies has somehow addressed such intra-word differences, we believe that a more transparent picture is needed. It would be useful to know the relative strength of different evidence for different senses, especially for handling conflicting evidence from the context.

In Chapter 6, we therefore attempt to study how the nature of target words of a given POS (nouns only) affect the difficulty of WSD (e.g. how the number and range of senses affects the response to different types of disambiguation information). We start by analysing the disambiguation results obtained in Chapter 5 with respect to the way senses are characterised in WordNet. We aim to find some patterns which distinguish senses with different disambiguation outcomes. We also tackle the problem from an opposite direction, using the cloze test format (see below) and disambiguating the words (for the blanks) with various types of disambiguation information. We then try to group the test samples into clusters of senses according to their responses to the different information, and infer some properties underlying such clusters.

3.3 Materials

The materials used in this work include a number of lexical resources and some natural language texts, which we briefly describe below.

3.3.1 Lexical Resources Used

As lexical resources, we used conventional dictionaries, thesauri, and some resources which share properties of both. Some of the conventional resources are in machine-readable form. For those which we had only the printed version, the parts we found useful for our study were manually keyed in.[2] In the following, *(e)* refers to electronic versions and *(p)* to printed ones. (See also the initial section of the References.)

- 1911 Crowell Roget's Thesaurus (ROG11) *(e)*
  (Published by the Crowell Company, 1911)
  (Electronic version prepared by P. Cassidy, MICRA Inc., 1991)

  This is one version of Roget's Thesaurus. Like other versions in the main tradition of the 1852 original, words are grouped into approximately 1,000 numbered and named categories of broad concepts. Under each category, words are listed according to POSs, then divided into paragraphs, and further into smaller, more closely related groups separated by semi-colons. Cross references between categories are marked. The categories are grouped into sections and then into classes, in a hierarchical organisation. The electronic version, although in ASCII, has additional features to facilitate computational use (e.g. common prefixes and suffices are expanded, a # to mark the start of a class, and so on). We converted the file into a Prolog-format database for use in this study.

- 1987 Penguin Roget's Thesaurus (ROG87) *(p)*
  (Kirkpatrick, 1987)

  This new version of the Roget's Thesaurus retains more or less the same classificatory scheme as that in earlier versions. The major differences are that more words in contemporary English (such as new concepts from politics, science, etc.) are included and the number of semantic classes is reduced from 1,000 to 990. Also, there are typesetting features such as each paragraph starting with an italicised keyword.

  We supplemented ROG87 with all noun entries in ROG11, and will refer hereafter to this combined version as Roget's Thesaurus (ROGET).

- Longman Dictionary of Contemporary English (LDOCE) *(e)*
  (Procter, 1978)

  LDOCE is an exemplar of conventional dictionaries. Words are listed in alphabetical order, with pronunciation, numbered definitions, grammatical codes,

[2] Several other lexical resources were examined during the study but not used in the work.

examples, and so on. The machine-readable version additionally has subject codes and box codes for NLP use. Sense definitions are ordered by frequency, and organised at the homograph level and a more fine-grained level. Moreover, LDOCE restricts the words used in the definitions to a defining vocabulary of about 2,000 words. In fact, although LDOCE is electronic, using it requires dealing with the typesetting codes embedded in the texts, so we just picked out the definitions for our test words by hand.

- Longman Lexicon of Contemporary English (LLOCE) *(p)*
  (McArthur, 1981)

  LLOCE is a lexicon organised in a classificatory way like Roget's Thesaurus, but uses relatively less abstract category labels. It groups words into everyday concepts under 14 semantic fields (e.g. Life and Living Things, People and the Family, etc.). The definitions are based on those in LDOCE, but the coverage of LLOCE is only about half of LDOCE.

- WordNet 1.5 (WN15) *(e)*
  (http://www.cogsci.princeton.edu/~wn/)

  WordNet is an online lexical database (also available in Prolog format) produced by the Cognitive Science Laboratory of Princeton University. The resource possesses some features of both a thesaurus and a dictionary. It contains nouns, verbs, adjectives, and adverbs, in four databases. On the one hand, words are organised into sets of synonyms ("synsets"). These synsets are related to one another through relational pointers for specific relations, such as antonymy, hyponymy, meronymy, and so on. In particular, nouns are organised into an IS-A hierarchy and a PART-OF hierarchy; and verbs into a hypernym/troponym hierarchy. On the other hand, each synset has an optional gloss definition, which is often less detailed than normal dictionary definitions. Version 1.5 was released in March 1995.

- WordNet 1.6 (WN16) *(e)*
  (http://www.cogsci.princeton.edu/~wn/)

  Released in March 1998, WordNet 1.6 is the newest version of the resource to date. Most change is on the interface, and only little on the content. Apart from the pilot tests in Chapter 4, where we used WN15, for the rest of our work we used WN16 since it was then available.

3.3.2 Test Data

Apart from the set of randomly chosen words used in the pilot studies in Chapter 4, the test data in our study mainly came from two sources:

- Semantic Concordances (SEMCOR) files

  Semantic concordances were created from texts in the Brown Corpus alongside the WordNet project (Landes et al., 1998). In the latest package, there are 186 SEMCOR text files in which all content words are tagged with WN16 senses. The files cover different types of text (e.g. fiction, press reportage, scientific writings, etc.) and each is about 2,000 words long. The letter code together with text examples and some statistics on noun distributions for each text type can be found in Appendix A. We ran our experiments on all the polysemous noun instances of the 186 files, amounting to more than 60,000 test instances for more than 5,000 distinct nouns, with number of senses ranging from 2 to 30, average at 3.41. Words were already lemmatised in SEMCOR, so no further stemming for handling word variations was needed in our tests.

- Cloze passages

  We also studied WSD in the form of cloze tests, texts with blanks to be filled in by the reader. The background and motivation are explained in the next section. In our study, 14 short cloze texts (about 200-300 words each) were chosen from real practice papers, giving 60 blanks for polysemous nouns to form our test set (further described in Chapter 6).

3.4 WSD and Cloze Tests

We mentioned the use of cloze passages for test data. In this section, we will discuss the relationship between WSD and cloze tests, and explain why we use the latter as a tool to study WSD.

3.4.1 Background of Cloze Tests

The concept of *cloze* was first introduced by Wilson Taylor in 1953. Oller (1994) gave a broad definition of cloze as "any procedure that omits portions of a text or discourse and asks readers or listeners to resupply the missing elements". The essence, however, is not just on "gap filling", but more importantly on the use of linguistic and pragmatic constraints from the surrounding context to help guess the missing words (Rye, 1982).

There are two major classes of cloze procedures, namely *fixed-ratio procedures* and *variable-ratio procedures*. For the fixed-ratio type, the distance between blanks is a constant, usually between 4 and 12 words. Therefore, for any $1 : n$ ratio, there will be n alternative ways of deletion of the same text. For the variable-ratio type, we can have *random deletion*, which has been shown to have similar effect as fixed-ratio procedures. However, we are more interested in *rational deletion*, where we can deliberately select the deletion points to test different hypotheses.

Cloze has been used as a measurement device for determining the readability of texts, language proficiency of students, intelligibility of authors, and so on (Rye, 1982; Harris, 1985). In particular, it has been widely used in classrooms in place of conventional reading tests, as cloze tests can avoid questions biased to what the question-setter thinks is important. Recent development has led to different response modes, such as multiple-choice (M-C) instead of open-ended blanks (Jonz, 1976). The M-C cloze allows the reader a limited range of choices to compare with his own response. Research showed that this was as effective as the conventional open-ended format. The following is an example of a blank in an M-C cloze text and a set of words from which the choice is to be made:

"... All the ______ should now be very frightened. ..."	Choices: boys / he / sound / time

3.4.2 Motivation

Relationship between WSD and Cloze

We can thus find many similarities between the process of disambiguating word senses and that of completing cloze tests. First, both depend on the effective use of (triggering) contextual information to arrive at the right answer. In WSD, different types of information are used to find the correct sense of a target word. In cloze, surrounding context provides hints to choose the right word to fill a target blank.

Second, work is only required at particular points of a text. In WSD, presumably only the polysemous words need to be disambiguated. This is comparable to the deletion points (i.e. the blanks) in cloze. Moreover, since there is no rule for where a polysemous word would appear in a text, it therefore resembles rational deletion in variable-ratio cloze.

Third, the candidate senses of a polysemous word in WSD can be thought of as the choices allowed for a blank in an M-C cloze. Although the word choices for a cloze blank could be polysemous themselves or of different POSs, we can restrict the choices to make the situation more like WSD.

Given the commonalities between WSD and cloze procedures, it seems reasonable to hypothesise that the two tasks share a similar mechanism.

Problems with Conventional Test Materials

There are two main types of test for WSD in past studies. One is to test for all word instances (all content words, or just words of a particular POS) in running texts, and the other is for only instances of several chosen words. The current trend is to collect such test data from large corpora. This leads to two ways of corpus sense tagging. The first is textual tagging (or sequential tagging), where annotators assign senses to the word instances in the order as they appear in a text. Resnik and Yarowsky (1997) pointed out a disadvantage of the method: since annotators are forced to repeatedly refamiliarise themselves with the sense inventories of each word, it would slow the speed and lower the agreement rate. The second is lexical sense tagging, where annotators tag all instances of one word regardless of where they occur in the running text, and then work on another word. This is the approach taken in SENSEVAL.

Manual tagging is by no means a comfortable task, but is often needed in WSD studies. In fact different modes of sense tagging may be required for different applications (Miller et al., 1999). As one of our objectives in this project is to show the differential preferences for disambiguation evidence from different target words, we therefore not only need to look at a large number of different test words (for generality), but also need to obtain the test words at reasonable cost (for manageability).[3] In addition, the test materials should be in a format which facilitates detailed qualitative

[3]This is partly a consequence of data limitation, so we have to focus on many different words rather than many instances of particular words.

analysis. The two types of test data gathering mentioned above thus do not really meet these criteria.

What we are looking for is some way in between, and the M-C cloze format seems a promising approach. First, we can control where the blanks should be, so annotators do not need to tackle every single word in a text as in sequential tagging. This makes the tagging more manageable while maintaining a reasonable amount of test samples, and annotators do not need to be bored with tagging thousands of instances of the same word as in lexical sense tagging. Second, we do not just focus on a few words, so generality is ensured. Third, and more importantly, cloze is a standard language test for humans. Preparing test material of this kind should therefore be less demanding.

Practical Advantages

The motivation for studying WSD in the form of cloze tests is therefore based on modelling and processing grounds. On modelling, we have discussed the resemblance between filling out cloze blanks and disambiguating word senses. Such similarity suggests that it is legitimate to model WSD with cloze tests. On processing, we have shown that preparing test data from cloze makes it easier to meet the requirements for test samples in our study.

In addition, there are certain practical advantages. The most obvious is that examples of cloze tests are abundant, and they usually cover balanced topics. Although not all the blanks will be suitable for the study, a certain amount of data is always available. It is also easy to prepare the data from large corpora. All we need is to mark disambiguation points (i.e. which words are to act as blanks) in a controlled manner. Such rational deletion further makes it convenient for us to manipulate the variables we want to study, and control the level of difficulty.

Compared to conventional evaluation methods, studying WSD as cloze gives a more transparent picture of the performance of the system. We can focus on each target blank and immediately visualise how effective a particular type of information is for disambiguating that test sample. As commented earlier, past studies tend to focus on the overall effectiveness of their algorithms, and do not really look at the individual needs of different words. It is therefore impossible to see exactly the specific individual and combined effect of different types of information. In this study, however, we intend to introduce a discriminative element in the process. Cloze tests thus serve as a useful analytical tool. They allow us better control of the test conditions, and therefore the effect of different variables can be more systematically studied. In particular, we can readily see how each type of disambiguation information works for different words.

Cloze tests have another advantage of being a natural and neutral language processing task. They are normal to humans as reading tests. Consequently, we can assume that the most general kind of sense distinction (as that in a normal dictionary) should be appropriate. This is unlike in other applications such as machine translation and information retrieval, where there may be different preferences for sense distinction. Evaluation of WSD via such a neutral language task is not seen in the literature. The closest studies we can find are those on word prediction (e.g. Boggess, 1988; Kakigahara and Aizawa, 1988; Even-Zohar and Roth, 2000). However, they are not directly addressing WSD.

In terms of evaluation, cloze has another strength. It gives a convenient alternative for estimating an upper bound for the performance of a WSD system. We can simply ask human subjects to complete the tests and take their average results. This is similar to Gale et al.'s (1992b) estimation from inter-tagger agreement, but is definitely a less artificial task. Such presumably more reliable results can also function as a rough indicator of the relative complexity of the test materials.

What We are not Doing

Despite the resemblance between WSD and cloze, our modelling of WSD with the latter deviates from genuine cloze procedures in some respects. The most notable difference is that we only handle blanks for nouns. Also, the choices for each blank are the candidate senses of the expected word only, instead of several different words.

This last difference marks the limitation of the achievement we should expect. In particular, we only aim at finding out the differential effectiveness of various types of disambiguation evidence on different words in different contexts. We therefore do not attempt to distinguish the following:

- *Synonyms*: As Jonz (1976) pointed out, it is inappropriate to give second-language users synonymous choices but only score one correct. Similarly, we do not expect our disambiguation mechanism to tell apart this kind of preference of usage, especially when some of the lexical resources we use actually organise words into synonyms and treat them in the same way.
- *Antonyms*: We have assumed that the word which we disambiguate is appropriate for the blank syntactically and semantically, but regardless of its plausibility in the real world. As a result, in a context of "someone escapes from his ______", "creditor" and "debtor" may be disambiguated in a similar way and receive similar scores.[4]
- *Multiple acceptable fillers*: This is a more general case of the last type. For instance, consider "The ______ catches the mouse". Although one might say "cat" sounds better than "dog", we would content ourselves as long as we manage to select the correct animal sense for both words when they fill the blank.

Although the above distinctions also belong to semantic analysis, they nevertheless fall outside the scope of WSD in our project.

So in this chapter, we have given an overview of our WSD model, introduced the materials and explained some of our approaches. In the next chapter, we will start the discussion on the first WSD parameter, namely *Lexical Resource*.

[4] This may not be a very good example since both "creditor" and "debtor" only have one sense (according to WordNet), but we can readily extend the situation to polysemous words. See also next point.

4

Extrinsic Parameter I: Lexical Resource

4.1 Introduction

In this chapter, we discuss the first extrinsic parameter of WSD: *Lexical Resource*. As a repository of lexical information, lexical resources are indispensable for almost every NLP task. Notably a lot of different types of lexical information are required to perform WSD. As reviewed in Chapter 2, combining different types of existing resource seems to be the most efficient way to gather a wide range of information.

Although one can also start from scratch and encode as much information as desired into a lexical database manually, the time and effort required make it so costly that coverage is bound to be severely limited. WordNet (Miller et al., 1990; Fellbaum, 1998) may be an exception, but its broad coverage is an effort of years and is at the expense of the range of information included. Alternatively, acquiring lexical semantic information semi-automatically from existing resources such as MRDs and corpora would not be entirely satisfactory either. First, a single resource may not contain all the necessary information. Second, even if it does (e.g. a comprehensive dictionary might contain thesaural information implicitly in its meaning definitions), the extraction of some information is not always straightforward. For instance, consider the transformation of dictionary definitions to taxonomies (Amsler, 1981; Calzolari, 1984; Chodorow et al., 1985; Markowitz et al., 1986; Nakamura and Nagao, 1988; Vossen et al., 1989; Guthrie et al., 1990; Klavans et al., 1990; Bruce and Guthrie, 1992; Vossen and Copestake, 1993). On the one hand, we need robust methods to handle the considerable variations exhibited in the definition patterns. On the other hand, since dictionaries are infamous for their non-standardised sense granularity, the resulting taxonomies may be ad hoc and unportable. It therefore seems that the obvious solution is to obtain different types of information from various types of existing resource (Klavans et al., 1990; Sanfilippo and Poznański, 1992; Knight and Luk, 1994; McHale and Crowter, 1994; Klavans and Tzoukermann, 1995; McHale, 1995; Chen and Chang, 1998). One way is to integrate resources of different types into a multi-component resource, for example, by aligning the senses in common to them.

By *integration*, we refer to bringing together different types of lexical information originally residing in different resources to a centralised repository. The resources are tied to one another by the set of senses they have in common, while the individual structures of the resources are preserved.

Although such resource integration has been suggested (e.g. Yarowsky, 1992) or implemented in some way (Knight and Luk, 1994; McHale and Crowter, 1994; McHale, 1995), surprisingly few studies have actually made use of multiple existing lexical resources simultaneously in WSD. One study which actually used information distributed across various resources is perhaps McRoy (1992), but the resources were tailor-made in her study and the linkage between them was manually imposed.

In this chapter, we will investigate the *automatic* combination of different types of lexical resource, in particular dictionaries and thesauri. In Section 4.2, we will first review more thoroughly the past studies which made a similar attempt to integrate resources, and then in Section 4.3, we will present our generalised, structurally-based sense-mapping algorithm. Performance measures for the algorithm will be introduced in Section 4.4, and a series of pilot tests reported in Section 4.5. A full-scale linkage of noun senses between two resources to form our *Integrated Lexical Resource* (ILR) will be discussed in Section 4.6, followed by a summary of this chapter in Section 4.7.

4.2 Related Work

McHale and Crowter (1994) attempted to map each LDOCE sense to one of the 1,024 categories in the Roget's International Thesaurus (RIT), in order to enrich the former with the latter. The enriched information was intended for multiple purposes, including its use in a principle-based parser. McHale and Crowter assumed that the way which words were defined in dictionaries would include enough semantic background for sense mapping and also that the physical distance of two words in a relational hierarchy (like RIT) would roughly correspond to the semantic distance between them. The mapping was therefore done by finding the best "measurement of relatedness" (MOR) between the words in the definition (in LDOCE) and the potential categories (in RIT) of the target word. The best MOR achieved 63% success for mapping 100 randomly selected LDOCE senses. McHale (1995) also studied the importance of various properties (transitivity, order, density, and hierarchy) of RIT for mapping by testing different MORs (e.g. method of quartet, number of intervening nodes, etc.). He found that the "hierarchy" property was clearly important, whereas transitivity seemed the least so.

Although McHale and Crowter's (1994) method seems to work to some extent, its complexity depends on not only how many potential categories the target word has, but also the number of words in the definition and how many potential categories these words have. Moreover, since the classification in RIT is broad, many short distances between the target word and the words in the definition may be spurious. Hence, we believe that the gap between a dictionary and a thesaurus is too big to cross in one step. This belief is supported by the apparently better results reported in studies which link dictionaries to some "in-between" resources which share properties with both a dictionary and a thesaurus, though in limited ways.

For instance, Sanfilippo and Poznański (1992) correlated senses in the dictionary LDOCE and the "in-between" resource LLOCE to derive a taxonomy for verbs. Although verb definitions are generally problematic since the genus terms in them are vague, less than 1% of the 500 closely examined matches were infelicitous. Similarly, Chen and Chang (1998) obtained very high precision for linking noun senses between LDOCE and LLOCE. In both cases, however, the substantial resemblance of the contents in LDOCE and LLOCE may give an advantage to the mapping.

Knight and Luk (1994) tried to acquire a large-scale ontology by merging a number of existing resources to enhance the performance of knowledge-based MT systems.

Part of their work involved combining LDOCE and WordNet (WN) automatically. By their "Definition Match Algorithm", each LDOCE definition (including example sentences) was compared with the possible senses of the word in WN.[1] A sense in WN was "defined" by its synonyms, gloss, superordinate, siblings, and super-superordinate, which were assigned weights of 1.0, 0.8, 1.0, 0.6, and 0.6 respectively. Scores were computed from 2-dimensional matrices containing the weights for the words involved in comparison for each sense in each resource, to determine the resultant mappings. At the lowest confidence level, mapping accuracy for low-ambiguity words (exactly two senses in LDOCE and WN) was 75% and that for high-ambiguity words (more than five senses in LDOCE and WN) was 47%. Accuracy increased with confidence level at a rather severe expense of coverage.

Thus, while it may be too difficult to link a dictionary and a thesaurus directly, we propose a two-part algorithm which indirectly links a dictionary and a thesaurus via an "in-between" resource (we will call this resource the *mediator* hereafter). The linkage between a dictionary and a mediator would thus be similar to the above studies (Sanfilippo and Poznański, 1992; Knight and Luk, 1994; Chen and Chang, 1998). However, their algorithms were designed to integrate specific resources and depended on the particular contents in the resources. For instance, they exploited example sentences and subject codes specific to LDOCE; or took advantage of the resemblance between LLOCE and LDOCE. In contrast, we intend to make our algorithm more portable by exploiting the *structural* inter-relatedness among the resources. Our proposal is thus unique in the following respects:

- The algorithm is designed with *types* of lexical resource rather than specific resources in mind, so it is relatively more general and portable.

- The algorithm is structurally based. It is therefore independent of any exclusive characteristics of individual resources and only minimally dependent on the actual contents of them.

- The mapping from a dictionary to a thesaurus is done in two parts, first from the dictionary to the mediator, and then from the mediator to the thesaurus. (The choice of mediator and the direction of mapping will be discussed later.)

In the following, where the mapping algorithm per se is discussed, it should be noted that the mapping applies to generic types of resource. Specific examples are used from time to time simply for concreteness and illustration.

4.3 Structurally Based Sense Mapping

We first discuss in Section 4.3.1 the structural inter-relatedness among the different types of resource which motivates our algorithm, and then describe our sense-mapping algorithm step by step with illustration in Section 4.3.2.

[1] There was also a "Hierarchy Match Algorithm" as a complementary step, which exploited the subject codes in LDOCE.

4.3.1 Inter-relatedness between Resources

Figure 4.1(a) shows the structure of a word entry (e.g. *word*) in a dictionary. All the essential elements in a dictionary, except the alphabetical ordering of words, are shown. Under each entry we find the pronunciation of the word, and one or more meaning definitions (usually numbered) grouped by parts of speech (pos1, pos2). Depending on the size of the dictionary, illustrative examples of the meaning definitions may also be present. Since our design is motivated from the structural properties of generic types of resource, we only show the structures which most resources of the same type share.[2]

Figure 4.1(b) shows the structure of a corresponding entry of *word* in a thesaurus. Unlike dictionaries, thesauri order words by meaning and group them into semantically related concepts. For example, there are word groups at the lowest level (4, as in the figure). Words are usually tied together at this level by specific semantic relations (often but not necessarily synonymy). These word groups are further grouped into more general concepts in level 3 which are in turn more generally categorised in level 2, and so on. Such grouping eventually ends up in some hierarchical structure.

The gap between a dictionary and a thesaurus is big because the effectiveness of sense mapping depends on how well one can work out the semantic relationships (implicitly embedded) among the words in a dictionary definition with respect to the thesaural classes. This is far too constrained by the actual dictionary definitions. For instance, McHale and Crowter (1994) found that results varied considerably with different dictionaries. This possibly means that a more robust algorithm should depend only minimally on the lengths and the exact wordings of the dictionary definitions.

We therefore introduce a mediator into the mapping process to bridge this gap, so that the load on processing dictionary definitions can be lightened. The foremost criterion for such a mediator is that it should share certain structural properties with both a dictionary and a thesaurus so that it can serve the intended purpose. In this case, WordNet seems to be an obvious choice for a mediator, although it is not the only possibility.

The direction of mapping is also important since the mappings are not necessarily one to one. We map senses *from* the dictionary (via the mediator) *to* the thesaurus because of the *specificity* of senses in these resources. Senses in thesauri are defined by classes of words, and they are less specific and less elaborately defined than those in dictionaries. It would be more appropriate to map the more precise senses to the more general ones, so as not to corrupt the generality of the latter.

A schema for the structural inter-relatedness among the various types of resource is shown in Figure 4.2. The arrows in the figure indicate the direction of mapping. In the diagram, we illustrate the generic types of resource with specific examples: Roget's Thesaurus (ROGET) for thesauri, Longman Dictionary of Contemporary English (LDOCE) for dictionaries, and WordNet (WN) for mediators. The version of the resources actually used will be specified later where appropriate (see Chapter 3 for their description).

In the middle of Figure 4.2 is the structure of WN, a hierarchy with synsets at the nodes. A word **x2** has a sense as in synset X. On the one hand, since classes

[2]For example, we do not exploit subject codes in LDOCE and CIDE although the codes are deliberately included in the machine-readable versions of these dictionaries for NLP purposes. However, we might consider using them in the mapping process if these codes become common at some point.

word */pronunciation/* **[pos1]** 1. Definition(word, pos1, 1) ......*(example)* ...... 2. Definition(word, pos1, 2) ......*(example1)* ...... *(example2)* ...... 3. Definition(word, pos1, 3) ...... **[pos2]** 1. Definition(word, pos2, 1) ......*(example1)* ...... *(example2)* ...... 2. Definition(word, pos2, 2) ......*(example)* ......

(a)

<Level 1>	<Level 2>	<Level 3>	<Level 4>
Class 1	*Class 1.1*	Class 1.1.1	*a1, a2, a3, a4, a5, ... a8, a9, a10, ...*
		Class 1.1.2	*b1, b2, ..., b7, b8, ... b13, 14, ...*
		Class 1.1.3	...
	Class 1.2	Class 1.2.1	*c1, c2, ...* ***word,*** *... c15, c16, ...*
		Class 1.2.2	...
Class 2	*Class 2.1*	Class 2.1.1	...
	Class 2.2	...	...

(b)

Figure 4.1 *Structure of Word Entries in (a) Dictionary (b) Thesaurus*

in ROGET are more general, we expect to find not only words in synset X, but also those in the coordinate synsets (i.e. M and P, with words m_1, m_2, p_1, p_2, etc.) and the superordinate synsets (i.e. C and A, with words c_1, c_2, etc.) in the same ROGET paragraph.[3] In other words, the thesaurus class to which **x2** belongs should overlap with $X \cup M \cup P \cup C \cup A$. On the other hand, the LDOCE definition corresponding to the sense of synset X (i.e. D_x) is expected to be similar to the textual gloss of synset X (i.e. $Gl(X)$). Also, since it is not unusual for dictionary definitions to be phrased with synonyms or superordinate terms, we would also expect to find words from X and C, or even A, in the LDOCE definition. That means $D_x \approx Gl(X)$ and $D_x \cap (X \cup C \cup A) \neq \emptyset$. For illustration, we instantiate the schema to a concrete example (**x2** = "player") in Figure 4.3.

[3]Hyponyms of X (i.e. words in R and T) may also appear in the same ROGET paragraph, but we leave them out in order to control for the processing burst they may cause.

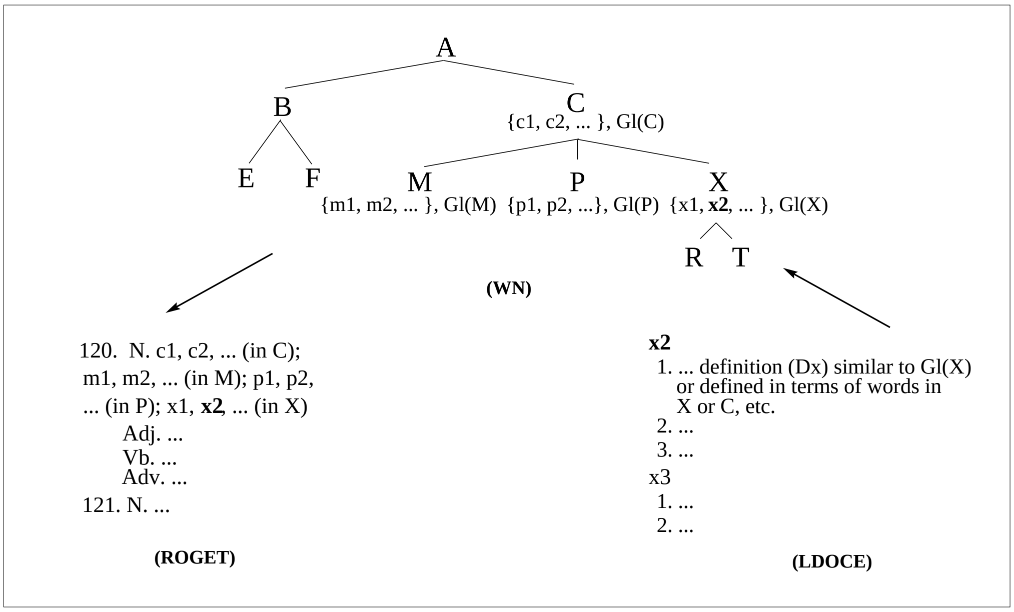


Figure 4.2 *Schema of Inter-relatedness among Resources*

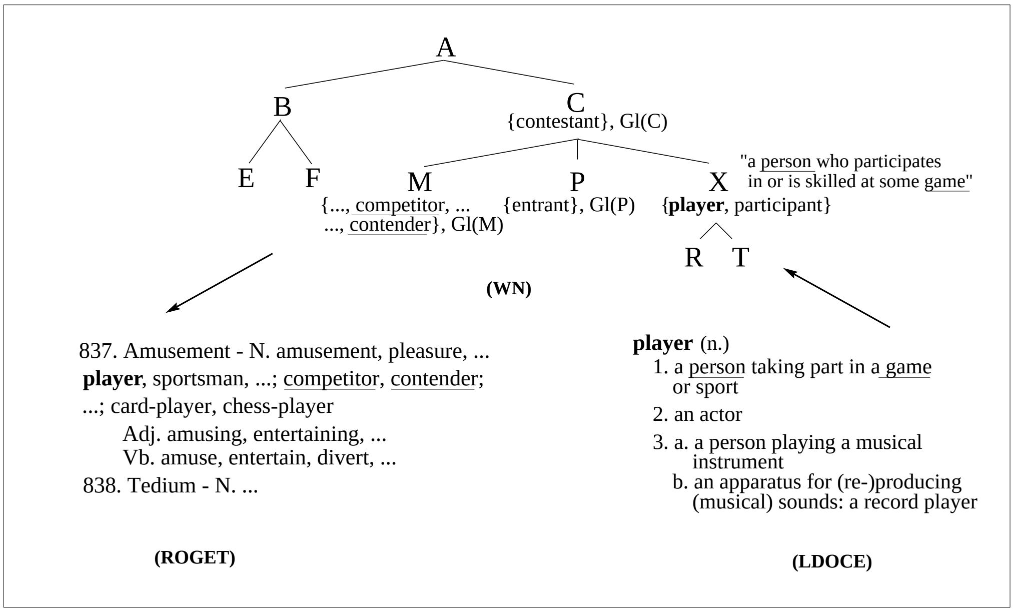


Figure 4.3 *Instantiated Schema of Inter-relatedness among Resources*

4.3.2 The Algorithm

Hence we can exploit the structural inter-relatedness to map senses from the dictionary to the mediator and from the mediator to the thesaurus separately. With the mediator as the backbone and assuming transitivity, a *mapping chain* from the dictionary to the thesaurus through the mediator can then be derived. Our algorithm, which only handles noun senses at present, is summarised in Figure 4.4 and more fully detailed as follows. Again, we use ROGET, LDOCE and WN as exemplars for the various resource types.

Step 1: From LDOCE, get the sense definitions $D_1 \ldots D_t$ under the word W. The definitions for W="copy" are shown in Table 4.1.

#	**Definition in LDOCE**
D_1	*a thing made to be exactly like another*
D_2	*a single example of a magazine, book, newspaper, etc.*
D_3	*written material ready to be printed*
D_4	*good copy, interesting news*

Table 4.1 *LDOCE Definitions for "copy"*

Step 2: From WN, get all the synsets $S_n\{w_{n1}, w_{n2}, \ldots\}$ such that $W \in S_n$. Collect the corresponding gloss definitions $Gl(S_n)$, the chain of hypernym synsets $Hyp(S_n)$, and the coordinate synsets $Co(S_n)$, if available. Such information for "copy" as in WordNet 1.5 (WN15) is shown in Table 4.2.

Step 3: Convert the above data to flattened lists of words, removing punctuations, function words, and repeated words from the definitions (LDOCE definitions and WN glosses).[4] Compute a similarity score matrix $\mathcal{A}$

$$\begin{bmatrix} \mathcal{A}_{11} & \mathcal{A}_{12} & \cdots & \cdots & \mathcal{A}_{1n} \\ \mathcal{A}_{21} & \ddots & \ddots & \ddots & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots \\ \mathcal{A}_{k1} & \ddots & \ddots & \ddots & \mathcal{A}_{kn} \end{bmatrix}$$

for the LDOCE senses and the WN synsets. $\mathcal{A}(i,j)$ is the similarity score between the i^{th} LDOCE sense (i.e. D_i) and the j^{th} WN synset (i.e. S_j). It is a weighted sum of the overlaps between the LDOCE definition and the WN synset, hypernym chain, and gloss respectively:

$$\mathcal{A}(i,j) = a_1|D_i \cap S_j| + a_2|D_i \cap Hyp(S_j)| + a_3|D_i \cap Gl(S_j)|$$

Unlike what Knight and Luk (1994) did, we do not include the WN coordinate terms (called "siblings" in the former) in this comparison. Although coordinate

[4]Words were also lemmatised with an implementation of the Porter stemmer (Porter, 1980) in the pilot tests.

terms are usually abundant, we find that the chance of getting a hit of them in an LDOCE definition is rarely high enough to justify their significance. Our method is also simpler than Knight and Luk's in the sense that we compute the similarity matrix directly from word overlaps, instead of multiplying matrices with weights for every word.

The weight setting was motivated by the relative significance of the terms (in the formula) as evidence for matching senses. Studies on the transformation of dictionary definitions to taxonomies, e.g. Amsler (1981), suggest that words are often defined in terms of their superordinate terms in dictionaries. Dictionary definitions can also be in terms of synonyms (Sparck Jones, 1986), but perhaps less often. Meanwhile, one would expect similar expressions between the LDOCE definition and the WN gloss for the matching senses. However, such expressions may also contain more noise. As a result, we crudely set $a_1 = 3$, $a_2 = 5$ and $a_3 = 2$, in attempt to capture this relative significance. The exact values of the weights may be an empirical issue, but we will not be concerned with it here.

For the example of "copy", we get matrix $\mathcal{A}$ as follows:

$$\begin{bmatrix} 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 \\ 7 & 0 & 9 & 0 \\ 7 & 5 & 3 & 5 \end{bmatrix}$$

Choosing the maximum element in each row and putting it in a symbolic form, we get the following mapping results between the LDOCE senses and the WN synsets of "copy":

	S_1	S_2	S_3	S_4
D_1				•
D_2				
D_3			•	
D_4	•			

Step 4: From ROGET, get all the paragraphs $P_m\{w_{m1}, w_{m2}, \ldots\}$ such that $W \in P_m$. Excerpts of paragraphs for "copy" in the 1987 Penguin Roget's Thesaurus are shown in Table 4.3.

Step 5: As in Step 3, compute a similarity score matrix $\mathcal{B}$

$$\begin{bmatrix} \mathcal{B}_{11} & \mathcal{B}_{12} & \cdots & \cdots & \mathcal{B}_{1m} \\ \mathcal{B}_{21} & \ddots & \ddots & \ddots & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots \\ \mathcal{B}_{n1} & \ddots & \ddots & \ddots & \mathcal{B}_{nm} \end{bmatrix}$$

S	**Gl(S)**	**Hyp(S)**	**Co(S)**
S_1 {*copy*}	*material suitable for a journalistic account; "catastrophes make good copy"*	{{*material*}, {*information, info*}, {*message, content, subject_matter, substance*}, {*communication*}, {*social_relation*}, {*relation*}, {*abstraction*}}	{{*rehash*}}
S_2 {*transcript, copy*}	*copy (e.g. of a legal or school record)*	{{*written_record, written_account*}, {*record*}, {*evidence*}, {*information, info*}, ..., {*abstraction*}}	{{*worksheet*}, {*translation, inter-lingual_rendition, rendering, version*}, {*casebook*}, {*log*}, ..., {*dossier*}, {*chronology*}, {*statute_book*}}
S_3 {*copy, written_matter*}	*matter to be printed: exclusive of graphical materials*	{{*text, textual_matter*}, {*matter, material*}, {*writing, written_material*}, {*written_communication, written_language*}, {*communication*}, ..., {*abstraction*}}	{{*stanza*}, {*machine-readable_text*}, {*draft, draft_copy*}, {*hard_copy*}, {*installment, instalment*}, {*letter, missive*}, {*lyric, words, language*}}
S_4 {*copy*}	*an imitation or reproduction of an original; "she made a copy of the designer dress"*	{{*representation*}, {*creation*}, {*artifact, artefact*}, {*object, inanimate_object, physical_object*}, {*entity*}}	{{*illustration*}, {*projection*}, {*map*}, ..., {*display, presentation*}, ..., {*picture, image, icon, ikon*}, {*pieta*}, {*rubbing*}, {*stage_set, set*}}

Table 4.2 *Information for "copy" from WordNet 1.5*

Class	**Example Words in Class**
P_1 (22/copy)	{*copy, exact copy, clone, reproduction, replica, facsimile, ...*}
P_2 (23/prototype)	{*prototype, archetype, type, norm, ..., copybook, copy, ...*}
P_3 (91/duplication)	{*duplication, ..., restoration, copy, carbon copy, ...*}
P_4 (166/reproduction)	{*reproduction, procreation, ..., copy, duplicate, Phoenix*}
P_5 (529/news)	{*news, good news, ..., information, bulletin, ..., copy, filler, ...*}
P_6 (548/record)	{*record, ..., file, dossier, ..., copy, spare copy, carbon copy, ...*}
P_7 (551/representation)	{*representation, personification, ..., copy, Xerox, ...*}
P_8 (586/script)	{*script, ..., signature, copy, transcript, ..., caption, heading, ...*}
P_9 (586/writing)	{*writing, creative writing, ..., literature, script, copy, works, ...*}
P_{10} (589/reading matter)	{*reading matter, ..., copy, ..., magazine, journal, ...*}

Table 4.3 *Classes for "copy" in 1987 Penguin Roget's Thesaurus*

for the WN synsets and the ROGET paragraphs. $\mathcal{B}(j,k)$ is the similarity score between the j^{th} WN synset (i.e. S_j) and the k^{th} ROGET paragraph (i.e. P_k), and is a weighted sum of the overlaps between the ROGET paragraph and the WN synset, hypernym chain, and coordinate terms respectively:

$$\mathcal{B}(j,k) = b_1|S_j \cap P_k| + b_2|Hyp(S_j) \cap P_k| + b_3|Co(S_j) \cap P_k|$$

As a ROGET class contains words expressing every aspect of the same idea, therefore synonyms, hypernyms, and coordinate terms from WN may be equally good evidence for sense matching. So in this case we set $b1 = b2 = b3 = 1$.

Matrix $\mathcal{B}$ for "copy" is shown below, followed by a representation of the mapping outcome:

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 2 & 1 & 1 & 1 & 1 & 1 \\ 4 & 1 & 1 & 1 & 1 & 9 & 1 & 3 & 3 & 2 \\ 2 & 3 & 1 & 1 & 1 & 1 & 3 & 4 & 2 & 5 \\ 7 & 2 & 1 & 1 & 1 & 3 & 6 & 1 & 1 & 1 \end{bmatrix}$$

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}
S_1					•					
S_2						•				
S_3										•
S_4	•									

Step 6: For $i = 1$ to t (i.e. each LDOCE sense), find $max(\mathcal{A}(i,j))$ from matrix $\mathcal{A}$. Then trace from matrix $\mathcal{B}$ the j^{th} row and find $max(\mathcal{B}(j,k))$. The final mapping is then established between D_i and P_k (or more precisely, the head of P_k). For "copy", the final outcomes are as follows:

$$\begin{array}{llll} D_1 \to S_4 & S_4 \to P_1 & \Rightarrow & D_1 \to P_1 \\ D_2 \to \emptyset & & \Rightarrow & D_2 \to \emptyset \\ D_3 \to S_3 & S_3 \to P_{10} & \Rightarrow & D_3 \to P_{10} \\ D_4 \to S_1 & S_1 \to P_5 & \Rightarrow & D_4 \to P_5 \end{array}$$

4.4 Performance Measures

In this section, we introduce the measures we use to evaluate the performance of the sense-mapping algorithm described above. In the following, *direct mapping* refers to the mapping from the dictionary to the mediator, and also the mapping from the mediator to the thesaurus; whereas *indirect mapping* refers to the mapping from the dictionary to the thesaurus.

It is not at all clear what an objective measure of accuracy for sense mapping between resources should be, mainly because resources do not necessarily have a one-to-one correspondence between the senses they contain. For this reason, we tend to be lenient in judging the appropriateness of the mapping. For both the direct and indirect mapping, we allow various degrees of success, e.g.:

Step 1: *Given word W, collect LDOCE senses.*

Step 2: *Collect WN synsets, hypernym chains, glosses, and coordinate terms for W.*

Step 3: *Compute similarity score matrix A.*

Step 4: *Collect ROGET paragraphs for W.*

Step 5: *Compute similarity score matrix B.*

Step 6: *Derive mapping chain from matrices A and B by transitivity.*

Figure 4.4 *Summary of the Two-Part Sense-Mapping Algorithm*

- *Accurate*: mapping can be judged as correct right away.
- *Alternatively Accurate*: mapping is not related directly by the genus term, BUT can be related by other properties conveyed by the sense.
- *Barely Accurate*: mapping does not fit the above two criteria, but the overlapping words are convincing enough AND there is apparently no better match possible.

However, though we have these rules of thumb for our judgement, with borderline cases we have but our intuition to count on; and since it is quite difficult to distinguish such varying degrees of accuracy in practice, we will group them into a single category called *Accurately Mapped* (A).

On the contrary, errors can be more easily marked as different types. For the direct mapping, we will identify four types of error:

- *Incorrectly Mapped* (I): mapping is unsatisfactory AND a better alternative can be found.
- *Forced Error* (F): mapping is unsatisfactory AND no better alternative is available.
- *Unmapped-a* (Ua): mapping is empty BUT a mapping is expected;
- *Unmapped-b* (Ub): mapping is empty AND none of the alternatives is good enough.

The error types for the direct mapping are summarised in Table 4.4. We can see that I and Ua are *misses* because they fail to identify the expected target. On the other hand, F and Ub are *false alarms* because they produce something when nothing is expected.

Most of the error types described above also apply to the indirect mapping, except that Forced Error would be replaced by *Carry-over Error* (C). The latter is defined as "mapping is wrong as a result of a faulty outcome from the first stage

Target Exists	**Mapping Outcome**	
	Wrong Match	No Match
Yes	*Incorrectly Mapped*	*Unmapped-a*
No	*Forced Error*	*Unmapped-b*

Table 4.4 *Error Types in Direct Mapping*

(i.e. dictionary→mediator)", in contrast to Incorrectly Mapped which is a fault due to the second stage (i.e. mediator→thesaurus).

The categorisation of the errors for the indirect mapping is thus slightly different from that for the direct mapping. The scheme of error categorisation for the indirect mapping, with respect to the outcomes in the two pairs of direct mappings, is shown in Table 4.5. The symbol U means that we need to look at the actual data to decide whether the empty mapping should be classified as Ua or Ub.

<table>
<tr><th rowspan="2">L→W</th><th colspan="5">W→R</th></tr>
<tr><th>A</th><th>I</th><th>Ua</th><th>F</th><th>Ub</th></tr>
<tr><td>A</td><td>A</td><td>I</td><td>Ua</td><td>I</td><td>Ub</td></tr>
<tr><td>I</td><td>C/A</td><td>C</td><td>U</td><td>C</td><td>U</td></tr>
<tr><td>Ua</td><td colspan="5">U</td></tr>
<tr><td>F</td><td>C/A</td><td>C</td><td>U</td><td>C</td><td>U</td></tr>
<tr><td>Ub</td><td colspan="5">U</td></tr>
</table>

Table 4.5 *Error Categorisation for the Indirect Mapping*

In addition, we will also present the results using more general numeric measures such as *precision*, *recall*, and *coverage*. They are defined for the direct mappings as follows: (For the indirect mapping, F is replaced by C in the definitions.)

$$precision = \frac{A}{A+I+F}$$

$$recall = \frac{A}{A+I+F+Ua+Ub}$$

$$coverage = \frac{A+I+F}{A+I+F+Ua+Ub}$$

4.5 Pilot Tests

In this section, a series of pilot experiments on our sense-mapping algorithm is reported. The specific resources used were LDOCE, WordNet 1.5 (WN15), and the 1987 Penguin Roget's Thesaurus (ROG87).[5]

[5]The relevant data for the test words from ROG87 were manually input.

4.5.1 Experiment P1

Test Materials

The algorithm was first tested on a block of 12 nouns, picked with reference to a list of English word frequency with respect to the LOB corpus.[6] The test nouns are listed in Table 4.6. Columns L, W, and R show the number of senses of the words (strictly headwords) in LDOCE, WN15, and ROG87 respectively. The average number of senses is around six, in all the three resources.

Word	**L**	**W**	**R**	**Word**	**L**	**W**	**R**
Country	5	4	3	Matter	7	5	8
Water	8	8	9	System	5	8	6
School	7	6	3	Interest	6	8	14
Room	5	4	3	Voice	9	8	4
Money	2	3	1	State	6	5	7
Girl	5	5	4	Company	9	8	10

Table 4.6 *Test Nouns in Experiment P1*

Results and Discussion

The performance of the algorithm on the 12 nouns is shown in Table 4.7. For the direct mappings, the accuracy was around 70% for L→W and 75% for W→R. The accuracy for the indirect mapping, L→R, was about 55%. This lower value is not surprising since it is contingent upon the success of the two direct mappings. Notwithstanding the peculiarities of the specific resources, these results do provide support for our hypothetical structural relationship among the resources and demonstrate the plausibility of the proposed sense-mapping algorithm. The sample size is small, but the test nouns are among the most common words and should therefore be representative enough for us to make meaningful observations from the results.

Performance Measures	**L→W**	**W→R**	**L→R**
Accurately Mapped (A)	68.92%	75.00%	55.41%
Incorrectly Mapped (I)	12.16%	1.39%	4.05%
Unmapped-a (Ua)	2.70%	6.94%	13.51%
Unmapped-b (Ub)	13.51%	5.56%	16.22%
Forced Error (F)	2.70%	11.11%	–
Carry-over Error (C)	–	–	10.81%
Precision	82.26%	85.71%	78.85%
Recall	68.92%	75.00%	55.41%
Coverage	83.78%	87.50%	70.27%

Table 4.7 *Results of Experiment P1*

[6] gopher://gopher.sil.org:70/00gopher_root%3A%5Blinguistics.info%5Dlob.freqs. The frequencies of the chosen words range from 262 to 685.

The actual results certainly depend on the versions of the resources used. For example, we cannot expect the 1852 edition of Roget's Thesaurus to cover terms developed in the past few decades. Similarly, senses which are less common nowadays may not appear in dictionaries at all. Nevertheless, our algorithm has the advantage of being structurally based and is portable for different versions of the resources. Therefore, once we are equipped with more edition-compatible resources, results should improve. There is, however, an inherent limitation of sense mapping because a resource will unavoidably have senses which are absent from other resources, and lack senses appearing in other resources. For instance, in this experiment we found that 18.92% of the LDOCE senses, 11.11% of the WN15 synsets, and many ROG87 classes for the 12 nouns were unattached to any linkage. We will discuss later how we can make use of such extra senses.

Meanwhile, the errors in fact provide support for our algorithm. In Section 4.4, we mentioned that the errors for the direct mappings could be divided into misses ($I + Ua$) and false alarms ($F + Ub$). A simple comparison shows that there are relatively more false alarms than misses, especially in W→R. This implies that failure is often due to missing senses in individual resources, rather than the inadequacy of the mapping process itself. Moreover, the algorithm can potentially be refined to reduce the number of errors, with respect to the following problems observed from the results (although we did not handle them in this study):

- Definition analysis

  In L→W, we did not parse the LDOCE definitions to locate the genus terms. This means certain definitions, including those with "empty heads" as discussed in Chodorow et al. (1985), cannot be properly handled. Here are examples of other problematic cases which result in sense mismatch:

  - Unresolved terms

    The relatively small defining vocabulary in LDOCE (about 2,000 words) and the style of its definitions do not allow senses to be verbosely defined.[7] Definitions thus often end up with a lot of deictic terms. For example, the second LDOCE sense of "water" is "a supply of this liquid ...", while "this liquid" refers to what is defined in the previous sense. Leaving such reference unresolved might result in a loss of information.

  - Hidden relationship

    For example, the LDOCE definition of "water" as "liquid produced by some part of the body ..." was incorrectly mapped. Although the WN15 senses of "water" include "salty fluid secreted by sweat glands" and "saline fluid secreted by lacrimal glands", the fact that "sweat glands" and "lacrimal glands" are "body parts" could not be disclosed without more refined definition analysis.

[7] McHale (1995) found the small defining vocabulary of LDOCE an advantage. Comparing the effect of using LDOCE and COLLINS, the former gave better results. It was suggested that although the defining vocabulary in LDOCE was small, most words appeared in RIT; whereas the one in COLLINS was richer, but there were also more extraneous words.

- Compound terms handling

  The W→R mapping introduces a dilemma. If we break up compound terms and compare individual words, we lose the specificity (e.g. obviously "geographical unit" should not overlap with just any "unit"). However, if we stick to the originals, we end up with far fewer overlaps and miss certain acceptable cases (e.g. "geographical area" would not overlap with "geographical unit").[8]

We now look at the indirect mapping more closely. Although its success is largely determined by the success of the two direct mappings, the accuracy (55.41%) turns out to be better than what one could possibly get by chance (68.92% $\times$ 75.00%). The results of the indirect mapping in Table 4.7, according to the error categorisation in Table 4.5, simply measure the final outcomes. They do not suggest the reasons for failure. Here we offer an alternative way of analysing the data, which pays more attention to what happens during the process. This alternative scheme of error categorisation for the indirect mapping is shown in Table 4.8.

<table>
<tr><th rowspan="2">L→W</th><th colspan="5">W→R</th></tr>
<tr><th>A</th><th>I</th><th>Ua</th><th>F</th><th>Ub</th></tr>
<tr><td>A</td><td></td><td colspan="2">BREAKDOWN IN W→R</td><td colspan="2">INADEQUACY OF ROGET</td></tr>
<tr><td>I</td><td rowspan="2">BREAK-DOWN IN L→W</td><td rowspan="4" colspan="4">OTHERS</td></tr>
<tr><td>Ua</td></tr>
<tr><td>F</td><td rowspan="2">INADE-QUACY OF WN</td></tr>
<tr><td>Ub</td></tr>
</table>

Table 4.8 *Alternative Categorisation of Errors in L→R*

The breakdown of errors in L→R according to the above scheme results in a distribution as shown in Figure 4.5. In contrast to what is suggested in Table 4.8, where there are 16 slots which could contribute to "OTHERS" but only two each for the rest, it turns out that "OTHERS" contains the least number of cases (only 12.12%). Instead, most failure comes from senses which appear in LDOCE but not in WN15.

[8] We can also consider using the cross references in ROG87 to increase the chance of finding word overlaps. We did not do so because cross references may not be present in every thesaurus, and how far they should be used is also a critical question. For example, the classes in ROG87 for "country" do not relate to people, but those for "nation" do. Given that "country" and "nation" belong to the same synset in WN15, does this relationship simply carry over to ROG87? The spread of links must be carefully controlled, or we will end up with information which is too diffused.

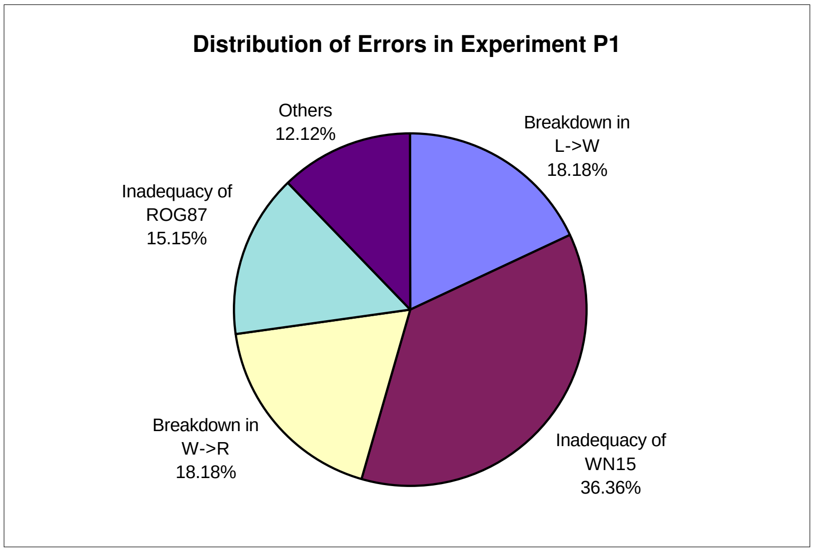


Figure 4.5 *Distribution of Errors in Experiment P1*

4.5.2 Experiment P2

Motivation

Despite its small sample, Experiment P1 has given us an idea of how well the algorithm might be expected to work. The results also suggest that WN15 has quite different senses from LDOCE (see Figure 4.5). Although our sense-mapping algorithm itself does not depend on particular instances of various resource types, the actual effect and hence the usefulness of the resultant integrated resource does. Therefore, in addition to showing the plausibility of the algorithm, we also want to look at how the particular resources differ in the senses they contain, to help us select the appropriate resources to combine (when we have a choice).

Test Materials

In this experiment, we will study how well WN15 serves as the mediator by seeing how it differs from LDOCE and ROG87 with respect to the polysemy of words, as defined by the number of senses a word has in WN15. Three groups of test samples were prepared: the "low polysemy group" (LO) for words with five or fewer WN15 senses, the "medium polysemy group" (MED) six to ten senses, and the "high polysemy group" (HI) eleven or more senses. There were 12 randomly chosen nouns in each group. The test words and the number of senses they have in the various lexical resources[9] are listed in Table 4.9.

[9] There is a slight discrepancy from what was reported in Kwong (1998). It was discovered after the paper was published that one ROG87 paragraph for the noun "copy" had been missed. Correcting

LO Group				MED Group				HI Group			
	L	**W**	**R**		**L**	**W**	**R**		**L**	**W**	**R**
cheque	2	1	1	step	9	9	9	life	20	13	8
plan	5	3	5	number	11	10	6	service	16	14	10
party	9	5	10	statement	4	6	9	field	12	15	11
bench	5	5	6	strike	3	6	5	card	8	11	10
industry	4	3	5	power	16	9	7	hand	12	13	13
society	6	3	6	picture	10	8	6	face	8	13	7
chance	6	4	2	sense	8	7	4	case	17	16	14
copy	4	4	10	credit	8	8	8	air	13	13	7
music	5	4	2	note	10	9	10	place	21	15	5
current	4	3	3	title	4	8	6	piece	16	11	10
letter	5	4	5	cast	8	9	13	call	15	12	5
ceremony	3	3	4	bridge	7	9	5	block	8	11	11
(Average)	4.83	3.50	4.92	(Average)	8.17	8.17	7.33	(Average)	13.83	13.08	9.25

Table 4.9 *Test Nouns in Experiment P2*

Additional Evaluation Measures

In addition to the performance measures in Section 4.4, we also compare the number of senses the test words have in different resources by simple correlation. Also, a *Polysemy Factor* ($\mathcal{P}$) was defined as follows to characterise the relationship between the source ($\mathcal{S}$) and the target ($\mathcal{T}$) resource in the direct or indirect mappings:

$$\begin{array}{rcl} Let & & \\ f & : & \mathcal{S} \rightarrow \mathcal{T}\ (human\ judgement) \\ g & : & \mathcal{S} \rightarrow \mathcal{T}\ (algorithm) \\ \mathcal{T}' & = & \{t \in \mathcal{T} : t = g(s) = f(s)\ for\ some\ s \in \mathcal{S}\} \\ \mathcal{S}' & = & \{s \in \mathcal{S} : g(s) = f(s)\} \\ \mathcal{P} & = & \dfrac{|\mathcal{T}'|}{|\mathcal{S}'|} \end{array}$$

$\mathcal{P}$ is therefore a ratio between the minimum number of target senses covering all mapped source senses and all mapped source senses. It reflects the granularity of senses in $\mathcal{S}$ with respect to those in $\mathcal{T}$. Consequently, by relating $\mathcal{P}$ to the mapping accuracy, we can tell whether any apparent difference in the number of senses between resources is due to more fine-grained senses in one resource or completely different senses in the resources.

Results and Discussion

The correlations of the number of senses between resources are shown in Table 4.10. The upper half of the matrix shows the overall correlations while the lower half shows those for individual polysemy groups. (An asterisk means the correlation is statistically significant at .05 level.) Overall significant positive correlations (at $p < .05$) were found between each pair of resources. This suggests that in general the more senses a word has in one resource, the more senses it tends to have in another resource. However, this relationship of polysemy is unstable with ROG87, especially for the MED and HI words, as seen in its extremely weak correlations with other resources for these polysemy groups.

---

this mistake increases the accuracy by a small margin, but the change is minor and does not affect our claims. The results reported hereafter are the updated results.

	LDOCE	WN15	ROG87
LDOCE	–	0.8227∗	0.4124∗
WN15	L: 0.7170∗ M: 0.6778∗ H: 0.4874	–	0.6004∗
ROG87	L: 0.5212 M: 0.1026 H: -0.2424	L: 0.5672 M: 0.2605 H: 0.1451	–

Table 4.10 *Correlation of Number of Senses between Resources*

Table 4.11 shows the results for the direct mappings in Experiment P2. First, the average accuracy for L→W was 61.12%. Results were better for the LO and MED groups than for the HI group. We found in Experiment P1 (and also this experiment, see below) that failure in L→W was often due to senses appearing in LDOCE but not in WN15. However, on the one hand, since $\mathcal{P}$ was always quite high, the mapping in L→W was almost one-to-one. On the other hand, LDOCE and WN15 correlated quite strongly in the number of senses they have for words (see Table 4.10), and the mean numbers of senses in the two resources were actually similar (see Table 4.9). We can therefore additionally infer that not only does LDOCE have senses which are not in WN15, the latter also has a similar amount of senses lacking in the former.

Second, the average accuracy for W→R was 73.46%, and the LO group performed much better than the others. On the one hand, we see that there are more false alarms than misses, which suggests that WN15 has senses which ROG87 lacks. On the other hand, $\mathcal{P}$ is much lower in MED and HI, showing that many senses in WN15 are too fine-grained to be distinguished by the coarser ROG87 classes. Although it also suggests that ROG87 has senses exclusive to itself, the extremely weak correlation between WN15 and ROG87 on the number of senses at MED and HI makes it difficult to estimate the amount.

Although WN15 has similar number of senses for words as LDOCE, and the success rate is higher for W→R than L→W, it does not follow that LDOCE has more senses than WN15 which are not in ROG87. It only suggests where WN15, as a mediator in the mapping process, falls short.

However, the average accuracy reported above may have been underestimated. Of the 88,200 nouns in the Prolog database of WN15, only 0.07% belongs to the HI group, with 13.18 senses on average; and 0.55% belongs to the MED group, with 7.05 senses on average. The rest 99.37% are all LO group nouns, with 1.18 senses on average. As a consequence, the weighted average accuracy should be much closer to that of the LO group. Thus a more realistic estimate of the average accuracy for L→W would be $64.77\% \times 0.9937 + 65.63\% \times 0.0055 + 52.96\% \times 0.0007 = 64.76\%$. This value is comparable to or even slightly better than Knight and Luk's (1994) average "pct correct" at their lowest confidence level. For W→R, the weighted average accuracy is 80.89%.

Performance Measures	L→W			W→R		
	LO	**MED**	**HI**	**LO**	**MED**	**HI**
Accurately Mapped (A)	64.77%	65.63%	52.96%	80.97%	69.63%	69.79%
Incorrectly Mapped (I)	15.37%	11.82%	15.17%	0.00%	6.94%	3.64%
Unmapped-a (Ua)	2.08%	3.13%	2.57%	2.08%	1.39%	2.55%
Unmapped-b (Ub)	11.25%	11.48%	8.46%	15.28%	14.03%	7.47%
Forced Error (F)	6.53%	7.95%	20.83%	1.67%	8.01%	16.56%
Polysemy Factor ($\mathcal{P}$)	0.79	0.88	0.91	0.80	0.66	0.59
Precision	74.91%	76.29%	59.36%	90.00%	82.55%	77.63%
Recall	64.77%	65.63%	52.96%	80.97%	69.63%	69.79%
Coverage	86.67%	85.40%	88.96%	82.64%	84.58%	89.98%

Table 4.11 *Results of Direct Mappings in Experiment P2*

Similar to Experiment P1, errors for the direct mappings in Experiment P2 are also mostly false alarms ($F + Ub$). Moreover, Ub always outnumbers F by a large margin except in the HI group. This is in fact a favourable phenomenon because we would rather have the target-less senses remain unmapped than have wrong mappings which may adversely affect subsequent operations on the mapping outcomes.

We now look at the results for the indirect mapping L→R, which are shown in Table 4.12. Again, the average accuracy of 54.07% may have been underestimated, as its weighted counterpart is 59.68%. This value is lower than those of the direct mappings, but is higher than what one can get by chance. The anomalously high C in the HI group once more suggests the idiosyncrasy of highly polysemous words, but fortunately such idiosyncrasy should not be fatal given the small proportion of HI words out of all words.

	L→R thro WN15			
	LO	**MED**	**HI**	**Average**
Accurately Mapped (A)	59.72%	55.69%	46.79%	54.07%
Incorrectly Mapped (I)	0.00%	8.88%	6.95%	5.28%
Unmapped-a (Ua)	12.22%	8.42%	7.97%	9.54%
Unmapped-b (Ub)	14.17%	15.48%	9.26%	13.66%
Carry-over Error (C)	13.89%	11.54%	29.02%	18.15%
Polysemy Factor ($\mathcal{P}$)	0.72	0.69	0.59	0.67
Precision	75.83%	73.88%	56.12%	68.61%
Recall	59.72%	55.69%	46.79%	54.07%
Coverage	73.61%	80.04%	63.00%	72.22%

Table 4.12 *Results for L→R in Experiment P2*

Figure 4.6 shows the results of breaking down the errors in L→R according to the scheme in Table 4.8. The distribution is very similar to the one found in Experiment P1. The absence of LDOCE senses from WN15 still accounts for many of the errors, and the analysis above apparently suggests that the highly polysemous words account for most of the discrepancy between the resources.

In short, we have further tested our sense-mapping algorithm in this experiment, with results like those for Experiment P1. In addition, we discovered that LDOCE and WN15 have similar number of senses lacking in each other, while the senses

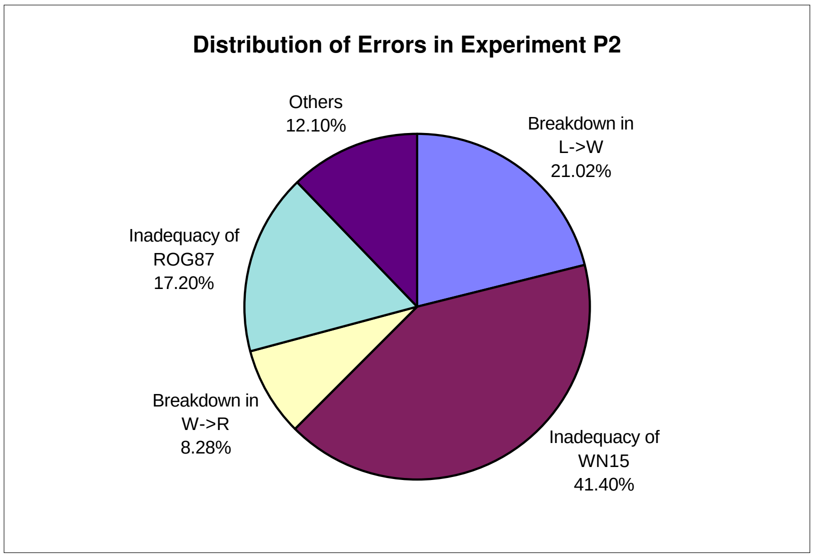


Figure 4.6 *Distribution of Errors in Experiment P2*

they have in common are similarly grained. Meanwhile, although WN15 and ROG87 have senses exclusive to themselves individually, mapping between them was more successful and senses were often more fine-grained in WN15 than in ROG87. So in the next pilot experiment (P3), we will explore the effect of replacing WN15 with another resource, which is similar to WN15 structurally but more closely related to LDOCE, as mediator. Before that, we will discuss the potential use of our results, regarding both the mapped senses and the sense difference between the resources.

Application of Results

Our study on combining resources was motivated from the need for enriched lexical information for WSD. We are therefore concerned with how the sense-mapping results could be used in the task. Two ways are suggested below, the first for obtaining a fuller set of senses for words, and the second for complementing two classifications of words with each other.

1. *Sense set expansion*

   We noted that there were senses which could not be mapped because they were only present in one resource. It is therefore possible to obtain a set of more representative senses for a word by conflating the "mappable" (and presumably correctly mapped) senses and adding to them the unattached senses from individual resources. For example, "note" has 10 senses in LDOCE and 9 in WN15, listed in Tables 4.13 and 4.14 respectively. The L→W mappings are shown in Figure 4.7, where a correct match is indicated by a bullet (•), an incorrect match by a cross (×), the expected target by a hollow bullet (∘), and a forced error by an asterisk (∗).

Sense	LDOCE Definition
D_1	musical sound, usu. of a particular length and PITCH
D_2	a written sign for any of these sounds
D_3	a quality of voice
D_4	any quality; ELEMENT
D_5	a record or reminder in writing
D_6	a remark added to a piece of writing and placed outside the main part of the writing (as at the side or bottom of a page, or at the end)
D_7	a short usu. informal letter
D_8	a formal letter between governments
D_9	a piece of paper money
D_{10}	PROMISSORY NOTE

Table 4.13 *LDOCE Definitions for "note"*

Synset	WN15 Synsets
S_1	eminence, distinction, preeminence, note
S_2	note, promissory note, note of hand
S_3	bill, note, government note, bank bill, banker's bill, bank note, banknote, Federal Reserve note, greenback
S_4	note (tone of voice)
S_5	note, musical note, tone
S_6	note, annotation, notation
S_7	note, short letter, line
S_8	note (written record)
S_9	note (emotional quality)

Table 4.14 *WN15 Synsets for "note"*

	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9
D_1					$\bullet$				
D_2					$\bullet$				
D_3				$\circ$					$\times$
D_4									$*$
D_5								$\bullet$	
D_6						$\bullet$			
D_7							$\bullet$		
D_8							$\bullet$		
D_9			$\bullet$						
D_{10}		$\bullet$							

Figure 4.7 *Results of L→W for "note"*

From Figure 4.7, we see that S_1 is an example of an unattached sense of "note" in WN15. S_9 in WN15 and D_4 in LDOCE should really be unattached too, although they were inappropriately mapped in one way or another. But ideally, with all mapping correct, we could obtain the following expanded set of senses for "note":

1. promissory note	6. letter
2. bank note	7. written record
3. tone of voice	8. eminence
4. musical note	9. emotional equality
5. annotation	10. any quality

where 1 to 7 are the mapped and conflated senses, 8 and 9 are extra from WN15, and 10 is extra from LDOCE.

Such an addition of extra senses from individual resources to the mapped senses should not affect the granularity of the latter. If the extra senses were closely related to the mapped senses, they would have been mapped already. On the contrary, we could adjust the granularity of the mapped senses, based on the mapping results and assisted by the symbolic matrix in Figure 4.7. Reading the matrix row-wise shows if a source sense is acceptably mapped to more than one target sense, and thus how vaguely the source sense is defined. Reading the matrix column-wise reveals how many source senses are mapped to the same target sense, and thus how fine-grained a word is in the source resource. In the "note" example above, D_1 and D_2 are both mapped to S_5 (musical note), whereas D_7 and D_8 are both mapped to S_7 (letter). Thus LDOCE has more fine-grained senses than WN15 in these cases. So if granularity is a concern in particular applications, one can treat D_1 and D_2 (and D_7 and D_8 alike) as one sense or as two senses accordingly.[10]

2. *Blending classification systems*

We found that the judgement for appropriateness of the W→R mappings was difficult when (i) the expected target and the mapped target for a WN15 sense were not identical but were nevertheless close neighbours in the ROG87 hierarchy, and (ii) more than one target was seemingly acceptable depending on which aspect of the sense was in focus.

This is to some extent a consequence of the difference in the classificatory structure of the two resources. WN15 has sets of synonyms as nodes and there is a strict IS-A relationship governing the hierarchy; whereas ROG87 may also have synonym or near-synonym groups at its lowest level, but the higher level classes contain broader conceptual relationships. That is perhaps why, for instance, we expect "field" ("a region where a battle is being fought") in WN15 to be mapped to the class "region" in ROG87, but we are unwilling to consider it wrong when the sense turns out to be mapped to "arena". Although it makes the mapping a little fuzzy, we can nevertheless capitalise on such a difference in

[10] While we suggest using the mapping results as a guide to adjust sense granularity, other ways have been suggested elsewhere, e.g. Sparck Jones' (1986) merging of "rows" by the degree of overlapping.

the classification between resources to fulfill our original goal of obtaining richer lexical knowledge for WSD via the integration of various types of resource.

Consider the following example:

> He sang a song last night. The notes were too high for a bass.

The musical sense common to the underlined words are self-suggestive enough, but these senses cannot be easily related via the WN15 hierarchy, as seen from their hypernym chains:

```
song -- (a short musical composition ...)
  => musical composition, opus, ...
    => music
      => art, fine art
        => creation
          => artifact, artefact
            => object, inanimate object, ...
              => entity

note, musical note, tone
  => musical notation
    => notation, notational system
      => writing, symbolic representation
        => written communication, ...
          => communication
            => social relation
              => relation
                => abstraction

bass, basso -- (an adult male singer ...)
  => singer, vocalist
    => musician, instrumentalist, player
      => performer, performing artist
        => entertainer
          => person, individual, ...
            => life form, organism, ...
              => entity
            => causal agent, cause, ...
              => entity
```

The nodes do not meet until the very top node "entity" (and for two of them only). However, our algorithm maps the corresponding senses of "song", "note", and "bass" to heads 412 (Music), 410 (Melody), and 413 (Musician) in ROG87 respectively. In this way, the relations among these senses become more apparent and this would possibly allow more effective disambiguation. We will return to this point later in this chapter.

4.5.3 Experiment P3

WN15 vs LLOCE

From the previous pilot experiments we see that WN15 is not perfectly efficacious as a mediator in the mapping process mainly because it differs too much from LDOCE in the senses they contain. However, this difference is not enough in itself to invalidate WN15 from the job, unless a better substitute can be found. So in this experiment, we replace WN15 with the Longman Lexicon of Contemporary English (LLOCE) and explore the effect of the latter on mapping senses between LDOCE and ROG87.

According to McArthur (1981), LLOCE "brings together words with related meanings and lists them in sets with definitions, examples, and illustrations ...". LLOCE is thus structurally similar to WN15, for they both possess dictionary and thesaurus properties. Nevertheless, LLOCE also differs from WN15 in the following respects. On the one hand, the meaning definitions in LLOCE are derived from those in LDOCE. So when there is a match, it can be easily and almost unequivocally located. On the other hand, the classification in LLOCE starts with 14 semantic fields, branching into smaller subsets of words. Although these semantic fields are less abstract than the ROG87 classes, the two classification are similar for having only weakly defined relations within and between classes. These characteristics thus bring LLOCE closer to both LDOCE and ROG87, suggesting that LLOCE might work better than WN15 as the mediator in the mapping process.

Since our sense-mapping algorithm is portable, only slight adjustments are needed for use with LLOCE. Comparison between LDOCE and LLOCE definitions would be more straightforward since a simple overlapping procedure will suffice (for the reason mentioned above). For the mapping of LLOCE to ROG87, words in the same LLOCE category (i.e. the small subsets under the 14 semantic fields) as the sense in question will be compared to the potential ROG87 paragraphs for the most overlaps. The algorithm was tested on the 36 nouns used in Experiment P2, irrespective of the polysemy groups.

Results and Discussion

The accuracies of L$\rightarrow$R through WN15 and LLOCE respectively are shown in Table 4.15. To be fair, we compare the results for the mapping of the LDOCE senses which appear in both WN15 and LLOCE (+LLOCE and the last column respectively). On average, only 56.33% of our LDOCE test senses were present in LLOCE. The Polysemy Factor $\mathcal{P}$ of LDOCE with respect to LLOCE is 0.62 and that of LLOCE with respect to ROG87 is 0.97. This is very interesting because the situation is almost reverse with WN15 (refer to Table 4.11). However, as seen in Table 4.15, it did not make much difference whether WN15 or LLOCE was used as mediator to map these senses to ROG87. Both cases led to about 35% overall success (out of all test senses). But that was all with LLOCE, whereas an extra 20% correct mappings could be obtained with WN15 (–LLOCE column). We can therefore see that although LLOCE may be more similar than WN15 to LDOCE in terms of contents and to ROG87 in terms of word classification, LLOCE does not turn out to be a better mediator between LDOCE and ROG87. Moreover, it is to a certain extent handicapped by its collection size.

The above results thus demonstrate the robustness of our algorithm. At the same time, they defend WN15 for its suitability as the mediator in addition to the fact that WN15 is freely available and is less biased than LLOCE to any particular resource.

Performance Measures	L→R thro WN15		thro LLOCE
	+LLOCE	–LLOCE	
Accurately Mapped (A)	34.47%	19.59%	35.30%
Polysemy Factor ($\mathcal{P}$)	0.76	–	0.73

Table 4.15 *Results for Experiment P3*

In this whole section, we have demonstrated the feasibility and robustness of our structurally-based sense-mapping algorithm via pilot experiments P1-P3. We have tested the algorithm on particular instances of various types of resource, namely LDOCE for dictionaries, ROG87 for thesauri, as well as WN15 and LLOCE for mediators. The mapping of LDOCE to WN15 compares favourably to Knight and Luk's (1994) results, despite the simpler procedures in our algorithm. The mapping of LDOCE to ROG87 via WN15 seems slightly worse than McHale and Crowter's (1994) 63% success in their direct mapping of LDOCE to RIT. However, their test samples contain different POSs, but ours are nouns only. Also, it is not clear whether their 100 LDOCE senses include target-less senses. If they do not, our results would then be much better since our precision was over 70%.

4.6 Full-Scale Linkage of Noun Senses

Given the results from the pilot experiments, and considering the availability of resources, it seems sensible for us to capitalise on the direct mapping of W→R to start up the Integrated Lexical Resource (ILR), part of our WSD model in Chapter 3, for use in the rest of our project. In this section, we will therefore describe a full-scale linkage of noun senses between WordNet and Roget's Thesaurus.

WordNet 1.6 (WN16) was released by Princeton sometime after our pilot experiments, so this new version instead of WN15 was used for the full-scale linkage. The Prolog version of WN16 has a total of 94,474 distinct nouns[11], which amount to 116,364 senses altogether.

There was more complication in using Roget's Thesaurus since we only had the machine-readable version of the 1911 Crowell Roget's Thesaurus (ROG11), but we preferred the 1987 Penguin Roget's Thesaurus (ROG87) for its bigger and more up-to-date collection of words. We therefore did some pre-processing to obtain information from both versions to form a joint version which we collectively call ROGET. This pre-processing stage will be described in Section 4.6.1. (See Chapter 3 for description

[11]In making the word estimates, capital letters are disregarded. So "March", "march", and "MArch" will be counted as different senses of the same word, just as they are displayed from the online interface.

of individual resources.) Then we will discuss the integration of WN16 and ROGET in Section 4.6.2, and envisage the potential application of the resultant ILR in Section 4.6.3.

4.6.1 Pre-processing for ROGET

To start with, *all* the nouns plus their head numbers and paragraph identifiers in the index of ROG87 were manually input as plain text, which was later converted to a Prolog-format database. However, as pointed out by the editor of ROG87:

> "*References to the same Head* are not usually repeated under different grammatical forms of the same word *[in the index].*" (Kirkpatrick, 1987: p.xiii)

So the index might under-represent the information in the thesaurus. For a shortcut remedy, we supplemented the ROG87 index with ROG11, which was old but complete and machine readable. A rough alignment of the head numbers and head names was first made between the two versions. A few heads in ROG11 were consequently left out as no straightforward correspondence in ROG87 could be found. The supplementation was done with the simple algorithm in Figure 4.8.

```
FOR every Noun-OldHead entry in ROG11
    Find OldHead-NewHead alignment
    IF Noun-NewHead NOT in ROG87 index
        Add Noun-NewHead to database
NEXT
```

Figure 4.8 *Simple Algorithm for Supplementing ROG87 with ROG11*

The ROG87 index itself contains 20,024 distinct nouns and 47,416 noun senses. The joint version, ROGET, has 35,718 distinct nouns and 75,013 senses. Sample entries indexed by word and by head are shown in Figures 4.9(a) and (b) respectively. The increase in words and senses in ROGET is a result of (i) the disclosure of entries in ROG87 which were missing in the index, and (ii) the addition of entries which were only in ROG11. We will illustrate these cases with some examples below.

Example 1 (Entries hidden in ROG87 recovered)

"Annotation" was not found in the ROG87 index, while the related verb "annotate" was indexed with heads 520 (Interpretation), 547 (Indication) and

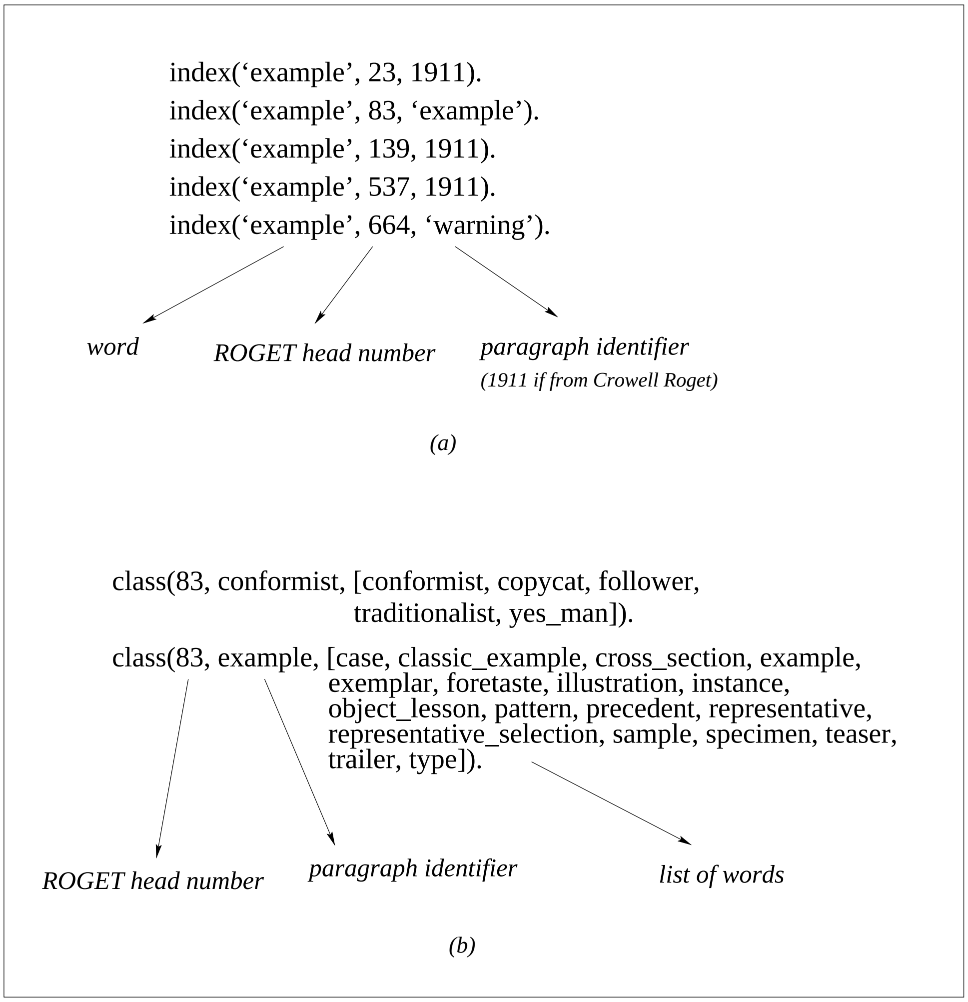


Figure 4.9 *Sample Entries in ROGET Indexed by (a) Word, and (b) Head*

591 (Dissertation). After supplementation, "annotation" was put in heads 520 and 547 in ROGET.[12]

Example 2 (Additional entries from ROG11)

"Anamorphosis" was apparently not found either in the index or the text of ROG87, but was added to heads 246 (Distortion), 440 (Dim-sightedness), and 552 (Misrepresentation) in ROGET after supplementation. Post-checking confirmed that these heads in ROG87 did not contain the word originally.

[12] Checking the complete text in ROG87, "annotation" is really in 520, but is only cross-referenced via "interpretation" in 547.

4.6.2 Integrating WN16 and ROGET

Practical Adjustments

The linkage between WN16 and ROGET was done with the algorithm (steps 2, 4, and 5) described in Section 4.3.2. Some adjustments were made to the algorithm. First, the hypernym chain from WN16 (i.e. $Hyp(S_j)$ in step 2) was limited to 2 levels from the target synset up the hierarchy (instead of a complete chain to the top node). This was because we observed that high level nodes were often too general to be helpful. Second, heads (with all paragraphs in them) instead of individual paragraphs in ROGET were taken as basic units, since we had actually disturbed the paragraph structure when we supplemented ROG87 with ROG11. In addition, the resultant mappings must have scores greater than 1, which means there must be at least one other word in common apart from the target word itself, otherwise the link would be left void. Ties in scores were preserved, allowing multiple links from the source sense to the target senses.

Compatibility of WN16 and ROGET

The full-scale mapping of nouns between WN16 and ROGET suggests the compatibility of the two resources as shown in Figure 4.10. The shaded area (including the netted area) represents the senses in WN16 of words which appear in both resources, and the netted area represents the senses among them which were mapped.

We can see that while the two resources differ considerably in collection size, about 90% of the nouns common to both resources could be linked in some way. The difference in size, however, was observed to arise from the peculiar contents of individual resources. For instance, WN16 contains a lot of specialised words (e.g. "genus Amphioxus", "genus Salpa", "genus Doliolum", "order Coniferales", "order Taxales", "order Arales", etc.) which are absent from ROGET. On the other hand, ROGET contains many multiple-word nominals (e.g. "amateur dramatics", "ancient monument", etc.) which are not found in WN16. In other words, although the number of linked words is just about half of the ROGET collection and only about one-fifth of WN16, these shared words are nevertheless among the commonly used words, unlike the above peculiar examples. Therefore we believe that the results from the linkage would form a usefully enriched resource for WSD, despite its limited size.

Sample Inspection and Further Observations

Although ideally we should make sure the mappings are all correct before putting the integrated resource to use, unfortunately full-scale checking is impractical given the huge amount of data. However, we could exploit the results obtained from the pilot tests (see Section 4.5) to estimate the success of the full-scale mapping. Since the samples in the pilot tests are among the most common words, we believe that they are representative of most words in the rest of the lexicon as far as the mapping effectiveness is concerned. Therefore, we do not expect much difference in results when the mapping is scaled up. It is thus reasonable to expect some 75% success rate from the full-scale linkage.

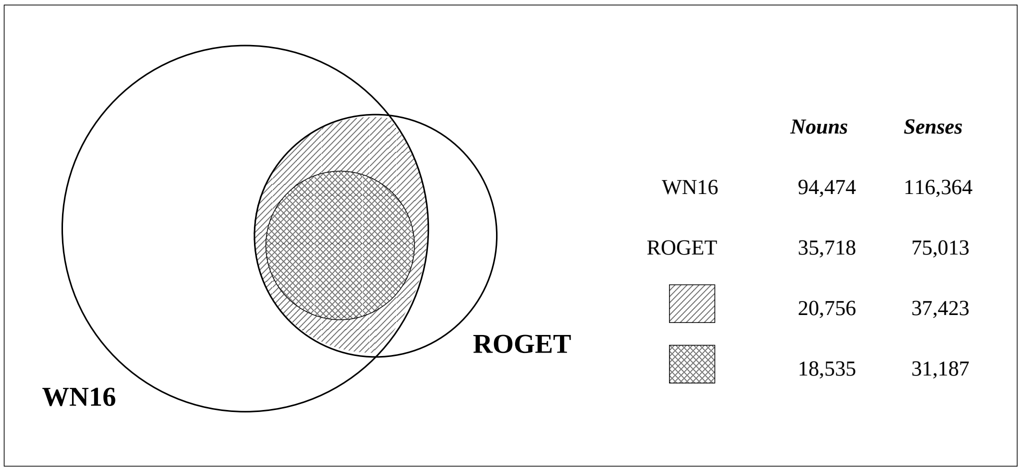


Figure 4.10 *Compatibility of WN16 and ROGET*

To investigate this claim further, we inspected the mappings for a random sample of nouns. These 19 nouns, listed in Table 4.16, were taken from a randomly selected excerpt from the Brown Corpus. The excerpt, the polysemy of the nouns with respect to WN16 and ROGET, and the judgement of appropriateness for the corresponding mappings, are recorded in Appendix B.

concert	point	rhythm	man
score	joust	wit	baton
fate	tune	kind	technique
time	misfortune	performance	style
thrust	texture	orchestra	

Table 4.16 *Nouns for Inspection of Full-Scale Mapping*

The analysis of the mappings for the samples in Table 4.16 corroborates most of the findings from the pilot tests. For instance, LO polysemous nouns (i.e. nouns with five or fewer WN16 senses) are most prevalent among the samples. The mapping accuracy is about 70% on average, which is close to our expectation. Also, false alarms (14% F + 9% Ub) outnumber misses (5% I + 2% Ua) by about three times, illustrating again that the major source of errors is from the difference in content between resources.

In addition, the current analysis uncovers further interesting differences between WN16 and ROGET. For instance, the construction of WN16 was to a certain extent corpus-based, and it contains word senses which were derived from actual usages of words in the Brown Corpus. As a consequence, there may be some non-central (and perhaps mildly technical) senses in WN16 which one might not find in an ordinary resource, e.g. the “fourth dimension” sense of “time” resulted in a forced error.

Apart from the origin of senses, WN16 and ROGET also seem to differ in the aspects of senses they emphasise. The latter difference may account for what we found in Experiment P2, that ROG87 correlated weakly with other resources in terms of the number of senses for words. Such a disproportion of senses across resources is also obvious in the current analysis. For example, "concert" has only one sense in WN16, namely "a performance of music". But it is under six heads in ROGET: 24 (Agreement), 181 (Concurrence), 410 (Melody), 412 (Music), 414 (Musical instruments), and 706 (Cooperation) – everything but "performance"! This variation of focus thus leads to unexpected mappings and demands some flexibility in our judgement of appropriateness. For example, given that "concert" is not under "performance" in ROGET, we might need to evaluate the mapping with respect to other properties of the sense. In this case, the next relevant property is apparently related to "music", and we will be at the crossroads again since there are three alternatives (heads 410, 412, and 414). However, in practice we will treat all of them as acceptable targets.

In this section we have performed a large-scale automatic sense linking between WN16 and ROGET to form a combined resource. About 18,000 nouns and over 30,000 senses were mapped. We have shown that it is reasonable to expect about 70% of the resultant mappings to be appropriate. This figure also suggests how reliable the mappings will be for subsequent use. In the next section, we will discuss the potential use of them.

4.6.3 The Integrated Lexical Resource

As mentioned at the beginning of this chapter, a single resource almost never contains all types of lexical information which WSD typically requires. This motivates us to combine various (types of) lexical resources so that different types of lexical information, which originally reside in different resources, could be used simultaneously and compatibly. We will now sketch how our ILR, despite its imperfection, could achieve this purpose in principle.

It happens that both component resources in our ILR (WN16 and ROGET) have a hierarchical classification of senses, although the classificatory criteria are quite different. With the ILR, the two classifications can be used together compatibly. We can thus take two senses as defined WN16 but compare them with respect to the ROGET classification, as in the example below.

Figure 4.11 shows the symbolic matrix for the mapping of "performance" from WN16 to ROGET. The leftmost column shows the WN16 senses while the top row shows the ROGET heads of "performance". On the one hand, we see that the last three WN16 senses of "performance" are closely related meanings in the ROGET classification by being under the same ROGET head. However, these senses are distinguished in the WN16 classification and in fact, they relate to head 676 in ROGET via different groups of words (see Appendix B). On the other hand, while senses 2 to 5 in WN16 are similar in the sense that they all descend from the top node of "human action", sense 2 goes to a different ROGET head from the others. Hence senses which are close in one classificatory system may be distant in another, depending on what semantic relations we are considering. More importantly, WSD not only requires these different types of semantic relation as disambiguation evidence, but also requires them to be applied appropriately (as we will explore in subsequent chapters).

We can adopt a quantitative way of doing such cross-resource similarity comparison, as suggested below. We first define $M(s)$ as "the set of ROGET heads that sense

	157 (Effect)	164 (Production)	413 (Musician)	551 (Representation)	594 (Drama)	676 (Action)	725 (Completion)	768 (Observance)	876 (Celebration)	917 (Duty)	988 (Ritual)
1. dramatic/musical entertainment					•						
2. act of presenting entertainment				•	•				•		
3. act of doing something						•					
4. process of functioning						•					
5. recognized accomplishment						•					

Figure 4.11 *Symbolic Matrix for "performance" in W→R*

s is mapped to", and $S(h)$ as "the ROGET section (one level above heads) subsuming head h". We can then define a function like this:

$$Sim_r(s_1, s_2) = \begin{cases} 2 & \text{if } M(s_1) \cap M(s_2) \neq \emptyset \\ 1 & \text{if } M(s_1) \cap M(s_2) = \emptyset \\ & \text{and } h_1 \in M(s_1), h_2 \in M(s_2), S(h_1) = S(h_2) \\ 0 & \text{otherwise} \end{cases}$$

Since the ROGET hierarchy is a relatively flat one, a step function should be appropriate. The actual values of the function could be determined empirically, as long as they correlate with the similarity between two senses with respect to the ROGET hierarchy. Applying the above function to "performance", we get:

$$\begin{aligned} Sim_r(performance/1, performance/2) &= 2 \\ Sim_r(performance/1, performance/4) &= 0 \end{aligned}$$

The ability to quantify such comparisons means that we can combine different resources into a single sense inventory, but preserve the structures of individual resources to allow the flexible use of different types of information for WSD. This is the most significant property of the ILR.

In actual WSD, the above comparison is normally between senses of different words. However, comparing senses of the same word, our example serves to illustrate another potential use of the mapping results, in terms of WSD evaluation. Resnik and Yarowsky (1997) suggested that wrong answers from WSD systems should be given different penalties depending on the closeness of the wrong answer to the expected one. Our sense-mapping matrix can therefore serve as a kind of confusion matrix, and the similarity function above offers an alternative way for determining the closeness of senses. In the case of "performance", a system which fails to distinguish between senses 1 and 2 should therefore receive less penalty than one which fails between senses 1 and 4.

4.7 Summary

In this chapter we have explored an extrinsic parameter of WSD, namely *Lexical Resource*. We start with the general assumption that a single lexical resource rarely contains all types of lexical information which WSD typically requires, and suggest enriching lexical knowledge by combining different existing lexical resources. A structurally-based, two-part algorithm for automatically combining a conventional dictionary and a thesaurus via a mediator was designed. Pilot experiments were done, using LDOCE, ROG87, and WN15 for the above types of resources respectively. In general, the direct mappings (dictionary to mediator and mediator to thesaurus) achieved accuracy of above 70%, although the indirect mapping (dictionary to thesaurus) was slightly inferior. However, our method was shown to be robust as the major source of errors was identified as the difference in the senses contained in individual resources.

Given the plausibility of the method, we have carried out a full-scale mapping of noun senses between WN16 and ROGET to form the Integrated Lexical Resource for use in the rest of this project. More than 18,000 nouns and 30,000 senses were mapped, with success rate expected at about 70%. How the ILR can be used in principle was discussed, although the actual effect remains to be seen in subsequent chapters.

5

Extrinsic Parameter II: Contextual Information

5.1 Introduction

In this chapter, we will study another extrinsic parameter of WSD: *Contextual Information*, or *context* for short. This parameter covers any information which is available from the triggering context (TC) as potential evidence for disambiguation.

It seems reasonable to think of context as a necessary condition for WSD. According to Hirst (1987), word sense selection can be treated as a problem of context recognition. Ide and Veronis (1998) also suggested that "context is the only means to identify the meaning of a polysemous word". In fact, words need not be disambiguated if they are not used in context. However, despite its importance, context in WSD is seldom given any formal definition. Usually it is taken loosely to refer to some neighbouring text surrounding a target word. Tabossi (1989) provided a relatively elaborate interpretation of context in psycholinguistic studies. She used context in the broad sense to refer to the extra-linguistic situation where conversation takes place; and in the restricted sense, to refer to the orthographical, phonological, syntactic, and semantic aspects of discourse within which a linguistic expression occurs. As mentioned in Chapter 2, the effect of context is one of the gaps in WSD studies requiring more serious attention. Few studies have explicitly investigated the role of context in WSD, although WSD studies in general may throw light on the subject implicitly. The lack of a well-defined notion of context is probably responsible for the situation to a certain extent.

We do not intend to give any formal definition of context, but will try to formulate a more systematic investigation of it. We will adhere to Tabossi's restricted sense of context and separate it into two aspects: the *size* and the *realisation* of context. By our definition, size refers to the amount of neighbouring text surrounding a target word from which disambiguation evidence is gathered, and realisation refers both to the type of lexical information drawn from the TC and its computational extraction and application.

We will discuss these two aspects of context and their related studies in Section 5.2. Then from Section 5.3 to Section 5.5, we will confine the realisation to one type of semantic relation, and experiment with various context sizes. In particular, we will design a recursive filtering algorithm in Section 5.4 which bypasses combinatorial explosion and thus allows full-text context to be tested. In Section 5.6, we will fix the context size, and study realisation via various types of semantic relation. An example illustrating our findings will be given in Section 5.7, followed by a summary of this chapter in Section 5.8.

5.2 Two Aspects of Context

As said, in most cases, context for WSD is only informally defined. For instance, Ide and Veronis (1998) distinguished three types of context, namely "microcontext", "topical context" and "domain". On the one hand, they defined "microcontext" as "some small window of words surrounding a word occurrence in a text or discourse, from a few words of context to the entire sentence in which the target word appears". On the other hand, they defined "topical context" as "substantive words that co-occur with a given sense of a word, usually within a window of several sentences". We can see that such a distinction of context has actually blended two considerations into one while they should really be separate. These considerations are (i) how much surrounding text is taken into account (size), and (ii) what type of information is evaluated and how (realisation). Ide and Veronis' definitions thus seem so minimal that we do not know, for example, what the type of context defined by "co-occurring words taken from a small window enclosing the target word" actually is. Mixing the two aspects of context also causes confusion in the interpretation of results, as will be seen below when we discuss Leacock et al's (1998) study. Therefore, to allow properly controlled comparison between different variations of context, we will divide context into two aspects: size and realisation, as discussed in Sections 5.2.1 and 5.2.2 respectively.

5.2.1 Size

Traditionally the size of context is often described as "local" (for short context) or "global" (for long context). However, although a ±3-word or one-sentence window may be typical for local context (Choueka and Lusignan, 1985; Leacock et al., 1998), and a window of ±50 words or more can be considered global context (Yarowsky, 1992), there is no unequivocal boundary between them. For example, a window of 300 words is already the upper limit for a 300-word abstract, but the same size would be relatively local for a book. It may not even cover an entire chapter. So the size of context should more appropriately be viewed in relation to the length and the type of the discourse in question. Moreover, whether context size should be measured in natural linguistic units (e.g. paragraphs, sections, etc.) or in some other way depends on how well one can identify meaningful units from the discourse. Since this is another debatable issue, we will follow the main practice of measuring context size in terms of window sizes.

One early study on context size was done by Choueka and Lusignan (1985). They asked informants to use one neighbouring word (1-context), or two (2-context) if one was insufficient, to disambiguate 31 ambiguous words (about 3,000 occurrences altogether).[1] It was reported that 77% of the test samples were successfully disambiguated with 1-context, and 91% with 2-context. Choueka and Lusignan thus claimed that "short contexts are in general sufficient for correct disambiguation", and their results were often cited in later studies as support for doing WSD with short contexts.

However, there are two points about the study which are noteworthy. First, the test was done with French words. Second, their test samples included POS ambiguity. Moreover, an analysis of their samples as in Table 5.1 shows that only about 16% of

[1] The neighbouring word(s) can come from pre-, post-, or symmetric context.

the samples correspond to word sense ambiguity within the same POS. Consequently, and reinforced by their remark that "auxiliaries and determiners are good clues", the results may more appropriately be treated as evidence for the sufficiency of short context for resolving categorial ambiguities. The support for the sufficiency of short context for WSD at large should therefore be much less strong than has generally been perceived.

Semantics	Morphology		
	Same	*Different*	
Similar	(no example given) [2]	DRESS/v (*put clothes on*) DRESS/n (*clothes*)	48%
Different	BAT/n (*wooden tool*) BAT/n (*animal*)	FALL/v (*act of falling*) FALL/n (*season*)	52%
	16%	84%	

Table 5.1 *Analysis of Sample Types in Choueka and Lusignan (1985)*

More recently, Leacock et al. (1998) studied the individual and combined effect of "topical" and "local" context on WSD. By their definition, "topical context" consists of all open-class words found in the same sentence as the target word, and the preceding sentence. "Local context", on the other hand, consists of open-class words within ± 3 positions from the target, close-class words within ± 2 positions, and POS tags within ± 2 positions. They used a Bayesian classifier to assign senses to target word instances. However, they only tested for three words, namely "serve", "line", and "hard" for verbs, nouns, and adjectives respectively. Moreover, they only considered senses for these words which had enough training samples (4/13 for "serve", 3/11 for "hard", and 6/29 for "line", with senses from WordNet). Their disambiguation results are summarised below:

	Local	Topical	Combined
serve (v.)	78%	68%	83%
line (n.)	67%	78%	84%
hard (adj.)	83%	60%	83%

Thus they found that "topical context" alone worked better for nouns, whereas "local context" alone worked better for verbs and adjectives. Combining both types of context was beneficial to verbs and nouns only. Given the effectiveness of "topical context" for nouns, it was also suggested that nouns might be more "topical" (which appears to mean how strongly a word relates to a particular subject area) than words of other POSs. However, this point is overgeneralising because senses of the same

[2] No example was given for this category in Choueka and Lusignan (1985), but we believe that it possibly corresponds to closely related senses in contrast to homonyms.

word can vary in how "topical" they are, as the following was also observed in the study:

> "When it happens that a nontopical sense accounts for a large proportion of occurrences, then adding topical context to local will have little benefit and may even reduce accuracy." (Leacock et al., 1998: p.152)

Towell and Voorhees (1998) did a similar study with a neural network approach, with similar results. Although such results might suggest that context has different effects for ambiguous words of different POSs, the tests were so severely limited by the small number of words and senses involved that it is impossible to generalise from them.

Yet another serious problem is that different aspects of context are not distinguished when interpreting the results. On the one hand, it is curious to contrast "topical context" with "local context" since the former seems to stress the information sought but the latter the context size, although "topical context" may implicitly mean a longer context than "local context". On the other hand, the comparison between "topical" and "local" context is in fact unfair since there were three types of information considered in "local context", but only one in "topical context". Hence one can only see the interaction of the variables but not their individual main effects from the results, and no valid cause-effect relationship can be inferred. For a controlled comparison, one should use the same type of information but from different amounts of neighbouring text, or different types of information from the same amount of neighbouring text. At least, this is what one should do in principle, although it is not always feasible in practice. For instance, one may like to compare trigram and 100-gram for local and global context respectively, but 100-gram statistics can hardly be satisfactorily obtained. Similarly, information like discourse focus would not be available with too short a context. Even so, one should take every care to control the experimental conditions to enable meaningful interpretation of results. We believe that separating context into size and realisation is a sensible step in this direction.

Thus concerning size alone, all of the above studies were not extensive enough. They only tested two sizes of no more than two consecutive sentences. Other studies have suggested various optimal context sizes associated with different types of disambiguation information (Yarowsky, 1992; Agirre and Rigau, 1996; Pedersen, 2000), but the range of sizes tested was still small and context size was not their focus.

5.2.2 Realisation

By context realisation, we refer to the type of information obtained from the TC and how it is related to the conventionalised context (CC) of the target word. (The CC is represented in the lexical resource.) We have thoroughly surveyed in Chapter 2 the various types of information used in WSD. Here is a summarised list:

1. Pragmatic Information
    - Inference: e.g. Wilks (1975a).
    - Domain knowledge: e.g. Hirst (1987).
    - Subject area: e.g. Stevenson and Wilks (1999).

2. Semantic Information

   - Broad semantic relations: e.g. Cowie et al. (1992), Yarowsky (1992).
   - Narrow semantic relations: e.g. Agirre and Rigau (1996).
   - Positional semantic relations: e.g. Yarowsky (1993), Resnik (1997).

3. Other Information

   - Syntactic structure: e.g. Dagan and Itai (1994), Lin (1997).
   - POS tags: e.g. Wilks and Stevenson (1996).
   - Morphology: e.g. Bruce and Wiebe (1994a), Ng and Lee (1996).

Semantic information is apparently the core type of information used in WSD. Recent studies tend to use more than one type of information simultaneously. Varied results are reported not only from studies using different types of information, but also from those using similar ones. As we pointed out in Chapter 2, this is a consequence of many other intervening aspects such as sense distinctions, test materials, and so on. In fact, from a practical point of view, one might only be concerned with the overall effectiveness. However, for a systematic investigation of the effect of context on WSD, it is important to compare the effect of different types of information in a more controlled setting. This would also provide useful hints on how different types of information could be optimally combined for WSD. Therefore in the next few sections, we will study the size and realisation of context separately, under controlled experimental conditions so as to obtain more meaningful results.

5.3 Context Size (Part One)

In this section, we will present a preliminary study on the effect of context size with respect to one type of disambiguation information. We will first describe briefly how we define context size and control the context realisation in Sections 5.3.1 and 5.3.2 respectively. Then we will report an experiment in Section 5.3.3, and suggest further investigation in Section 5.3.4.

5.3.1 Defining Context Size

As mentioned in Section 5.2.1, whether context size should be measured in terms of natural linguistic units depends on how well one can identify meaningful discourse units with them. For relatively local context, one might still reliably base on sentential boundary (e.g. 1-sentence or 2-sentence contexts). However, for longer contexts, natural linguistic units may become a less satisfactory measure. For example, paragraphs might differ considerably in length. So even if both correspond to a discourse segment, the actual amount of evidence for WSD may be less in the shorter paragraph. In the absence of obvious ways to compare the amount of information from two chunks of text of different lengths, it seems reasonable to assume that the amount of information available is more or less proportional to the text length.

In practice, context size is often defined as a *window* of $\pm n$ words from the target word, though the precise definition of window size may vary from study to study. For instance, a window of size n in Agirre and Rigau (1996) means $\pm n$ nouns from the target word, regardless of how far it extends over the neighbouring text. In another case, Yarowsky (1992) used windows of ± 50 words literally, but compared the target word with not only nouns but also words of other POSs (e.g. verbs) in the window. Thus the former exerts more control over the number of nouns while the latter exerts more control over the actual window size, but both ways are valid.

We will adopt the simple and straightforward definition of window size, i.e. $\pm n$ words means n words on both sides on a target word. However, since we only handle noun ambiguities and restrict the information to that available from nouns only, we also assume that the amount of information available from nouns alone is proportional to window size.

5.3.2 Controlling Context Realisation

As said, we need to fix at a particular realisation of context and study the effect of size with respect to it. Consequently, the amount of the information involved must potentially vary with context size and the CC in our resources should be available in that form of information. We therefore use a narrow semantic relation, namely the IS-A relation as exemplified in WordNet.[3] Although there are different ways for measuring semantic closeness/distance with respect to a hierarchical classification (Richardson et al., 1994; Agirre and Rigau, 1996), we aim to make it simple so as to keep any confounding effect to a minimum. So in this section (and the next two), we will measure semantic similarity by simple taxonomic distance, with the metric suggested in Rada et al. (1989). Thus the similarity between two concepts x and y is given by

$$S(x, y) = \frac{1}{1 + Dist(x, y)}$$

where $Dist(x, y)$ is the minimum number of edges separating the nodes representing x and y.

5.3.3 Experiment WBA-1

Materials

In this experiment, 186 texts from the SEMCOR data were used for both testing and evaluation. Each noun occurrence (excluding proper nouns) was treated as a target word.

As described in Chapter 3, SEMCOR texts are categorised into different text types, as listed in Appendix A with sample texts. These text types and letter codes are repeated below for quick reference:

[3] For empirical reason, we limit the hierarchy to one level below the top nodes for comparison, in order to reduce the number of spurious closeness.

A	Press: Reportage	J	Learned and Scientific Writing
B	Press: Editorial	K	Fiction: General
C	Press: Reviews	L	Fiction: Mystery and Detective
D	Religion	M	Fiction: Science
E	Skills and Hobbies	N	Fiction: Adventure and Western
F	Popular Lore	P	Fiction: Romance and Love Story
G	Belles Lettres, Biography, etc.	R	Humour
H	Miscellaneous		

Each text is about 2,000 words in length, and contains about 300 noun occurrences, giving more than 60,000 test samples.

Method

To lessen the computational load posed by combinatorial explosion, we modified the standard window-based WSD process in the Viterbi manner. So instead of making an exhaustive comparison of every possible sense combination of the target word with all words in the context, each sense of the target word was compared to each sense of each noun in context independently, and the results from these independent comparisons were combined.[4] This *Window-Based Algorithm* (WBA) is shown in Figure 5.1. In the figure, $S(x, y)$ is as defined in Section 5.3.2.

FOR each noun occurrence w in the discourse

$\quad$ $Context = \{$nouns within $\pm n$ words of w (duplicates removed)$\}$

$\quad$ FOR each sense s_{wi} of w

$\quad\quad$ $Score(s_{wi}) = 0.0$

$\quad\quad$ FOR each noun $c \in Context$

$\quad\quad\quad$ $Score(s_{wi}) = Score(s_{wi}) + \max_j(S(s_{wi}, s_{cj}))$

$\quad\quad$ NEXT

$\quad$ NEXT

$\quad$ $Tag(w) = \arg\max_i(Score(s_{wi}))$

NEXT

Figure 5.1 *The Window-Based Algorithm (WBA)*

Test Conditions and Hypotheses

The WBA was tested with window sizes of ± 5, ± 20, ± 50, and ± 70 words. We use the shorthand notations *wb05*, *wb20*, *wb50*, and *wb70* for these test conditions

[4] Duplicates for both target and non-target nouns in the context were first removed.

respectively. A context of ± 5 words is relatively local. Contexts of ± 20 and ± 50 words cover the range of optimal window sizes in various past studies with different types of disambiguation information (Yarowsky, 1992; Agirre and Rigau, 1996). Meanwhile a context of ± 70 words would be relatively global compared to these sizes, although it is still far from the upper limit with respect to our test texts.

Our hypothesis is that disambiguation performance would improve as context size increases, since more context would presumably provide more evidence for disambiguation. However, the improvement may decrease gradually to a stable level.

Performance Measures

The principal measures of performance are *precision* and *recall*, which we defined for this experiment as follows:

$$precision = \frac{no.\ of\ correctly\ disambiguated\ test\ samples}{no.\ of\ disambiguated\ test\ samples}$$

$$recall = \frac{no.\ of\ correctly\ disambiguated\ test\ samples}{no.\ of\ test\ samples}$$

where "disambiguated test samples" refers to cases with one or more candidate senses, but fewer than the number of possible senses, having a non-zero maximum score.

Results

Table 5.2 shows the precision and recall at different window sizes with respect to individual text types. The maxima of both measures for each text type are shaded.

The best results among the various test conditions are about 36% and 35% on average for precision and recall respectively. While precision remains relatively consistent across different context sizes, it is evident that recall tends to increase with context size (at least up to ± 50 words), although the effect is not as dramatic as one would expect. Thus the results support, though not very strongly, the claim that WSD performance improves with increasing context size and levels off at some point. At *wb50*, text categories M (Fiction: Science), N (Fiction: Adventure and Western), and R (Humour) have relatively poorer performance (both precision and recall) than other text types. Also, the fictional texts (e.g. L, M, N, and P) apparently have better precision at very small context, but otherwise not much difference is observed among text types.

5.3.4 Effect of Context Size

The most striking point of the above experiment is that WSD performance does not seem to be significantly affected by context size. We can think of two possible reasons.

First, the range of context sizes tested may not be wide enough. As suggested earlier on, how global a context is should be considered in relation to the length of the discourse. Given that our test texts are about 2,000 words in length, it seems reasonable to conjecture that even a context size of ± 70 words might still be relatively

Cat	Precision (%)				Recall (%)			
	wb05	***wb20***	***wb50***	***wb70***	***wb05***	***wb20***	***wb50***	***wb70***
A	38.10	37.66	39.14	38.80	22.08	34.54	37.71	37.61
B	38.55	38.75	38.81	36.28	24.12	35.88	37.65	35.40
C	33.34	34.57	36.40	36.71	19.82	31.85	35.00	35.38
D	34.32	33.56	34.18	34.64	22.05	31.33	33.39	34.16
E	37.21	37.82	37.16	37.64	25.37	35.30	35.95	36.58
F	36.64	37.15	37.49	37.38	23.01	34.25	36.38	36.55
G	36.49	36.60	36.78	35.66	22.96	33.92	35.61	34.77
H	37.29	37.96	39.48	39.73	26.18	35.73	38.43	38.87
J	35.36	36.34	37.35	37.08	24.64	34.47	36.63	36.51
K	33.98	34.49	34.15	34.07	20.55	31.60	33.09	33.17
L	38.03	36.53	37.04	36.78	21.08	32.94	35.78	35.70
M	34.67	33.43	31.72	31.20	19.27	29.07	29.80	29.46
N	34.07	34.00	32.47	32.63	21.88	31.99	31.85	32.10
P	37.13	34.50	36.40	36.03	20.82	31.33	35.63	35.61
R	31.75	32.66	32.05	31.78	19.63	30.18	31.11	31.00
Ave	35.75	36.04	36.39	36.19	22.89	33.47	35.38	35.38

Table 5.2 *Results of Experiment WBA-1*

local. A coherent discourse may have ideas developed throughout a wide span of text, and related words are not necessarily near neighbours.[5] Consequently, we should consider testing even larger context sizes.

Second, the problem may lie in the realisation of context which we have controlled in the study. As observed from the results, for instance, many words which were obviously used in some particular senses in a text were not always correctly disambiguated (e.g. "space" was consistently used in the sense of "topological space" in text br-j18, but not all occurrences of "space" in the text were correctly disambiguated). Also, some words which were not related to the theme of the text in any obvious way were somehow correctly disambiguated (e.g. "idea", "hold", etc. are relatively neutral in a text about mathematics, but are nevertheless correctly disambiguated).

So we will further investigate these two possibilities in the following section, exploring the expansion of context first, and then the modification of the taxonomic distance metric (an element of context realisation).

5.4 Context Size (Part Two)

In the last section, we identified the range of context sizes tested and the measurement of taxonomic distance as potential limitations on the results obtained in Experiment WBA-1.

To test context expansion, the foremost difficulty we need to overcome is that of combinatorial explosion. The increase in processing time was already noticeable as context was enlarged to ±70 words in Experiment WBA-1. Thus it seems impractical to test even bigger context size in the same way. Cowie et al. (1992) used simulated annealing to optimise the WSD process, but their method was still confined to senten-

[5] See e.g. Morris and Hirst (1991) on lexical cohesion.

tial context. Hence we propose a recursive filtering algorithm, which expands context at lower cost but in a slightly restricted way. We will demonstrate the effectiveness of our method with the upper limit of context size from our test materials (i.e. about 2,000 words).

On the metric for measuring taxonomic distance, we may have implicitly made too strong assumptions about the WordNet hierarchy. For instance, our metric assumes that the distances between pairs of nodes are identical regardless of the depth of the nodes in the hierarchy. Agirre and Rigau (1996) counteracted this assumption by measuring similarity in terms of "conceptual density". We motivated our choice for a simple metric in Section 5.3.2. Nevertheless, we are still able to improve our metric by grounding our measure of similarity on localised nodes from which we can safely assume uniform distances between nodes. We will search for such a local scope empirically.

So we will introduce the recursive filtering algorithm for expanding context in Section 5.4.1, and a corresponding experiment in Section 5.4.2. Then we will discuss an empirical approach to pre-define a local scope for measuring taxonomic distance with respect to the WordNet hierarchy in Section 5.4.3, and see how the modified metric affects WSD performance in Section 5.4.4.

5.4.1 A Recursive Filtering Algorithm

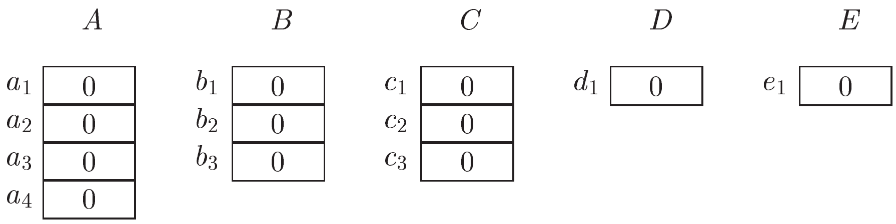
The idea of the recursive filtering algorithm is to gradually purge the irrelevant senses and leave only the relevant ones, within a finite number of processing cycles. Instead of examining all sense combinations exhaustively, the algorithm reduces the amount of processing by only selecting a different subset of senses from the whole context in each cycle to provide the disambiguating context for the rest. Hence context is made available progressively instead of in a burst, and in this way it is practical to use much larger context sizes because only a useful part of the context is exploited in each cycle. So ideally after each cycle, some polysemous words are disambiguated, and their correct senses provide further context to help disambiguate other words in subsequent cycles. On convergence, that is when no further improvement is found, we should end up with a set of correctly disambiguated senses.

However, there are certain practical concerns. First, the correctness of the intermediate results is not guaranteed. This can seriously affect the final results because the outcome of one cycle will influence the outcome of the next cycle. Second, words are not necessarily left with one single sense in the end because there may be ties, or the word is not disambiguated at all. So during processing, a history of the scoring should be kept to accumulate evidence for senses from different cycles. With such a history list, falsely preferred senses in earlier cycles can be overridden by more favourable candidates in later ones. Also, ties should be preserved so that in a hypothetical complete system they can be subjected to further refining disambiguation with other types of information.

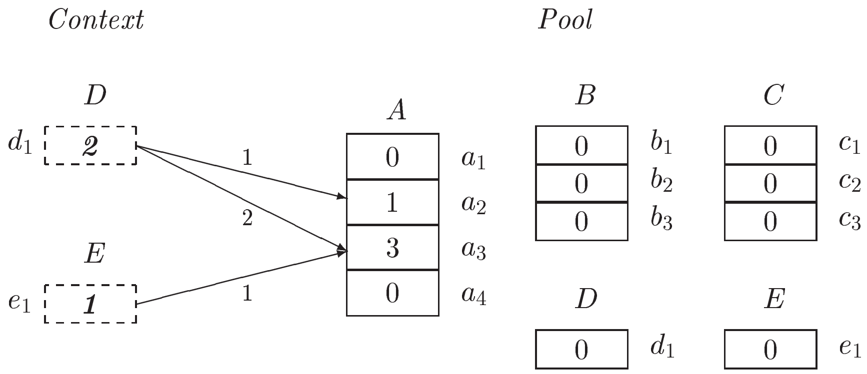
Like the unsupervised method in Yarowsky (1995), which needs a set of manually tagged seed collocations to start with, we also need a set of (correct) senses to get the ball rolling. Here the *monosemous* words in a text might be a good choice because they are already unambiguous and do not require further tagging. In addition, they offer the most economical processing since no matter how much context we consider, we always have n combinations to evaluate for a target word with n senses. Moreover, the SEMCOR data show that monosemous nouns can amount to as much as 30% of all nouns in a 2000-word text, and are sometimes domain-suggestive. Hence monosemous nouns should define good initial drivers for our recursive filtering algorithm. This *Monosemous Filtering Algorithm* (MFA) is shown in Figure 5.2.

We illustrate the algorithm with a hypothetical example below. The example contains five nouns: $A(4)$, $B(3)$, $C(3)$, $D(1)$, and $E(1)$, with the number of senses in brackets (so D and E are monosemous).

At initialisation, all senses have 0 for all types of score:



In the first round, $Context = \{d_1, e_1\}$, $Pool = \{a_1 \ldots a_4, b_1 \ldots b_3, c_1 \ldots c_3, d_1, e_1\}$, and suppose we set initial threshold $T = 2$. Every possible pair of senses from different words between $Context$ and $Pool$ are evaluated for the type of information interested.[6] Imagine d_1 and e_1 are found related to senses of A like this:



There is a distinction between *Accumulated Score* (AS, in solid frames) and the actual *Supportive Score* (SS, in dashed frames). An AS is kept for each sense in $Pool$, while an SS is kept for each sense in $Context$, computed respectively as shown in Figure 5.2. In this example, the AS for a_3 is 3 (2 from d_1 + 1 from e_1) and the SS for d_1 is 2 (maximum among 1 to a_2 and 2 to a_3).

In the second round, only a_3 has a score above the threshold T (i.e. ≥ 2), so it is filtered out as $Context$, and the comparison goes on:

[6]This is a measure to simulate the mutual inhibition among senses of the same word, as in the neural network approach.

Initialisation

Let $\mathcal{S}_n = \{s_{n1}, \ldots, s_{nj}\}$ be the set of possible senses for a word w_n

Let $\mathcal{W} = \{w_1, \ldots, w_k\}$ be the set of nouns in the discourse

Set *Pool* $\mathcal{P} = \{p \in \mathcal{S}_i \mid w_i \in \mathcal{W}\}$

Set *Context* $\mathcal{C} = \{c \in \mathcal{S}_i \mid w_i \in \mathcal{W},\ |\mathcal{S}_i| = 1\}$

Set *Filtered* $\mathcal{F} = \emptyset$, Threshold T, Interval t, Min Threshold min

Set Accumulated Scores $AS(\cdot) = 0.0$, Supportive Scores $SS(\cdot) = 0.0$

Recursion

WHILE $(T > min)$ {

FOR every sense $p \in \mathcal{P}$

$$AS(p) := AS(p) + \sum_{c \in \mathcal{C},\ c \neq p,\ \{c,p\} \not\subseteq \mathcal{S}_n} S(c, p)$$

NEXT

FOR every sense $c \in \mathcal{C}$

FOR every word $w_i \in \mathcal{W} \mid c \notin \mathcal{S}_i$

$$SS(c) := SS(c) + \max_j(S(c, s_{ij}))$$

NEXT

NEXT

$\mathcal{F} := \mathcal{F} \cup \mathcal{C}$

$\mathcal{C} := \{p \in \mathcal{P} \mid AS(p) \geq T,\ p \notin \mathcal{F}\}$

WHILE $((\mathcal{C} = \emptyset)$ AND $(T > min))$ {

$T := T - t$

$\mathcal{C} := \{p \in \mathcal{P} \mid AS(p) \geq T,\ p \notin \mathcal{F}\}$

}

}

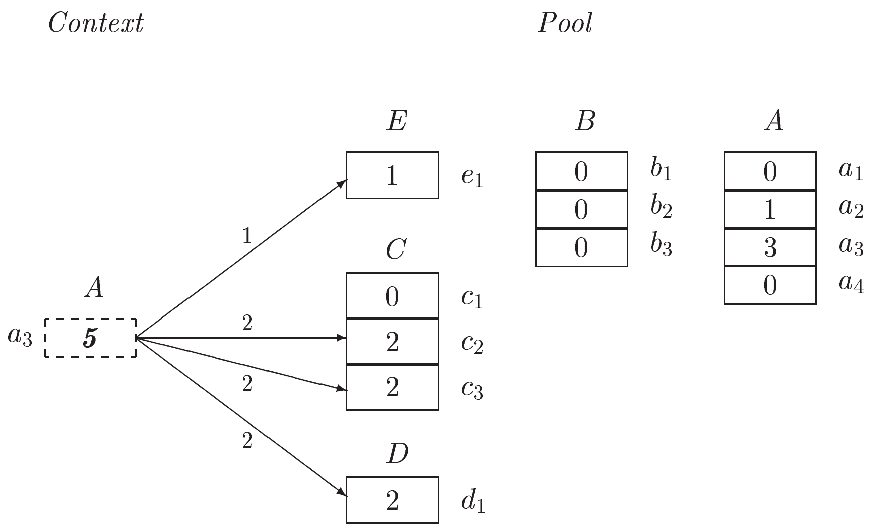
Termination

FOR each word $w_i \in \mathcal{W}$

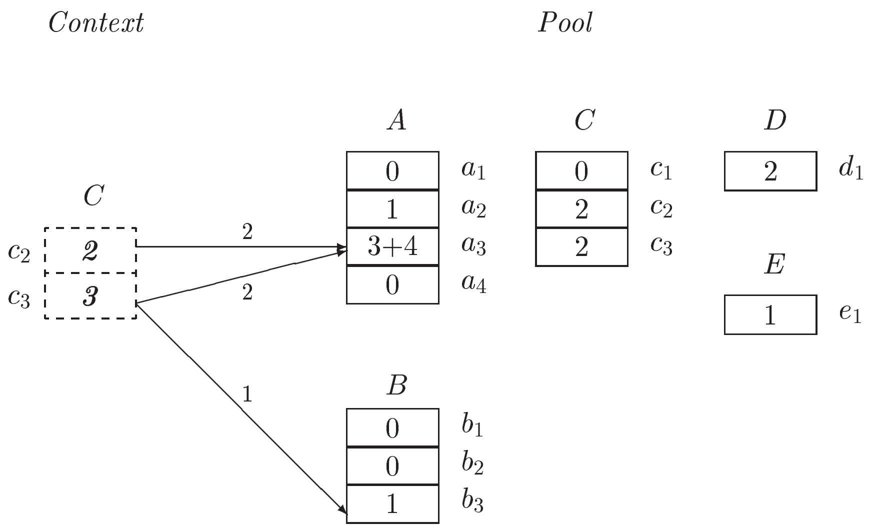
$$Tag(w_i) = \arg\max_j(SS(s_{ij}))$$

NEXT

Figure 5.2 *The Monosemous Filtering Algorithm (MFA)*

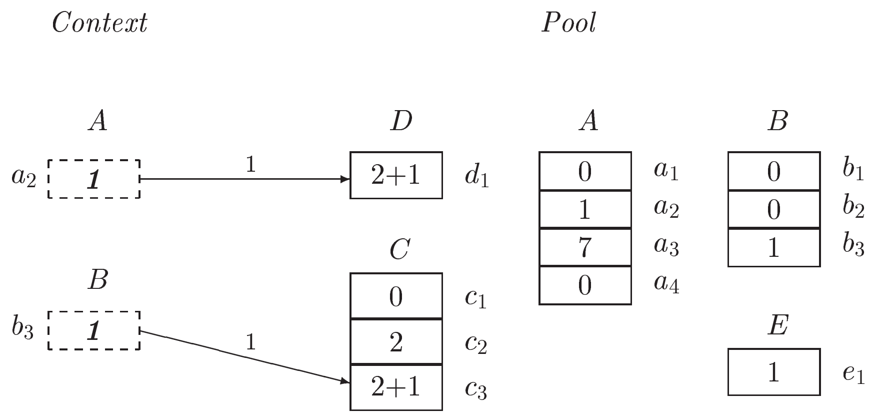


In the third round, c_2 and c_3 additionally reach the threshold, and are hence filtered out:

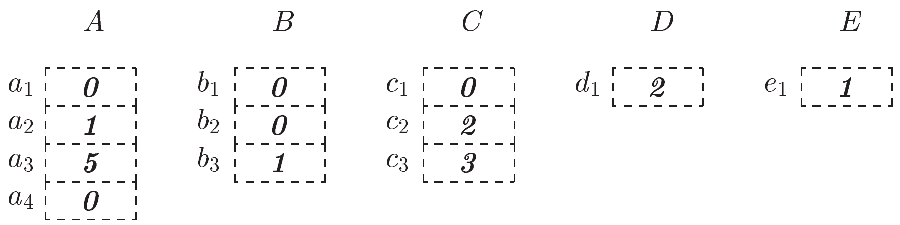


Even though $AS(a_3)$ is increased in this round, it will not be selected again since it has already been filtered earlier.

In the fourth round, nothing else crosses the threshold T, so we lower it to $T = 1$, and find that a_2 and b_3 now qualify:



Here comes the end of the process because no additional sense from *Pool* crosses the threshold, and when lowered again, T reaches its minimum. So we end up with these Supportive Scores:



Choosing the sense with maximum SS for each polysemous noun, A, B, and C will be assigned senses a_3, b_3, and c_3 respectively.

5.4.2 Experiment MFA-1

Objectives

This experiment has two aims: (1) To test how well MFA approximates large contexts from WBA (see Section 5.3.3). (2) To continue the study of the effect of context size by testing full-text context with MFA.

Materials and Method

For the first objective, the WBA with ±3000 words[7] (*wb3000*) and MFA with full-text context (*mf3000*) were compared. Since *wb3000* was estimated to be extremely time-consuming, the two methods were only compared with respect to the seven SEMCOR texts in category A (Press: Reportage).

[7]As each SEMCOR text is about 2,000 words long, a context size of ±3000 words would cover the whole text.

For the second objective, MFA with full-text context was run on the entire 186 SEMCOR files as in Experiment WBA-1.

Throughout this experiment, the metric for measuring taxonomic distance was the same as the one used in Experiment WBA-1.

Results and Discussion

The disambiguation results for the seven category A texts at test conditions *wb3000* and *mf3000* are shown in Table 5.3. It can be seen that with context size at its upper limit (for SEMCOR texts) and with respect to the type of disambiguation information used, MFA in general performs slightly better in both precision and recall than WBA (except in br-a02 and br-a11, where MFA is much better). Moreover, MFA apparently takes less time than WBA at a context size as large as full-text. The results thus suggest that MFA approximates WBA quite well at this particular context size and with everything else held constant. It therefore seems reasonable to replace WBA with MFA for testing larger contexts.

File	***wb3000***		***mf3000***	
	Precision (%)	**Recall (%)**	**Precision (%)**	**Recall (%)**
br-a01	40.88	40.76	40.41	40.18
br-a02	30.86	30.58	40.19	38.84
br-a11	35.97	35.62	41.53	40.85
br-a12	22.78	22.64	24.76	24.53
br-a13	25.57	25.09	27.10	26.59
br-a14	39.86	38.82	40.61	39.14
br-a15	43.27	42.96	45.09	44.77
Ave	**34.17**	**33.78**	**37.10**	**36.41**

Table 5.3 *Comparison between wb3000 and mf3000*

Consequently, running MFA with full-text context on all SEMCOR files produced the results in Table 5.4. The column **Round** in the table shows the average number of cycles which the algorithm takes to converge. On average, both precision and recall reach about 33%. Again, disambiguation performance on categories M (Fiction: Science), N (Fiction: Adventure and Western), and R (Humour) compares less favourably with others.

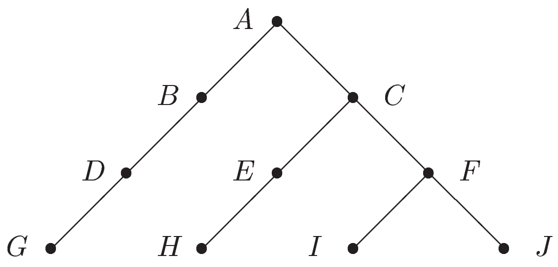
Experiments WBA-1 and MFA-1 showed that increase in context size did not seem to produce much significant effect on WSD performance, even when full-text context was tested. Although it is possible that an optimal context size may have been reached in Experiment WBA-1, we have yet to check the second possible reason, the nature of the distance metric used, for such an unexpected effect of context size.

Cat	Round	Precision (%)	Recall (%)
A	33	37.10	36.41
B	25	33.85	33.64
C	28	33.13	32.32
D	20	34.38	33.74
E	30	34.25	33.55
F	28	34.92	34.40
G	27	34.28	33.46
H	27	37.80	37.06
J	24	35.27	34.44
K	20	31.34	30.71
L	21	33.22	32.31
M	26	27.29	26.72
N	23	29.60	28.87
P	21	32.92	32.15
R	23	29.66	29.29
Ave	**25**	**33.84**	**33.13**

Table 5.4 *Results of Experiment MFA-1*

5.4.3 Modification of the Metric

The metric that we have been using since Experiment WBA-1 assumes that all the links in the WordNet hierarchy are uniform, representing identical distances between nodes. So according to a hypothetical segment of the hierarchy like this one[8]:



our metric would render J be equally similar to B and H, since there are four intervening edges in both cases. But in reality, such equality in edge numbers cannot guarantee that the actual semantic distances are the same. The main reason is that the fineness of classification may vary with different parts of the hierarchy. For example, Agirre and Rigau (1996) assumed concepts in a deeper and denser part of the WordNet hierarchy are closer together, and they measured semantic similarity in terms of "conceptual density". Hence, all we can assume safely is that the links in the hierarchy would be relatively uniform provided that they are within a local scope and the concepts linked by them do not differ too much in specificity, i.e. hierarchical level.

[8]This is just a general example, and the discussion is not necessarily restricted to dichotomous hierarchies.

So we can modify our metric to cope with the peculiarity of WordNet by predefining the size of such a local scope. Within a local scope, we will be confident enough to say that two nodes are reasonably similar and such similarity can thus be taken as evidence for disambiguation.[9] Besides, restricting comparisons to local scopes can help cut short the comparison between senses which only meet after a long chain of connection or never meet anywhere in the hierarchy at all. We will find the best scope by empirical means, and will explain how in this section. Note that we say "best" instead of "optimal", and the reason will be explained shortly.

Local Scope Specification

A *scope* of a sense s is characterised by a two-digit code. The first digit, which we call the *Ancestor Level (AL)*, refers to the maximum order of ancestors included. A first-order ancestor of s is its direct hypernym, a second-order ancestor is its super-hypernym, and so on. The second digit, which we call the *Relative Level (RL)*, refers to the maximum order of relatives (down to the generation of s) included. A first-order relative of s relates to s via the direct hypernym of s, a second-order relative relates to s via the super-hypernym of s, and so on. So with reference to the hierarchical segment above, scope *20* of J includes F and C, and scope *13* of J includes F, B, D, G, E, H, and I.

To maintain reasonable accuracy and efficiency, we hold the following principles when choosing the best scope:

1. *Precision weighs more than recall.*

   Theoretically we want to maximise both precision (P) and recall (R), but in practice there is always a trade-off. If we choose to optimise on one of them, we run the risk of a very low score of the other. We intend to find a compromise based on an effectiveness index E, following van Rijsbergen (1979):

   $$E = 1 - \frac{1}{\alpha(\frac{1}{P}) + (1-\alpha)\frac{1}{R}}$$

   Different levels of α will be compared and contrasted, but we will take $\alpha = 0.90$ for the selection. As pointed out earlier, correctness of the intermediate results of the MFA would be unknown, but the effect would be relayed from one cycle to the next. Too low a precision is thus unfavourable to subsequent rounds. We therefore need to spare some recall for a little more precision, and set a high α. This also explains why we use "best" to describe the scope to be chosen because it does not necessarily give "optimal" value for P or R.

2. *AL has priority over RL.*

   In the event of a tie between two or more scopes, the one with minimum AL and RL would be chosen. If the differences with respect to the two digits contradict each other, the scope with lower RL would be preferred. For example, if scopes *01* and *10* have the same E, scope *10* would be given priority. The reason is

[9] cf. Strength of relation and allowable paths in lexical chains (Hirst and St-Onge, 1998).

that one order of RL usually involves many more nodes than one order of AL (especially if the hierarchy is more than dichotomous). So by preferring AL to RL, computation can be kept low.

Pre-defining Best Local Scope

The files in the first half of each SEMCOR category were used as a trial sample for testing different local scopes. Both the AL and the RL were varied from 0 to 3, giving a total of 16 scopes to compare. The MFA was run on the text samples for the first cycle only.

The average results given by different scopes are shown in Table 5.5, with the maximum precision and recall shaded. The effectiveness indices of different scopes at various levels of α are plotted in Figure 5.3, and the average best scope for each text category at $\alpha = 0.90$ is shown in Table 5.6. The number of files tested for each category is indicated in brackets in the table.

It is obvious from Table 5.5 that optimal precision was obtained from the most constrained scope (apart from 00) whereas optimal recall from the least constrained one. This outcome is normal because when more distant relationships are allowed, there will be more hits but also more noise.

As mentioned, one order of RL (*Relative Level*, the second digit of a scope code) enlarges a scope more than one order of AL (*Ancestor Level*, the first digit). Consequently, we see from Figure 5.3 that at a low α (e.g. 0.25) where recall is deemed more important, scope effectiveness improves relatively more slowly as AL increases than as RL does. Moreover, the best scopes at low α levels are all in the order of $RL = 3$. This suggests that the gradient of recall between more constrained and less constrained scopes is much greater than that of precision. The effectiveness indices approach a constant as RL is 2 or higher, showing that at some point, further expansion of scope will not help. It may be a ceiling effect with respect to the depth of the WordNet hierarchy, but it also justifies our motivation to pre-define scopes to avoid unnecessary comparisons.

The best scope varies considerably with text categories as seen in Table 5.6. However, since the number of files tested for each category was quite different, it is not certain how reliable these results are. The best scope on average is scope *31* at $\alpha = 0.90$. So in the following experiments, we will modify our metric for measuring taxonomic distance with respect to WordNet by constraining the valid distances between two senses to a local scope of *31* from each sense.

5.4.4 Experiment MFA-2

Objectives

The aim of this experiment is to test the effect of the modified metric on WSD in general, and on the MFA for approximating large contexts.

Scope	Precision (%)	Recall (%)
00	42.53	1.46
10	53.48	8.60
20	50.07	11.58
30	48.85	12.88
01	41.11	9.05
11	46.76	14.60
21	47.28	17.24
31	46.74	18.28
02	39.61	26.65
12	39.55	26.75
22	40.72	28.28
32	41.15	29.06
03	37.83	33.38
13	38.47	34.09
23	38.69	34.41
33	39.01	34.81

Table 5.5 *Average Results from First Cycle of MFA with Different Scopes*

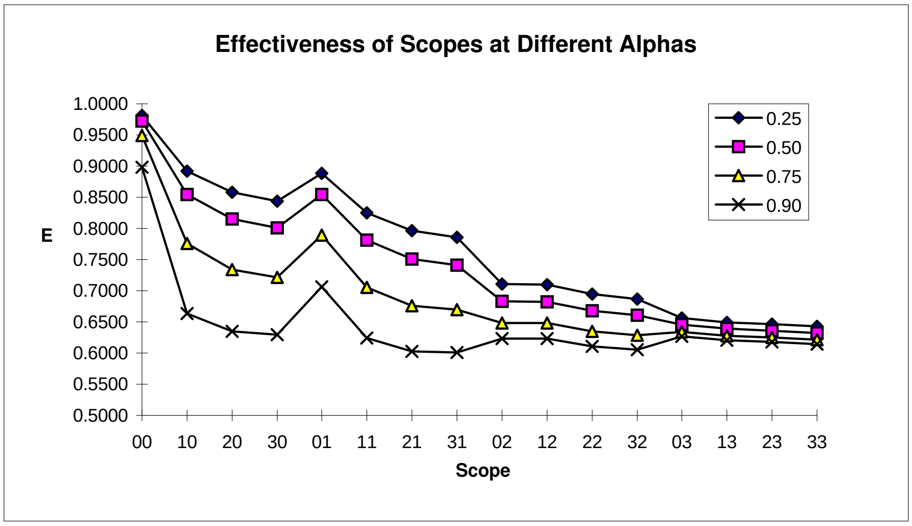


Figure 5.3 *Scope Effectiveness at Different Levels of* α

RL \ AL	0	1	2	3
0				H(6)
1			B(1),F(9),J(21)	A(4),N(5),P(3)
2	D(2),E(7)			G(9),K(15),R(3)
3		L(5),M(1)		C(2)

Table 5.6 *Best Scope for Individual Text Categories at* $\alpha = 0.90$

Materials and Method

In this experiment, all 186 SEMCOR files were used, as in Experiment MFA-1. The MFA with full-text context was run on these files, but this time the measurement of taxonomic distance between senses was constrained within a local scope (*31*, selected in Section 5.4.3) from each sense. Relations outside the local scopes of individual senses were disregarded. We will call this test condition *mf3000c* (*c* for "constrained").

For comparison, the WBA (see Section 5.3.3) with ±3000 words as context but with the modified metric was also tested on the seven texts in category A, as in Experiment MFA-1. We call this condition *wb3000c*.

Results and Discussion

The results from *mf3000c* are shown in Table 5.7. Compared to *mf3000* in Experiment MFA-1 (Table 5.4), it is obvious that constraining valid taxonomic distances to local scopes can cut short the disambiguation process, as it now takes on average 11 rounds fewer for the MFA to converge. Also, average precision has improved a lot, from 33.84% in the unconstrained condition to 42.02% in the constrained one. Nevertheless, average recall has dropped slightly, from 33.13% to 29.05%. The increase in precision suggests that relations found between senses within local scopes are less spurious, or more reasonably grounded. The reduction in the number of spurious connections is also reflected in the drop in recall since spurious connections can nevertheless deliver correct outcomes.

Cat	Round	Precision (%)	Recall (%)
A	14	56.90	37.86
B	14	40.63	29.46
C	18	40.27	27.31
D	14	40.35	27.88
E	16	46.75	35.49
F	16	46.60	34.77
G	17	39.80	30.17
H	16	45.61	34.30
J	14	44.04	32.94
K	11	36.32	21.26
L	10	34.60	19.35
M	13	37.83	23.43
N	10	37.88	20.54
P	9	42.88	22.46
R	13	35.27	23.37
Ave	**14**	**42.02**	**29.05**

Table 5.7 *Results of Experiment MFA-2*

The results of *wb3000c* and *mf3000c* with respect to the texts in category A are shown in Table 5.8. If we compare the results in Tables 5.8 and 5.3 (from Experiment MFA-1), we see that with both WBA and MFA, local scope gives both better precision and recall than simple unconstrained taxonomic distance. However, unlike in Experiment MFA-1 where both precision and recall were slightly better with MFA than with WBA, we found that while precision is higher in *mf3000c* than in *wb3000c*, recall is lower. This may suggest that the difference between the actual full-text context (with WBA) and its approximated counterpart (with MFA) is greater when the modified metric is used, as compared to Experiment MFA-1. This effect will be further discussed in the next section.

File	***wb3000c***		***mf3000c***	
	Precision (%)	**Recall (%)**	**Precision (%)**	**Recall (%)**
br-a01	51.96	50.44	61.95	41.06
br-a02	49.50	45.26	57.60	44.04
br-a11	52.61	49.35	66.04	45.75
br-a12	47.92	43.40	47.62	28.30
br-a13	49.40	46.07	59.06	37.83
br-a14	48.73	44.08	51.20	35.20
br-a15	51.89	49.46	54.82	32.85
Ave	**50.29**	**46.86**	**56.90**	**37.86**

Table 5.8 *Comparison between wb3000c and mf3000c*

So in this whole section we have investigated the two possible reasons for the limitation of the results in Experiment WBA-1 (Section 5.3.3). We have designed a data-driven and self-corrective recursive filtering algorithm, the MFA, to enable larger contexts to be used in WSD, which is otherwise impractical due to combinatorial explosion. We have also modified the metric for measuring taxonomic distance with respect to the WordNet hierarchy by pre-defining empirically a local scope to constrain the measurement. While WSD performance at expanded context with the original metric bears little difference from the results in Experiment WBA-1, there is a difference between the original metric and the modified metric for WSD at expanded context. So we now have to consider the relation between the modified metric and context size.

5.5 Context Size (Part Three)

In this section, we will resume our discussion on the effect of context size left from Section 5.3. We will repeat the experiment with the WBA on various context sizes, but constraining the measurement of taxonomic distance by local scopes. This new experiment will be reported in Section 5.5.1, followed by a summary of what we learned from the whole series of experiments (from Section 5.3 to Section 5.5) regarding context size for WSD.

5.5.1 Experiment WBA-2

Materials and Method

The test materials, test conditions, and performance measures for this experiment are basically the same as those in Experiment WBA-1 (Section 5.3.3). The only difference is that we now use local scopes (pre-defined empirically in Section 5.4.3) to constrain the measurement of taxonomic distance with respect to the WordNet hierarchy. So the various test conditions (i.e. various window sizes) in this experiment will be indicated as *wb05c*, *wb20c*, *wb50c*, and *wb70c* correspondingly.

To further explore the effectiveness of the MFA for approximating large contexts with the modified metric, we focus our analysis on two text types, namely J (Learned and Scientific Writing) and K (Fiction: General). These were deliberately chosen as intuitively very different types in terms of how strongly a text relates to a particular subject area. Since there were 43 texts available in J but only 29 in K, we based the analysis on all texts from K and 29 texts chosen at random from J. Also, as a control, two random samples from J (J1 and J2) were studied. These selected texts gave more than 18,000 test samples, representing occurrences of about 3,000 different words ranging in number of senses from 2 to 30, with average 3.8. These samples were additionally tested for the *wb3000c* condition (see Experiment MFA-2), and results were compared to those from *wb05c*, *wb20c*, *wb50c*, *wb70c*, and *mf3000c*.

Results and Discussion

Table 5.9 shows precision and recall at the various test conditions with respect to individual text types for the 186 SEMCOR files. The maxima of both measures for each text type are shaded. At *wb70c*, the average precision and recall are about 43% and 26% respectively. It is also apparent that with our modified measurement of semantic distance, increasing context size has a more dramatic effect on WSD performance. On the one hand, average recall increases by more than 20% from ±5-word to ±70-word context, yet only with slight loss of average precision. On the other hand, the rate of increase in recall slows down as context size increases.

Cat	**Precision (%)**				**Recall (%)**			
	wb05c	***wb20c***	***wb50c***	***wb70c***	***wb05c***	***wb20c***	***wb50c***	***wb70c***
A	56.79	52.46	54.04	52.97	5.21	14.28	25.77	30.03
B	46.83	46.80	44.62	43.96	5.34	14.42	21.24	24.11
C	50.92	42.67	42.04	43.23	4.43	11.53	20.65	24.64
D	39.77	39.27	42.13	42.03	5.62	14.58	23.95	26.91
E	52.57	46.49	47.18	48.10	7.23	17.38	26.62	30.42
F	50.30	47.67	46.79	46.73	5.70	15.15	24.50	27.78
G	51.78	45.92	42.88	42.87	6.16	15.79	23.97	26.98
H	49.71	48.96	48.30	48.79	7.85	18.66	28.03	31.46
J	47.05	44.30	44.59	44.59	6.27	16.48	25.91	29.14
K	46.41	40.77	40.43	40.03	4.00	10.85	17.92	20.58
L	42.12	37.43	36.34	36.03	3.88	10.07	15.99	18.24
M	41.84	41.01	43.60	42.74	5.78	11.72	20.60	22.34
N	44.47	36.10	35.91	37.75	3.93	9.51	15.58	19.30
P	48.71	40.10	39.22	38.31	4.60	10.06	16.21	18.75
R	32.89	34.91	36.02	36.15	3.84	9.12	15.54	19.16
Ave	47.71	43.60	43.28	43.33	5.52	14.13	22.49	25.66

Table 5.9 *Results of Experiment WBA-2*

The results for the selected samples from categories J and K at all test conditions are shown in Table 5.10.

Condition	Precision (%)			Recall (%)		
	J		K	J		K
	J1	J2		J1	J2	
wb05c	48.19	48.57	46.41	6.32	6.47	4.00
wb20c	44.54	45.67	40.77	16.80	16.92	10.85
wb50c	44.80	45.52	40.42	26.23	26.21	17.97
wb70c	44.90	45.42	40.03	29.60	29.43	20.58
wb3000c	46.48	45.58	37.51	44.60	43.46	34.46
mf3000c	42.80	45.51	36.32	31.83	34.17	21.26

Table 5.10 *Results for Selected Samples from J and K*

We first compare all the *wbXc* conditions for the effect of context size on WSD. In general, average recall improves (but at a decreasing rate) as context size increases. However, the improvement in recall is accompanied by a drop in average precision. More interestingly, the drop in precision from *wb05c* to *wb3000c* with text type K (fiction) is about three times greater than that in J (scientific). Precision drops initially and then remains quite stable from *wb20c* to *wb70c* for both text types; but while with J this level of precision is maintained at *wb3000c*, the precision in K suffers a further drop at *wb3000c*. This suggests that noise may increase more severely in K than in J as more context is used.

As for the performance of the MFA, we see from Table 5.10 that *mf3000c* has similar precision but lower recall compared to *wb3000c*, although the recall for *mf3000c* is somewhat greater than the values for the WBA with smaller window sizes. Looking at the actual results on individual texts, we found that precision was rather consistent across texts, but there was noticeable variation in recall.

In order to analyse our results more thoroughly, particularly since our text set size was not really large, we ranked the texts by their recall values in *mf3000c* and compared the averages over the top N texts for *wb20c*, *wb3000c* and *mf3000c*, as shown in Figure 5.4.

From Figure 5.4 we observe the following. First, under every test condition, the absolute recall value for J is always better than that for K. This suggests that the particular disambiguation information we used (i.e. the IS-A relation) may be inherently less suitable for disambiguating common, multi-purpose words like those typical of fictional texts.

Second, although recall for *mf3000c* is not as good as for *wb3000c*, it is nevertheless much better than recall for *wb20c*. Since a context of ± 20 words is often considered large by WBA standards, the MFA appears to be a rather good strategy for expanding context.

Third, the behaviour of MFA is different for the two text types. On the one hand, the distance between *mf3000c* and *wb3000c* is much smaller in J than in K. On the other hand, this distance in J seems to increase as we average over more texts, whereas in K the corresponding distance between *mf3000c* and *wb3000c* is rather

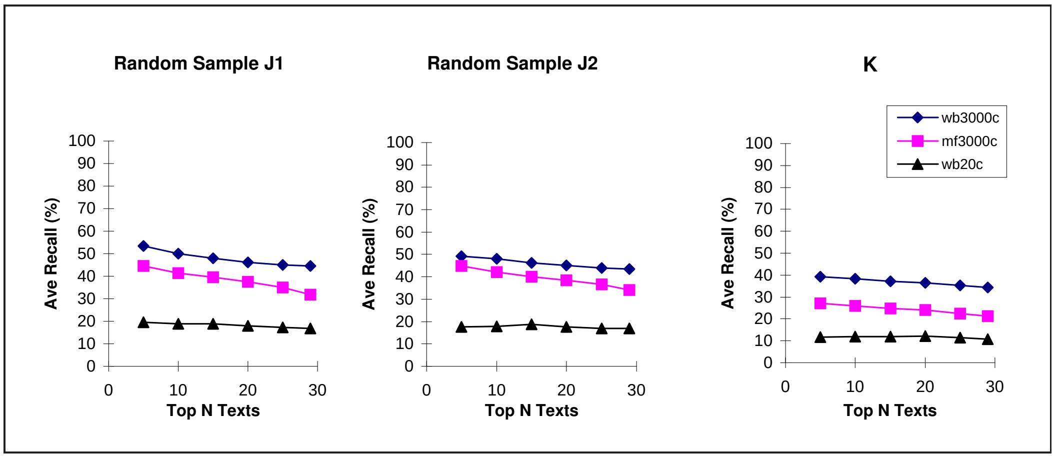


Figure 5.4 *Average Recall over Top N Ranked Texts in J and K*

stable.[10] Since the MFA depends on the number of monosemous nouns in the text and how good they are for disambiguating the rest, if they in fact provide a poor anchor at the beginning, this would cause the algorithm to stop prematurely and fail to expand context. Hence, to account for the divergence between *wb3000c* and *mf3000c*, we can hypothesise that some texts in J have exceptionally few monosemous nouns, or are in fact like those in K which do not have sufficiently useful monosemous nouns to drive the MFA. Nevertheless, where good anchors are available, the MFA is effective in approximating large contexts. Such a contrast of word uses in different text types will be further addressed below and in the next chapter.

5.5.2 Effect of Context Size Revisited

The lesson we have learned about the effect of context size can be summarised as follows. In general, with a narrow semantic relation as disambiguation information, recall for WSD increases with context size at a decreasing rate and seems likely to level off at some point. However this increase is at the expense of precision, though variably with different text types. More interestingly, the difference in the magnitude of recall between different types of text is large enough to suggest that the same type of disambiguation information may not be equally useful for disambiguating words typical of different text types. We will discuss this more in the next section where we study the realisation of context. Also, the effectiveness of the MFA for approximating large contexts seems to depend on not only how strongly a text relates to particular subject areas but also on the actual words in individual texts. This relation between individual words and WSD will be thoroughly investigated in Chapter 6.

5.6 Realisation of Context

In the last few sections, we have studied the effect of context size with respect to a particular type of semantic relation, and found that full-text context produced optimal

[10] The distance between *mf3000c* and *wb3000c* differs for J1 and J2. This suggests larger samples should be investigated, but the general difference from K is similarly supported by both samples.

results compared to smaller sized windows. In this section, we will thus control the context size at full-text, and investigate another aspect of context, namely realisation.

Apart from studying the effect of different types of realisation, another important objective in this section is to demonstrate the practical usefulness of the Integrated Lexical Resource (ILR) developed in Chapter 4. So we will first describe in Section 5.6.1 a spectrum of semantic relations, i.e. semantic relations of various degrees of specificity, and how such relations can be realised from the ILR. Then we will test the individual and combined effect of these semantic relations on WSD in Sections 5.6.2 and 5.6.3 respectively.

5.6.1 A Spectrum of Semantic Relations

In this section, we will describe some semantic relations and how they can be obtained from the ILR. Those are not the only ways this can be done but are nevertheless some obvious ones. We organise the semantic relations in terms of their specificity, from the narrowest type to the broadest type. This ranking is essentially intuitive, taking also the number of component lexical resources involved into account.

Narrow Semantic Relation: WN16 Relations (*wn*)

This is the relation used in our experiments on context size, and is measured by taxonomic distance constrained by local scopes with respect to the WN16 IS-A hierarchy as described in Section 5.4.3.

Broad Semantic Relation 1: ROGET Relations (*rog*)

Since ROGET also has a hierarchical structure, we can measure the distance between two senses with respect to the ROGET hierarchy as we do with WN16. However, the flatness and broad classification of the former require some other constraints on the measurement, or there would be too much noise.

For this purpose, we first find the set $TopC$ which contains ROGET heads covering most words in the text. Heads at the top two positions with eight or more words in the text are included (determined empirically), and ties are preserved. The similarity of two WN16 senses x and y with respect to the ROGET classification (now a component of the ILR) can then be measured by a function as suggested in Chapter 4. The exact function used in the following experiments was determined empirically as:

$$S(x,y) = \begin{cases} |M(x) \cap M(y)| & \text{if } M(x) \cap M(y) \neq \emptyset \\ & \text{and } (M(x) \cap M(y)) \cap TopC \neq \emptyset \\ & \\ 0.25 & \text{if } M(x) \cap M(y) = \emptyset \\ & \text{and } h_1 \in M(x), h_2 \in M(y), \\ & h_3 \in TopC, S(h_1) = S(h_2) = S(h_3) \\ & \\ 0 & \text{otherwise} \end{cases}$$

where
$M(s)$ = the set of ROGET heads that sense s is mapped to, and
$S(h)$ = the ROGET section (one level above heads) subsuming head h.

Broad Semantic Relation 2: Definition Match (*def*)

This is the word overlap between pairs of sense definitions in the tradition of Lesk (1986). Sense definitions are based on the WN16 glosses, with stop words and duplicated words removed.[11] The similarity between senses is measured by the number (≥ 2) of words common to their definitions, normalised with the Dice function:

$$S(x,y) = \begin{cases} \frac{2|X \cap Y|}{|X|+|Y|} & \text{if } |X \cap Y| > 1 \\ 0 & \text{otherwise} \end{cases}$$

where
X is the definition of sense x, and
Y is the definition of sense y.

Broad Semantic Relation 3: Mixing WN16 and ROGET (*com*)

This relation directly measures the relevance of a sense to the text as a whole. Each sense (in the text) is assigned a score based on the length of overlap between its WN16 definition and the set of words covered by the ROGET heads in *TopC* (as defined above).

5.6.2 Experiment REA-1

Materials and Method

The 186 SEMCOR files covering various text types were used as test materials, as in Experiments MFA-2 and WBA-2. Each polysemous noun occurrence (lemmatised) was treated as a target word.

The test files were disambiguated with the different realisations described in the last section. While *wn*, *rog* and *def* were run with MFA[12], *com* was run as a one-step heuristic since it does not involve any transitive comparison between pairs of senses. In addition to these types of information, and for the purpose of comparison, we also tested tagging the nouns simply with their first sense from WN16, which is expected to be the most frequent sense (*freq*).

As in the previous experiments, the measures of performance were precision and recall, as defined in Section 5.3.3.

[11] Lacking reliable mechanisms, no stemming was done. However, perfect stemming should be beneficial for definition matching.

[12] To compensate for the possible inadequacy of the sense linkage (see Chapter 4), we took an additional step when running *rog* with MFA. Apart from driving the algorithm with the monosemous nouns (more precisely the linked ROGET heads of these nouns), we also assigned credits to the candidate senses which were under the ROGET heads in *TopC*, in case the mapping has masked some likely senses and thus prevented them from being filtered.

Results and Discussion

Table 5.11 shows the average precision and recall from different context realisations. Results for *wn* are the same as those in Experiment MFA-2. The best outcomes (excluding *freq*) in each row are shaded. We will first summarise the performance of individual semantic relations in general, and then discuss their effect on different text categories.

Cat	Precision (%)					Recall (%)				
	wn	*rog*	*def*	*com*	*freq*	*wn*	*rog*	*def*	*com*	*freq*
A	56.90	38.14	42.35	39.30	69.33	37.86	12.47	29.55	29.82	69.33
B	40.63	38.31	30.23	27.65	64.99	29.46	12.14	19.38	19.02	64.99
C	40.27	46.96	28.72	33.07	63.37	27.31	21.24	19.96	24.48	63.37
D	40.35	47.59	31.65	33.31	63.27	27.88	14.46	19.59	24.41	63.27
E	46.75	45.08	38.83	39.93	67.74	35.49	18.06	27.07	29.55	67.76
F	46.60	46.44	38.31	40.29	67.36	34.77	17.49	26.94	30.29	67.36
G	39.80	42.12	32.20	34.03	63.69	30.17	16.24	21.92	25.16	63.69
H	45.61	44.70	38.62	43.26	71.08	34.30	16.75	25.62	32.04	71.08
J	44.04	42.34	38.89	38.85	69.97	32.94	16.59	23.11	28.96	69.97
K	36.32	38.55	31.25	33.68	71.34	21.26	12.40	17.61	23.32	71.35
L	34.60	41.84	25.58	35.44	67.33	19.35	13.05	12.68	26.46	67.35
M	37.83	54.87	34.93	41.69	69.18	23.43	15.34	21.39	28.41	69.31
N	37.88	37.20	30.77	34.04	71.65	20.54	14.79	15.40	25.79	71.65
P	42.88	44.96	32.46	30.70	69.78	22.46	13.59	18.81	23.30	69.78
R	35.27	37.79	29.09	29.95	69.53	23.37	15.29	16.94	22.38	69.57
Ave	42.02	42.28	34.90	36.81	68.79	29.05	15.50	21.68	27.15	68.80

Table 5.11 *Results of Experiment REA-1*

On average, *wn* and *rog* both achieve over 40% precision. However, there is a nearly two-fold difference in recall. The inferiority of *rog* can be attributed to two factors. First, as seen in Chapter 4, ROGET has a much smaller collection size than WN16. The number of words which both resources share is even smaller. Second, the results depend on not only the actual effectiveness of the semantic relation but also on the accuracy of the sense linkage between WN16 and ROGET. However, *wn* and *rog* have similar precision, suggesting that where the information is applicable, both types of information are similarly effective for disambiguation.

Meanwhile, the other broad semantic relations which are realised from the definitions of the senses, i.e. *def* and *com*, appear to result in lower precision than *wn* and *rog* but comparable recall. This is probably because definitions contain "noisy" words, although an effort has been made to keep them to a minimum (e.g. by removing stop words).

As for *freq*, it gives almost 70% precision and recall on average. Compared to other studies which used most frequent sense / first sense as baseline, what we get appears extraordinarily high. For instance, Wilks and Stevenson (1998) got a baseline of only 49.8% using first sense in LDOCE. We believe that the anomaly of *freq* is mostly due to the circularity between SEMCOR and WordNet: the former is tagged with senses from the latter while the sense ordering in the latter at least partly depends on what is observed from the former.

As suggested from our previous experiments on context size, narrow semantic relations may not be a very suitable type of information for disambiguating words in fictional texts. The results in Table 5.11 not only support this conjecture, but also show further interesting phenomena. In terms of recall, for instance, while *wn* does not work well for the fictional categories including K (Fiction: General), L (Fiction: Mystery and Detective), M (Fiction: Science), N (Fiction: Adventure and Western), and P (Fiction: Romance and Love Story), these categories are all best disambiguated by *com*, the broadest semantic relation according to our ranking in Section 5.6.1. This contrast thus suggests that different types of semantic relation are needed as they may complement one another at different situations (e.g. text types), leading us to the next experiment on the use of multiple types of information for WSD.

5.6.3 Experiment REA-2

Materials and Method

The test materials (i.e. 186 SEMCOR files) and method (i.e. application of different realisations) were the same as in Experiment REA-1. As mentioned in the previous chapters, most WSD studies which use multiple types of information usually constrain the information within a single resource. Few have actually used multiple lexical resources simultaneously. We will cover both possibilities in this experiment: for multiple types of information within a single resource, we tested *wn+def*; for multiple types of information from multiple resources, we tested *wn+rog*, *wn+rog+com*, and *wn+rog+com+def*. Moreover, for all these combinations, two variations – with and without *freq* (*+freq* and *–freq* respectively) – were tested.

The actual scores from individual realisations were first converted to *relative scores* to normalise the different scoring scales as follows:

$$S_r(w_i) = \frac{S_a(w_i)}{\max_j(S_a(w_j))}$$

where $S_r(w_i)$ is the relative score for the i^{th} sense of word w, and $S_a(w_i)$ is the *actual score* for the i^{th} sense of word w. The relative scores from different realisations for each sense were then added up for various combinations of realisations correspondingly. The measures of performance were precision and recall as in the last experiment.

Results and Discussion

The overall average precision and recall for the various combinations of context realisations are shown in Table 5.12. We will discuss the overall results below, while the results for individual text categories can be found in Appendix C.

Comparing Tables 5.11 and 5.12, we see that all combinations of realisations (at *–freq*) outperform their individual component realisation (e.g. *wn+def* performs better than *wn* or *def* alone). However, the effect of realisation combination is not necessarily additive. In other words, more types of information do not always give proportionally better results. For example, *wn+rog+def+com* only gives a very minor increase in recall compared to *wn+rog+com*. This situation is especially obvious when *freq* is

Realisation Combination		Precision (%)	Recall (%)
wn+def	*−freq*	43.24	35.50
	+freq	67.97	66.35
wn+rog	*−freq*	44.40	33.97
	+freq	70.82	68.91
wn+rog+com	*−freq*	42.59	37.65
	+freq	65.14	62.93
wn+rog+def+com	*−freq*	42.13	38.63
	+freq	61.22	59.36

Table 5.12 *Overall Results of Experiment REA-2*

included. For instance, given the extraordinary performance of *freq* alone in Experiment REA-1, we found in Experiment REA-2 that only *wn+rog+freq* outperforms *freq* on its own. The recall in the former is marginally better and the precision is 2 to 3% higher.

There are at least two possible causes for the above results regarding the combination of information types. On the one hand, two or more different types of information may in fact be good for disambiguating a similar set of words, so they do not offer much additional advantage to WSD when combined. On the other hand, the effect of different types of information may offset one another when undiscriminatingly combined. That means one type of information may be good for disambiguating particular words while another type of information may be adverse for those words. Therefore, for the combination of information to be effective, conflicting evidence must be resolved. Handling conflicting evidence thus requires the preferences of the target words (and their senses) to be known. This is an issue which has not been adequately addressed in past studies, and we will further explore it in Chapter 6.

All in all, the results in this experiment suggest that using multiple types of information for WSD is often beneficial to WSD. In particular, narrow and broad semantic relations are most useful in different text categories and may therefore complement each other. Most importantly, we have demonstrated that the two types of semantic relation can be effectively obtained and used together by combining two existing lexical resources, despite the limited success of that combination at present.

5.7 An Example

To illustrate the operation of the Monosemous Filtering Algorithm (MFA), and to show the benefits of using multiple types of knowledge from the Integrated Lexical Resource (ILR), we give a real example from one of the test files: br-j03. It is a 1944-word text under “learned and scientific writing”, with 461 noun occurrences, of which 100 are monosemous. The text describes some physics experiment of fluid using a band viscometer.

Figure 5.5 shows the precision and recall at each cycle of the MFA running on br-j03 (using full-text context and narrow semantic relation, as in Section 5.4.4). The dotted vertical lines indicate where the threshold was lowered. The graph thus shows that the algorithm converges after 14 cycles in this particular example. The gradually increasing recall shows that more words get disambiguated after each round. Also, the minimum threshold can be initially set to aim for particular desired levels of precision (and recall).

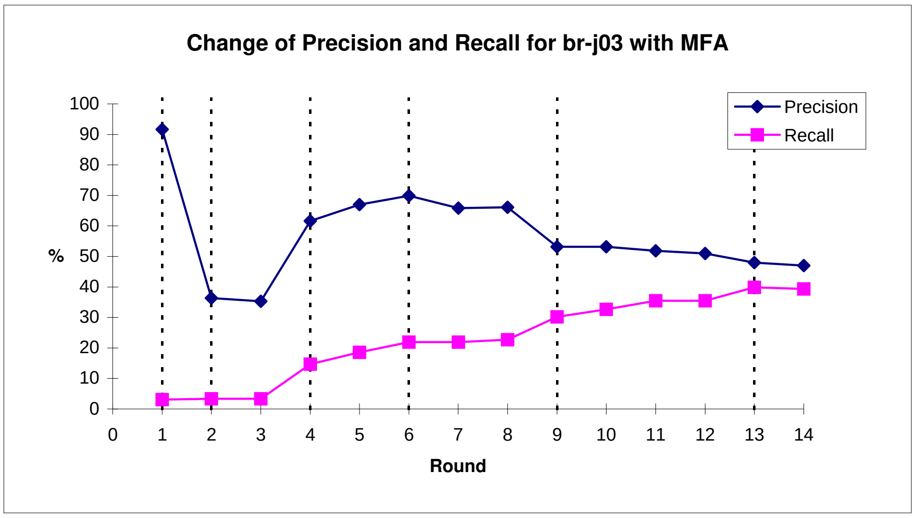


Figure 5.5 *MFA with Narrow Semantic Relation on br-j03*

Another property of the MFA which is not seen in the figure is its being self-corrective. As said, falsely preferred senses can be overridden by better candidates in later rounds. In br-j03, for instance, "heat" was wrongly tagged with sense 2 ("hotness") in round 10, but the expected sense 1 ("heat energy") emerged in round 12.

Table 5.13 shows the results of certain context realisations tested in Sections 5.6.2 and 5.6.3 on br-j03. The benefit of using multiple types of information from the ILR is evident. Combining *wn* and *rog* gives precision and recall which are superior to either *wn* or *rog* alone. Adding *freq* to *wn*+*rog* moreover outperforms *freq* itself. More importantly, we have demonstrated that our ILR does provide enriched information which can be used in practice to the benefit of WSD.

Realisation		**Precision (%)**	**Recall (%)**
Narrow Semantic Relation *(wn)*		47.02	39.34
Broad Semantic Relation *(rog)*		57.98	38.23
Frequency *(freq)*		65.65	65.65
Combined (*wn*+*rog*)	−*freq*	58.58	54.85
	+*freq*	74.08	72.85

Table 5.13 *Results of MFA with Different Semantic Relations for br-j03*

To see the contribution of individual types of information, consider the actual words being correctly disambiguated, which are shown below. Senses marked with $*$ cease to be correct with *wn*+*rog*, whereas those marked with $\star$ were only correct with *wn+rog*. The number after the slash (/) indicates the expected sense with respect to WN16.

wn	*wn/rog*	*rog*
angle/1, area/6,	axis/1,	*amount/2, band/6,
assumption/2, basis/2,	energy/1,	breakup/2, calculation/1,
change/2, concentration/2,	force/2,	chance/2, chemist/1,
conclusion/2, diameter/1,	head/6,	design/1, direction/2,
difference/2, *equation/1,	pressure/1,	effect/1, entropy/1,
field/5, fluid/2, heat/1,	rate/2,	exception/1, *experiment/2,
height/1, hole/4,	reason/2,	literature/1, paper/2, run/2,
*inclination/3, instrument/1,	suction/1,	shear/1, *solution/1,
manner/1, material/1,	system/7	strain/1, tape/1, *volume/1,
measurement/1, mechanics/1,		work/7
part/1, plane/2, poise/1,		
*property/2, relation/1,		
relationship/1, shortness/1,		
statement/1, temperature/1,		
tension/5, theory/2, type/1,		
unit/1, upper_limit/1, use/1,		
*value/1, way/1, *work/1		
	*stress/5	

We can see that there are some words which can be disambiguated with either resource (e.g. axis, energy), but more are exclusively disambiguated by information from one resource (e.g. tape, strain, and entropy with *rog*, and field, tension, and poise with *wn*). Using narrow and broad semantic relations together can sometimes overcome the "tennis problem" mentioned in Fellbaum (1998). For example, energy/1, force/1 and suction/2 are all close to heat/1 and field/5 in WN16 for being "physical phenomena". Although shear/1 and strain/1 are also obviously physics-related, they are classified as deformation, i.e. event, in WN16. Nevertheless, ROGET groups the latter two senses together with force, energy, and suction from another perspective, i.e. power (head 160), thus enabling them to be disambiguated.

Correct disambiguation can also result from evidence accumulation. For example, "stress", which is wrongly tagged with either *wn* or *rog* alone, is correctly disambiguated with *wn+rog*. The fact is that stress/5 (the expected sense) lost to different senses with *wn* and *rog* alone, but with *wn+rog*, the correct sense ended up with most accumulated evidence.

The importance of using multiple types of information for WSD can also be inferred from the example as follows. Although we found that full-text context in general works best for certain types of semantic information, using full-text context is in effect forcing the words to have "one sense per discourse", which is not always appropriate. For example, "work" is tagged with sense 1 once and sense 7 four times in br-j03. However, all instances of the word were tagged as work/1 with *wn* and all as work/7 with *rog* as well as *wn+rog*. This suggests that we do need other types of information (perhaps from a more local context) to fine-tune the WSD process.

5.8 Summary

In this chapter, we have explored another extrinsic parameter of WSD, namely *Contextual Information*. It is a vital element in WSD because nearly every method makes use of contextual information of one kind or another. However, little systematic study has been done on this parameter, which makes the role of context in WSD remain largely unclear.

We distinguished two aspects of context: size and realisation, and investigated them separately. We first studied the effect of context size with respect to a type of narrow semantic relation. The Monosemous Filtering Algorithm (MFA) was designed to allow larger context sizes to be tested, which was otherwise impractical due to combinatorial explosion. It was found that with the particular realisation, WSD performance improved with context size initially but levelled off at some point. Moreover, the effectiveness of the MFA for approximating large contexts appeared to depend on text types.

We then controlled the context size at its upper limit and tested the effect of different context realisations. In particular we studied the individual and combined effect of semantic relations of varied specificity, which were obtainable from our Integrated Lexical Resource (ILR). It was found that different words seemed to be better disambiguated with different types of semantic relation, and this interaction was closely related to text types. We have shown that different types of information might complement one another and demonstrated the practicality of our ILR.

The results from the combination of various types of semantic relation suggested that different relations might counteract each other. Therefore the requirements of different target words have to be known in order to handle such conflicting evidence. This will be the focus in the following chapter.

6

Intrinsic Parameter: Nature of Target Words

6.1 Introduction

In this chapter, we will study a third parameter, which is also an intrinsic parameter, of WSD: *Nature of Target Words* (abbreviated to *target nature* in the following discussion). In the last two chapters, we have examined *Lexical Resource* and *Contextual Information* in WSD. Our experimental results in Chapter 5 have implied that different target words make different demand on disambiguation information. In fact, as Resnik and Yarowsky (1997) pointed out, "disambiguation seems highly lexically sensitive, in effect requiring specialized disambiguators for each polysemous word". However, the impact of the target words themselves on WSD has not been sufficiently studied in the past.

By target nature, we refer to some intrinsic properties of the target words at the word level as well as the sense level. At the word level, the number of senses a word has (*polysemy*) and the closeness in meaning of these senses (*dispersion*) may affect the difficulty of the disambiguation task. At the sense level, as mentioned in Chapter 1, we suggested that different types of lexical information vary in their effectiveness to characterise a sense and distinguish it from other senses of the same word, i.e. senses have varied *information susceptibility* (or just *susceptibility* for short). The effect of polysemy and dispersion on WSD has occasionally been suggested in past studies, but susceptibility is still not well understood, and we attempt to investigate this property in greater depth in this chapter.

In Section 6.2, we will first briefly discuss what past studies suggest about the effect of target nature on WSD. Then in Section 6.3, we will describe our hypotheses regarding this parameter and our approaches to study it. In particular, we will investigate the subject from two directions, which are reported in detail in Sections 6.4 and 6.5 respectively, followed by a summary of the chapter in Section 6.6.

6.2 Target Nature in Past Studies

Although past studies did not explicitly address target nature, how this parameter affects WSD can perhaps be inferred from the relation between the test data and the disambiguation results in some studies.

According to Stevenson and Wilks (1999), there are two major kinds of test data in WSD studies. One is based on a small set of selected words with many instances for each word (e.g. Leacock et al., 1993; Yarowsky, 1995). For these cases, the number of candidate senses considered is also often limited, which may be a limitation of the algorithm itself, but is more often the inevitable consequence of data sparseness. The other kind of test data is based on unrestricted texts (e.g. Harley and Glennon, 1997; Stevenson and Wilks, 1999), where all words in a running text would be disambiguated. For these cases, the sense inventory is usually provided by some lexical resource (e.g. a dictionary) and is often more comprehensive.

The implications of the results from testing different data are certainly not the same, but those results could suggest something about the effect of target nature on WSD. Numerically speaking, for instance, the disambiguation accuracy reported from the first type of test data appears higher in general. So in terms of polysemy, it seems that the more the number of senses, the more difficult it is for WSD. In fact, Leacock et al. (1993) observed degradation in WSD performance when more candidate senses were considered, although they only tested for one target word. However, the degradation was not consistent, which made them hypothesise that some senses might be more difficult to disambiguate than others. Moreover, the studies testing smaller numbers of senses and getting apparently better results may in fact be testing with more distinct senses, so dispersion is a potentially confounding factor. Indeed, many studies found that if they relaxed the correctness criterion from the polysemy level to the homonymy level, much better results could often be obtained (e.g. Cowie et al., 1992; Agirre and Rigau, 1996; Harley and Glennon, 1997; Wilks and Stevenson, 1997). However, although the interaction between polysemy and dispersion can hardly be disentangled, the general trend suggested by past WSD studies seems obvious: more senses and more closely related senses tend to make the WSD task more difficult.

Compared to polysemy and dispersion, our understanding of information susceptibility is even more limited. As mentioned in Chapter 2, although many studies have shown that using multiple types of knowledge in WSD produces better performance (e.g. McRoy, 1992; Ng and Lee, 1996; Harley and Glennon, 1997), the contribution of individual types of information was overlooked because they were uniformly applied to every target word. Even though some studies combined information in optimal ways (e.g. Stevenson and Wilks, 1999), the combination was still restricted to "positive evidence". As observed by Leacock et al. (1998), for instance, the benefit of adding "topical context" to "local context" was discounted if the test samples contained a large proportion of "non-topical senses". In other words, a particular type of information may be more useful for disambiguating particular senses but less so for others, i.e. senses have varied susceptibility to different types of disambiguation information.

It is therefore quite obvious that the difficulty of WSD depends on target nature as well as the other extrinsic factors discussed in the previous chapters. This also suggests that one might improve the effectiveness of a WSD system by making the disambiguation process sensitive to the demands of individual target words. However, the relationship between target nature and WSD has not been extensively and satisfactorily studied in the past. It is therefore our primary aim in this chapter to find out more about this relationship. Secondarily, we also intend to respond to the issue discussed in Chapter 1 regarding the sufficiency of the lexical knowledge associated with the sense inventory in existing lexical resources for WSD, in the light of the results obtained in this chapter.

6.3 Studying Target Nature

Given the primary objective, we would in particular investigate the plausibility of separating words, or more precisely word senses, into fairly distinct groups according to their responses to disambiguation with different types of information. We call such groups *sense types*, in accord with the *Sense Type Classifier* (STC) in our Integrated Model of WSD as described in Chapter 3. The obvious problem is "what" these different types will be. One way is to take a shortcut by using readily available sense classifications and see if they can account for certain disambiguation results. Meanwhile, how senses are best characterised by different types of lexical information can be inferred from the differential responses of the former to the latter in WSD, which would be independent of whether or not we can name the sense types. Therefore merely identifying such response differences can be another starting point.

In Section 6.3.1, we outline the two approaches to investigating target nature which we follow in this chapter, and in Section 6.3.2, we state some general hypotheses on the effect of polysemy, dispersion, and susceptibility on WSD.

6.3.1 Our Approaches

Target nature is thus studied in two directions here, which we call the *Denotative Approach* and the *Connotative Approach* respectively.

The Denotative Approach

A taxonomic classification like WordNet's captures what is referred to as the "denotation" of words.[1] With such a classification, senses already fall into different groups (i.e. different branches from a small set of top nodes) according to what they refer to in the real world. Consequently, we can use these top nodes in WordNet to approximate sense types, and perform some post-hoc analysis of WSD results (as those in Chapter 5) to see if the correctness/incorrectness of disambiguation correlates to these sense types in any way. This approach will be pursued in Section 6.4.

The Connotative Approach

The other direction from which we study target nature is based on the normal understanding of "connotation" of words.[2] In this case, we start with no particular named sense types. Instead, by testing for the differential responses of words (with their expected senses) to various types of disambiguation information, we try to infer some underlying attributes or commonality among the senses to account for the observed differences. For this kind of study, the validity of the groupings depends very much on the amount as well as the variety of the test data, which means that manageability and generality are of particular concern. To this end, we suggest modelling

[1] According to Lyons (1981), denotation is the means by which a lexeme is related (in the relevant way) to the outside world.

[2] Connotation refers to some association between a lexeme and its typical contexts of occurrence: "That frequent use of a word or phrase in one range of contexts rather than another tends to create a set of associations between that word or phrase and whatever is distinctive about its typical contexts of occurrence." (Lyons, 1981: p.150)

WSD in the format of cloze tests. The background of cloze tests and our motivation of introducing them in the study were discussed in Chapter 3. The detailed account of this approach is given in Section 6.5.

6.3.2 General Hypotheses

Our general hypotheses regarding the effect of the three aspects of target nature: polysemy, dispersion, and susceptibility, are stated in H1, H2, and H3 respectively:

H1 The difficulty of WSD should correlate positively with the number of candidate senses. For instance, we expect it is more difficult to disambiguate w_1 than w_2 if w_1 has 10 senses but w_2 has only 2.

H2 The difficulty of WSD should correlate negatively with the dispersion of the candidate senses. In other words, it would be less difficult to disambiguate w_3 than w_4 if the former has two very different senses but the latter has two very closely related senses. As mentioned in Section 6.1, polysemy and dispersion may not be independent of each other. "More senses" does not necessarily mean "more fine-grained senses". For instance, if the additional senses (for a group of senses) cover a bigger conceptual space, the dispersion among all these senses may increase correspondingly. However, if the conceptual space if fixed, listing more senses within it may result in more fine-grained senses.

H3 Senses should fall under fairly distinct groups (may or may not be named sense types) according to how effectively they can be disambiguated by different types of information. Past studies like Wilks et al. (1989) and Leacock et al. (1998) already pointed out some difference between words of different POSs. Our results in the last chapter further suggest that such differential susceptibility can also exist even among words with the same POS. Further, the observed difference among text types also suggests that the distribution of such sense groups in texts may vary with text types.

The "difficulty" of WSD (and its quantification) is often inferred a posteriori from disambiguation results. Thus the way in which results are derived limits the possibilities for subsequent analysis. On the one hand, if the test data is made up of many instances for a small set of test words, then task difficulty for each test word can be more properly correlated with precision and recall which are computed separately for each word. On the other hand, if words in running texts are disambiguated, and precision and recall are computed from all test instances to give an overall percentage, then only descriptive statistics can be given. Our data belong to the second type.

6.4 The Denotative Approach

In this section, we further analyse our disambiguation results in Chapter 5 with respect to the various aspects of target nature. At the word level, we focus on the effect of polysemy and dispersion on the difficulty of disambiguation using the narrow type of semantic relation (as we did in the last chapter). At the sense level, we start with the WordNet classification of senses and use the classes at its most general level, i.e. clusters of words covered by the top nodes, to represent our sense types, and hence study the types' susceptibility to different kinds of disambiguation information.

The operationalisation of the various aspects of target nature in the Denotative Approach is described in Section 6.4.1, followed by an account of our post-hoc analysis in Section 6.4.2. The findings from the analysis will be discussed in Sections 6.4.3 and 6.4.4. Also, we will show some text examples in Section 6.4.5 to illustrate the effect of target nature on WSD.

6.4.1 Operationalisation of Variables

This section describes the measures we define for studying the various aspects of target nature: polysemy, dispersion, and susceptibility. These definitions give us some leverage to start the investigation.

Polysemy

For polysemy, we measure the number of candidate senses at three levels of granularity, defined as follows:

- *Coarsely-grained* (N_c): the number of top nodes covering all the candidate senses of a word in WN16, e.g. 7 of the 10 senses for "space" descend from the top node "abstraction" and 3 from "entity", so $N_c = 2$.
- *Fine-grained* (N_f): the number of candidate senses in the same cluster as the expected sense in WN16, e.g. if the expected sense is space/2 (topological space), then as there are 7 senses of the word "space" in this WN16 cluster, $N_f = 7$.
- *Mixed-grained* (N_m): the number of all candidate senses of the word as listed in WN16, e.g. "space" has a total of 10 senses in WN16, so $N_m = 10$.

Dispersion

For dispersion, we measure the closeness of the candidate senses by the following:

- *Intra-Cluster Distance* (D_{in}), defined as:

$$D_{in} = \frac{\sum_{i=1, i \neq e}^{N_f} Sim(s_e, s_i)}{N_f - 1}$$

  where $s_e = \textit{expected sense}$ and $s_i = i^{th}$ *sense in the same cluster*. The similarity $Sim(x, y)$ is as defined in Rada et al. (1989) (see Chapter 5).

- *Dispersion Ratio* (DR), defined as $\frac{N_c}{N_m}$, a ratio between 0 and 1 which reflects the relative dispersion, independent of the total number, of senses. For instance, 2 senses in 2 clusters and 9 senses in 9 clusters both give 1 for this measure.

Susceptibility

For susceptibility, we need some hypothetical underlying construct (sense types) to account for the different responses of the test instances to disambiguation. In the current approach, we define sense types as the most general classes in the WordNet IS-A hierarchy. There are nine top nodes in WN16:

Entity (Ent)	Human Action (Act)
Psychological Feature (Psy)	Group (Gro)
Abstraction (Abs)	Possession (Pos)
State (Sta)	Phenomenon (Phe)
Event (Eve)	

The above definitions are based on the information available from WN16 and may not be perfect. However, since we actually did the disambiguation using WN16 senses, it is reasonable to take the sense properties in WN16 for granted (e.g. sense granularity, conceptual space) and define the measures for target nature from them.[3]

6.4.2 Experiment NAT-1

In Chapter 5, we observed that the effectiveness of the narrow type of semantic relation for disambiguation and the MFA for expanding context varied with text types. In particular, the absolute recall was found to be much higher for scientific writings (J) than for fictional texts (K). Also, the difference between *mf3000c* (MFA with full-text context) and *wb3000c* (WBA with full-text context) was much smaller for J than for K, suggesting that the MFA was more effective for approximating large contexts in the former. To account for such differences, we shall perform a post-hoc analysis of those disambiguation results with respect to the various aspects of target nature.

Again, we will focus on categories J (scientific) and K (fiction), two intuitively very different types. The results for J reported below are based on the average of the two random samples J1 and J2 (see Chapter 5). Additionally, we will also show the results from analysing categories A (Press: Reportage) and L (Fiction: Mystery and Detective), which are similar to categories J and K respectively, as further support for our arguments. However, since A and L did not have as many texts for comparison, they will only be subsidiary in our discussion.

Results from *wb70c* and *mf3000c* in Chapter 5 were analysed to explore why the MFA seemed more beneficial to texts in categories A and J than those in K and L. The outcome for each test instance in the texts under these conditions was first identified as either "correct" ("y"), "incorrect" ("n"), or "no response" ("z"). Descriptive statistics were computed from these results in relation to the variables described in the last section. Results regarding the effect of polysemy and dispersion are discussed below in Section 6.4.3, and those regarding susceptibility in Section 6.4.4.

[3]In doing so, our analysis is potentially confounded since our disambiguation information also came from WN16. However, since the two methods we compare (WBA and MFA) both used the same type of disambiguation information, they should be equally confounded, and this introduces no interference to our analysis of their difference.

6.4.3 Polysemy and Dispersion

Tables 6.1 and 6.2 show the average polysemy and dispersion (in various measures) for different outcome categories in *wb70c* and *mf3000c* respectively. As pointed out, the difficulty of the task can be inferred retrospectively from the disambiguation outcomes, and how these outcomes are derived may lead to different interpretations. With our materials, consequently, we can only view the difficulty of the task dichotomously, i.e. it is relatively easy if the instance is correctly disambiguated ("y"), and relatively difficult otherwise ("n").[4]

Cat	Outcome	N_c	N_f	N_m	D_{in}	DR
A	y	2.63	2.83	5.43	0.1752	0.5401
	n	3.36	3.28	7.74	0.1558	0.5079
	z	2.20	2.39	4.27	0.1645	0.5889
J	y	2.73	2.55	5.33	0.1570	0.5984
	n	3.19	2.97	6.98	0.1470	0.5334
	z	2.38	2.30	4.32	0.1521	0.6276
K	y	2.40	3.36	5.50	0.1642	0.5126
	n	3.05	4.11	7.91	0.1509	0.4440
	z	2.20	2.74	4.54	0.1447	0.5576
L	y	2.50	3.32	5.56	0.1673	0.5296
	n	3.12	3.97	7.97	0.1500	0.4557
	z	2.35	2.82	4.84	0.1463	0.5589

Table 6.1 *Effect of Polysemy and Dispersion on WSD in wb70c*

Cat	Outcome	N_c	N_f	N_m	D_{in}	DR
A	y	2.51	2.68	5.00	0.1826	0.5691
	n	3.21	3.11	7.40	0.1589	0.4988
	z	2.30	2.56	4.57	0.1504	0.5798
J	y	2.69	2.47	5.07	0.1532	0.6068
	n	3.10	2.98	6.88	0.1494	0.5329
	z	2.40	2.25	4.25	0.1529	0.6368
K	y	2.25	3.07	4.91	0.1585	0.5379
	n	2.97	4.11	7.70	0.1487	0.4438
	z	2.22	2.67	4.50	0.1491	0.5631
L	y	2.29	3.29	5.17	0.1651	0.5095
	n	3.09	4.11	8.01	0.1454	0.4540
	z	2.40	2.61	4.70	0.1504	0.5783

Table 6.2 *Effect of Polysemy and Dispersion on WSD in mf3000c*

We can see that very similar patterns of task difficulty with respect to polysemy and dispersion were found for both test conditions. First, at all levels of sense granularity (N_c, N_f, N_m), targets are on average more polysemous for "n" than for "y". At the mixed-grained level, for instance, there are about two more candidate senses for the former. In other words, the more senses a word has, the more difficult the task

[4]Since the precision and recall are not for individual words, the same word might be easy at one occasion but difficult elsewhere. However, we are only looking at the average results for general patterns.

tends to be. Although we cannot tell the exact shape of this variation, the average pattern is in line with hypothesis H1.

Second, evidence is also present to support hypothesis H2 regarding the effect of dispersion on WSD. The dispersion ratio (DR) suggests, across the board, that words with more dispersed senses have an advantage for being successfully disambiguated.[5] The intra-cluster distance (D_{in}) also shows a similar effect of dispersion, although the contrast of D_{in} between different outcome categories is not very great because the measure is constrained by the total depth of the WordNet hierarchy. Nonetheless, it is evident that the "y" group tends to have targets with senses further separated within the same cluster.

More importantly, the above patterns are observed across (i) the various text categories, and (ii) the two test conditions. The significance of such generality is that it has left the observed WSD performance, which varied with context sizes as well as text types, unexplained. If we say that it is the target nature which makes it more difficult to disambiguate certain words, it is then obvious that, given the results in the above analysis, the difficulty is not only dependent on polysemy and dispersion. Moreover, so far we have not discussed the "no response" ("z") outcome. In fact the patterns shown by this group are not clear. For example, N_m for "z" is usually closer to that for "y", but D_{in} for "z" is always similar to that for "n". The presence of "z" again suggests that while the same type of information may not be equally effective for disambiguating all target words, the distinction is not likely to be entirely a function of polysemy and dispersion of the words. What other properties underlie this distinction is precisely what we hope to find out from the third aspect of target nature, i.e. susceptibility.

6.4.4 Susceptibility

In this section, we will define sense types using the nine top nodes in the IS-A hierarchy in WN16, as listed in Section 6.4.1. When we refer to a particular top node as a sense type in the following discussion, we enclose it in double quotes (e.g. "entity"). We start by analysing the average distribution of sense types in different text categories, as shown in Figure 6.1.

The most notable feature in Figure 6.1 is probably the variation in the size of the "entity" slice. It occupies only a quarter of all nouns in A and J, but as much as half of all nouns in K and L. Although "entity" (e.g. pulley, mosquito)[6] dominates the WordNet lexicon, the variation observed does suggest that words are used in quite different senses with different kinds of text. Another frequent type is "abstraction" (e.g. molecular_weight, viscosity). It is particularly abundant in J, occupying 30% of all noun instances, and slightly less so in A (26%), but it occupies only about 20% in K and L. Nouns in A are also quite largely composed of "human action" (e.g. buildup, disassembly), up to 21%, which is about three times as much as the corresponding slices in K and L. Also, although in relatively small proportions, J have four times as many "phenomenon" senses (e.g. surface_tension, weather) than others.

[5] This is in fact not as trivial as it might seem because even if N_c is always hooked to N_m, there is every possibility that the relationship is not linear. Thus N_m and N_c on their own do not imply dispersion because more clusters does not necessarily mean greater dispersion.

[6] The word examples for each sense type given in this section are monosemous.

Compared to the above-mentioned sense types, the rest (e.g. "psychological features", "possession", "state", etc.) appear relatively insignificant. Not only do they account for few instances, but their quantities are also rather constant across text categories.

All in all, the patterns in Figure 6.1, though not undisputedly distinct, do support the part of hypothesis H3 regarding the dependency of sense type distribution on text category.

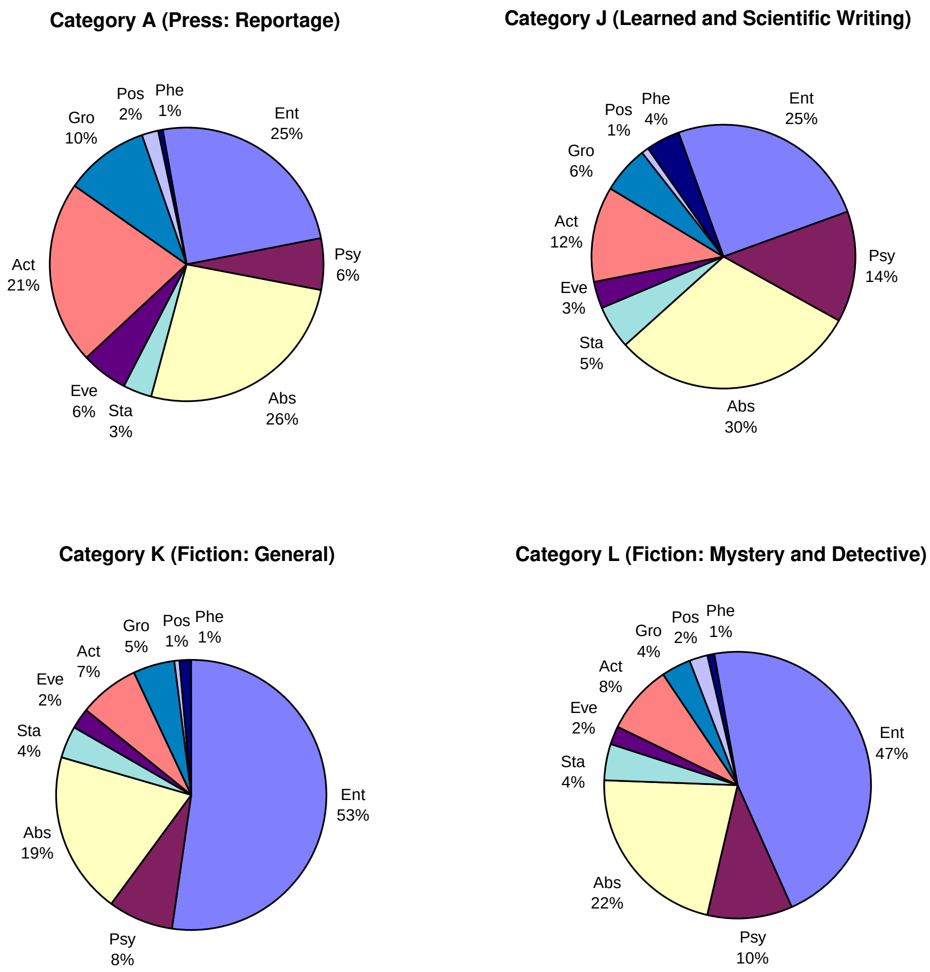


Figure 6.1 *Sense Type Distribution in Various SEMCOR Categories*

We can now try to account for the disambiguation results obtained in Chapter 5, summarised below:

1. In general, absolute recall and precision are higher for category J (and A alike) than K (and L alike).

2. From *wb70c* to *mf3000c*, precision is maintained at a similar level for J (and actually increased for A in *mf3000c*), but precision drops for K and L.

3. From *wb70c* to *mf3000c*, recall for A and J has increased but that for K and L remains similar.

Figures 6.2(a) and (b) show the correctness and error distribution respectively, with reference to sense types and text categories, in *wb70c*. Figures 6.3(a) and (b) show similar information but in *mf3000c*.

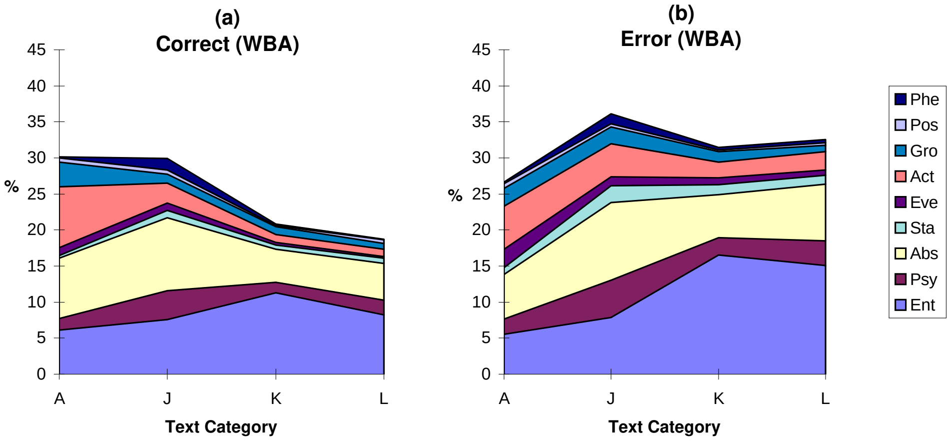


Figure 6.2 *Correctness and Error Distribution by Top Nodes in wb70c*

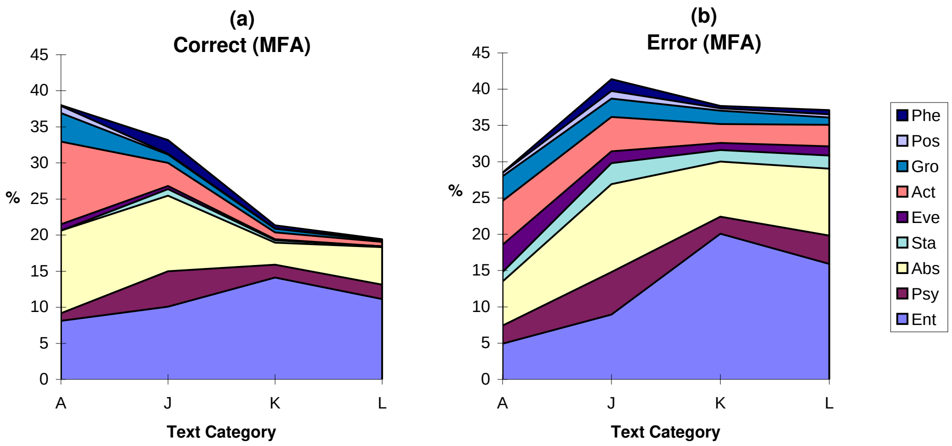


Figure 6.3 *Correctness and Error Distribution by Top Nodes in mf3000c*

Comparing Figures 6.2(b) and 6.3(b) for the error distribution, we see that generally the total error for every text type is about 5% higher in *mf3000c* than in *wb70c*.

This may have been a result of the increasing noise which accompanied expanded contexts. More interestingly, errors appear more or less proportional to the distribution of sense types under both test conditions. This means two things: (i) the number of errors for each sense type is proportional to the total number of that type in the same text category (for instance, in Figure 6.2(b), "abstraction" almost equals "entity" in A, but is about 2/5 of "entity" in K); and (ii) the proportions are also maintained across text categories (for instance, K and L have about twice as many errors for "entity" as A and J).

However, these proportions do not hold for the correct instances, as shown in Figures 6.2(a) and 6.3(a). First, at both test conditions, while the proportions of sense types within a text category seem to be maintained, the proportions across text categories do not hold. Most obviously, for example, K and L do not have many more "entity" senses correct than A and J, although the former have twice as many "entity" senses overall.[7] The number of correct senses for other sense types like "psychological feature", "abstraction", and "human action" in K and L are also less than expected (if it was in proportion with other text categories). This is probably why the absolute recall and precision are always lower in K and L.

As mentioned, from *wb70c* to *mf3000c*, total errors increase rather uniformly across text types. For category K, errors mostly increase in "entity" and "abstraction", but since total correctness does not change much[8], this results in similar recall but a drop in precision. For category J, on the other hand, total errors increase as with other text types, but correctness for senses of "entity", "psychological feature", and so on, also increases in *mf3000c*. It therefore leads to a rise in recall but precision remains similar.

The above analysis also suggests that the role of sense types is as follows. First, as mentioned earlier, related words in a coherent discourse are not necessarily near neighbours. Moreover, words do not only relate paradigmatically (as is mostly captured in WordNet) to one another within a text. Therefore it does not follow that frequency of sense type is always capitalised on, given the way we define sense types and the disambiguation information we use. However, the improvement for individual sense types in *mf3000c* (e.g. "human action" in A, "psychological feature" in J) or their degradation (e.g. "abstraction" in K) does show that a wider context is useful to uncover related words for certain sense types in certain text categories.

Second, following our presumption that fictional texts are less strongly related to particular subject areas than scientific writings, this property may correspondingly be reflected in the monosemous nouns as well. In other words, the monosemous nouns in K and L may possess relatively weaker selective power towards the rest of the ambiguous words in the texts. Moreover, if this is the case, other reliable disambiguation information must be sought and disambiguation has to be made sensitive to the target words and the text types in question.

So far, top nodes in the WordNet hierarchy do not seem to make very satisfactory definitions for sense types for our purpose, because no contrasting pattern could be found for them with respect to their correctness and error distribution. They therefore

[7] This outcome of "entity" may be a consequence of the abundance of sub-branches, so senses under this type are relatively diverse.

[8] There are subtle changes with individual sense types, e.g. fewer "abstraction" senses but more "entity" senses are correct in *mf3000c*. See implication below.

do not allow us to say conclusively how sense types relate to the difficulty of disambiguation. Nevertheless, through the above analysis, we can identify a significant distinction among different text categories in terms of their sense type composition, and can account for the disambiguation results obtained in Chapter 5 in terms of these sense types. In the next section, we will show some text examples to illustrate the contrast suggested between scientific writings and fictional texts.

6.4.5 An Example

Table 6.3 compares two texts from J and K respectively by their noun composition and results from the two test conditions (*wb70c* and *mf3000c*). The text br-j03 is a scientific writing about physics experiments with viscometers and br-k22 is a fictional text in which the author writes about his landlord.

		br-j03 (physics)		**br-k22** (fiction)	
WBA Results	Precision	53.14%		42.38%	
(*wb70c*)	Recall	35.18%		23.44%	
MFA Results	Precision	47.02%		39.62%	
(*mf3000c*)	Recall	39.34%		23.08%	
N (Poly|Mono)	Total	361	100	273	105
	Distinct	129	50	191	91
Average no.	Total	6.07	–	5.29	–
of senses	Distinct	6.09	–	4.85	–
Noun Sense	Ent	36.00%	24.38%	50.92%	56.19%
Composition	Psy	5.00%	8.86%	6.23%	1.90%
(Poly|Mono)	Abs	34.00%	31.30%	24.54%	22.86%
	Sta	0.00%	3.05%	5.49%	8.57%
	Eve	1.00%	7.48%	2.93%	0.00%
	Act	2.00%	4.16%	4.03%	3.81%
	Gro	4.00%	0.00%	2.20%	2.86%
	Pos	0.00%	0.00%	2.56%	1.90%
	Phe	18.00%	20.78%	1.10%	1.90%
Examples of Monosemous Nouns		velocity, ideal_gas, potential_energy, manometer, thermodynamics		husband, hallway, swimming_pool, stranger, autumn, desk, hotel, wife	
Polysemous	distance	1	1	1	2
Nouns	end	2	1	1	1
common to	function	4	1	1	2
both texts	hole	1	4	1	3
(Num|Sen)	literature	2	1	1	1
	manner	1	1	1	2
	means	1	2	1	1
	part	1	1	2	7, 12
	plane	1	2	1	1
	property	2	2	1	1
	rise	1	2	1	6
	time	2	3	2	1, 4
	way	1	1	1	2
	work	5	1, 7	3	1, 3

Table 6.3 *An Example Contrasting Two Text Types*

Apart from the difference in results with *wb70c* and *mf3000c* as discussed earlier, we observe the following differences regarding noun composition which also mark the contrast between the two text types. First, nouns are far more repetitive in br-j03 (physics) than br-k22 (fiction), as seen from the differences between the number of distinct nouns and that of all noun instances. Second, it appears that the nouns which both texts share tend to be relatively domain-neutral, common, and multi-purpose. They are also relatively more polysemous. We find that the number of their senses ranges from 3 to 14, averaging at 7.43, which is on the high side. Moreover, the two texts tend to use these words in different senses, and br-k22 tends to use the same word in different senses within the text (compare the 4th and the 6th columns in the lower half of Table 6.3). In other words, the "one sense per discourse" assumption (Gale et al., 1992a), quite apart from other criticisms, cannot be assumed to hold when the text category varies.

In addition, we mentioned that the success of the MFA depended to some extent on the quality of the monosemous nouns in a text, and hypothesised that those in fictional texts might be weaker in this respect. We have included some examples of monosemous nouns from br-j03 and br-k22 in Table 6.3. Strictly speaking, br-k22 has more monosemous nouns to start with, but they do not give any advantage to the MFA. This shows that the monosemous nouns in br-k22 (as compared to those in br-j03) may not be sufficiently useful for driving the MFA. In general, the monosemous nouns in br-j03 appear much more domain-suggestive than those in br-k22, and are more closely bound to their conventionalised contexts as expressed in WordNet. This possibly explains why, when using the same lexical resource and the same type of contextual information, disambiguation turns out to be more difficult for words in fictional texts. This interaction among *Lexical Resource*, *Contextual Information*, and *Nature of Target Words*, will be further explored in the next section, where we study the problem through the Connotative Approach.

6.5 The Connotative Approach

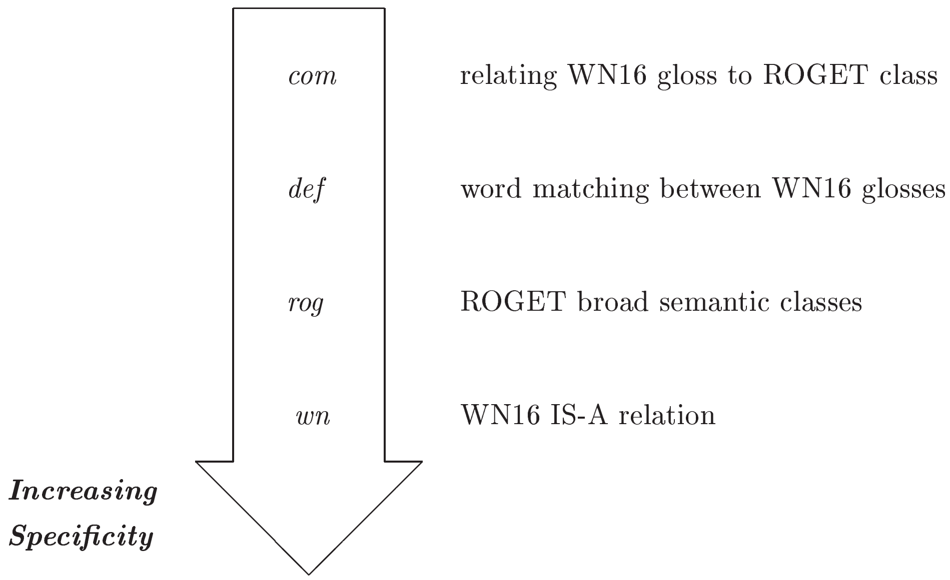
In the last section, we saw the difference in the distribution of sense types (defined with WordNet top nodes) for various text categories and noted how the disambiguation performance in Chapter 5 correlated with this difference. However, we have not got very far regarding the information susceptibility of sense types. Since hits and misses spread all over the various sense types, the contrast among the sense types is not big enough for any meaningful conclusion to be drawn. In this regard, we try to tackle the problem from the opposite direction. We start with no particular named sense types, but derive plausible clusters of senses according to their response patterns to different types of disambiguation information in WSD. Then retrospectively we look for the underlying characteristics in these clusters.

In Section 6.5.1, we will explain the design of the Connotative Approach and the rationales behind, followed by a report of an experiment in Section 6.5.2. We will then discuss in Sections 6.5.3 and 6.5.4 what we learn from the experiment, and in Section 6.5.5 the prospect of the approach.

6.5.1 Design and Rationale

The idea of the Connotative Approach is to find out from a sample of nouns (i) if senses do fall into different clusters based on their responses to particular types of disambiguation information, and (ii) what proportion of these clusters are capable of being disambiguated by exploiting the available lexical resources. The sample should thus represent some commonly used nouns in everyday English without any bias towards specific domains.

For the first purpose, we need to test different types of disambiguation information and find patterns from their effectiveness on disambiguating the noun samples. In Chapter 5, we have looked at various types of semantic relation obtained from our Integrated Lexical Resource (ILR), and organised these semantic relations of varied specificity into a spectrum as below:



This spectrum of semantic relations gives us some leverage in the investigation of information susceptibility. For the second purpose, however, we need further information which is outside our ILR. So we introduce two types of information obtained from corpus data, namely simple collocation (*col*) and a flexible type of POS bigram (*f-2g*).

According to the above specificity scale, *col* would be expected to cover semantic relations broader than *com*, as it may or may not include information contained in the ILR. In other words, it resembles "free association", where the relations governing the associations are not readily predictable.[9] On the other hand, *f-2g*, unlike the others, considers non-semantic information which is usually not covered by the lexical semantic information in lexical resources, including our ILR.[10] With the information

[9] For example, Rapp and Wettler (1991) and Wettler and Rapp (1993) modelled free association with co-occurrence data from corpora.

[10] Though some dictionaries like LDOCE might contain usage tips such as "*usu. followed by in*" or "*esp. in* **in the event**", and so on, for one thing this is not usually the focus when researchers make

additional to the types obtainable from the ILR, our spectrum of semantic relations is modified and embedded in a wider range of lexical information as below:

Nature		**Information**	**Source**
Semantic		*col*	Corpus
Semantic		*com*	ILR
Semantic		*def*	ILR
Semantic		*rog*	ILR
Semantic		*wn*	ILR
	Increasing Specificity		
Syntactic		*f-2g*	Corpus

Consequently, by applying these different types of lexical information in disambiguation, the hypothesis that different groups of senses are best characterised with different types of information can be put to test. Grouping the test samples with respect to their responses to the various types of information can reveal the differential information susceptibility of the senses. Also, to find out how sufficient the lexical information in the ILR is for WSD, we can analyse the results according to the following "formula":

$$\left\{ \begin{array}{c} What\ can\ be \\ achieved\ with \\ ILR \end{array} \right\} = \left\{ \begin{array}{c} All \\ Samples \end{array} \right\} - \left\{ \begin{array}{c} No\ Response \\ OR \\ All\ Wrong \end{array} \right\} - \left\{ \begin{array}{c} Correct \\ ONLY\ with \\ col\ OR\ f\text{-}2g \end{array} \right\}$$

6.5.2 Experiment NAT-2

Materials

As introduced in Chapter 3, we will bring in the cloze test format for studying target nature, based on both modelling and processing grounds. In this experiment,

implicit information from dictionaries explicit, and for the other, this kind of "tips" only covers the extremely distinct usages.

14 short texts, each containing about 200 to 300 words with blanks to be filled in, were chosen from real practice papers. Only blanks for polysemous nouns[11] were kept for the test and others were filled with the known correct answers during the preparation. The 14 texts thus had 60 blanks, giving us a sample of 60 nouns (in their expected senses) to study and analyse. The average number of candidate senses for the samples was 5.4. The texts are shown in Appendix D, and here is an example blank (the word in bracket is the known preferred answer):

... The <u>(rhythm)</u> of the music also helps people to remember ...

Method

Different combinations of context sizes and context realisations were tested for disambiguating the targets at the cloze blanks. The combinations are summarised below:

Realisation	Size			
	±3	±5	±50	Full
col		√	√	
com				√
def		√	√	√
rog		√	√	√
wn		√	√	√
f-2g	√			

Full-text context will again be approximated with the MFA. The realisations *com*, *def*, *rog*, *wn* are used in the same way as in Chapter 5 (refer to Experiments MFA-2, WBA-2, and REA-1 where appropriate). The other two types of information, *col* and *f-2g*, were not from the ILR but obtained statistically from corpus data. Their extraction and application are explained below.

Handling *f-2g* and *col*

Parameters for *f-2g* and *col* were estimated from the SEMCOR data which have about 200K sense-tagged content words. About 90K of them are nouns. However, since data sparseness was particularly pertinent, the parameter estimation was slightly modified from the standard approaches in attempt to improve the situation.

First, close relatives of the targets were used to increase the training data. Where there were too few or even no samples for a particular sense in the corpus, samples for its synonymous or direct hypernym senses (according to WN16) were also extracted from SEMCOR and added to the training data. For example, hands/2 has only one sample in SEMCOR, so we added samples from manpower/1 and work_force/1 (synonyms of hands/2) to our training data.

Second, senses without sufficient training samples were discarded. Given that we only used a 200K corpus, unlike other studies using a 10M corpus, many senses were

[11] The expected answers were checked to be present in WordNet 1.6.

bound to have too few occurrences. In this study, only the candidate senses which had five or more occurrences in the training data were considered.[12]

Third, the estimation of the left- and right-context bigram was made more flexible. Under normal circumstances, one would estimate $p(t_{i-2}t_{i-1}|s_i)$ for a left-context bigram, where s_i is the sense at position i, and t_{i-2} and t_{i-1} are the POS tags of the words at positions $i-2$ and $i-1$ respectively. In this study, however, we look for all possible bigrams within a range of three consecutive positions on the left of the target word. So we estimate all $p(t_j t_k|s_i)$ where $i-3 \leq j < k < i$. Similarly for the right context, we estimate all $p(t_j t_k|s_i)$ where $i < j < k \leq i+3$. We call these bigrams *flexi-bigrams*, hence the abbreviation *f-2g*. In this way, we intend not only to identify distinct syntactic patterns for particular senses, but also to accommodate modifiers or minor stylistic variations in the training samples. Consequently the idiosyncrasy resulting from data sparseness can be lightened.

Subsequently in the testing phase, we combine evidence from both the left and right context. As in Leacock et al. (1998), given s_i and the information at positions $-k$ to k (i.e. $c_{-k}, ..., c_k$),

$$p(s_i|c_{-k}, ..., c_k) = \frac{p(c_{-k}, ..., c_k|s_i)p(s_i)}{p(c_{-k}, ..., c_k)}$$

and the independence assumption leads to the approximation of

$$p(c_{-k}, ..., c_k|s_i) = \prod_{j=-k}^{k} p(c_j|s_i)$$

In our case, we only have two positions, namely left and right, and we further replace $p(c_j|s_i)$ with the best supporting flexi-bigram:

$$p(c_j|s_i) = \max p(t_j t_k|s_i)$$

where $p(t_j t_k|s_i)$ is as defined earlier.

The advantage of flexi-bigrams will be obvious with an example. Suppose we have the following training samples for way/2:

There is no	way ...	(EX BE DT)[13]
There is no easy	way ...	(BE DT JJ)
They could see no	way ...	(MD VB DT)

If we consider trigrams, we only find each pattern once, similarly for the immediately preceding bigrams. However, it is apparent that these samples embed a likely pattern which is "there is ...(some)... way ...".

[12] Although such modification of the training data may invoke certain hidden bias in the disambiguation process, we assume it is random enough across all samples so as not to invalidate the study.

[13] The POS syntactic tags in SEMCOR were condensed to the following set in our experiment: BE= "be", CC=Coordinating conjunction, CD=Cardinal number, DO= "do", DT=Determiner, EX=Existential "there", HA= "have", IN=Preposition, JJ=Adjective, MD=Modal, NN=Noun, PRO=Pronoun, RB=Adverb, RP=Particle, TO= "to", UH=Interjection, VB=Verb, XX=Others.

Now consider a test sample: "There is absolutely no way ...", in which case we would end up in an unobserved trigram (BE RB DT) or bigram (RB DT). Nevertheless, with flexi-bigrams, we obtain probabilities for EX BE, BE DT, EX DT, BE JJ, DT JJ, MD VB, and MD DT. When we look for

$$\max \left\{ \begin{array}{c} p(BE\ RB) \\ p(BE\ DT) \\ p(RB\ DT) \end{array} \right\}$$

the likely pattern BE DT for the test sample can be identified.

For *col*, we follow Yarowsky's (1992) approach for estimating $p(word|s_i)$ where $word$ is any noun or verb (in root form) occurring within a window of ±50 words from s_i.

It turned out that only 34 out of the 60 test samples had enough training data for the expected sense (and at least one other candidate sense); and for these 34 samples, the average number of qualified candidate senses was reduced from 5.8 to 3.2. Since the prior probabilities, $p(s_i)$ in the above equation, might have been affected by the way we produced the training data, two conditions were tested. One assumed uniform prior probabilities of all senses, and the other accounted for them as observed. We call these conditions *up* and *np* respectively. As for context size, *f-2g* was tested with ±3-word contexts only, while *col* was tested with ±5-word and ±50-word contexts. In addition, for *col*, we separately tested co-occurrences of nouns only, verbs only, and nouns plus verbs.

Aside 1: Results for *f-2g*

For *f-2g*, the accuracy was 13/34 with *up* and 17/34 with *np*. Samples which were correctly disambiguated by *f-2g* under *up* and still so under *np* suggest that prominent patterns for the usage of the corresponding senses were successfully captured. This is especially important if the expected sense is not the first sense of the word, since the latter usually has higher prior probability and therefore a built-in advantage in *np*. Our results suggest that even though all samples are nouns and syntactic knowledge has constrained the variability of usage patterns, some word senses do exhibit very specific patterns which provide useful information for disambiguation. Although such distinctness may only affect a small proportion of all noun senses, this is still useful for sense elimination. On the other hand, samples which failed at both *up* and *np* imply the absence of significant patterns associated with the senses. The observed syntactic structures for them are likely to be confused with those for other candidate senses. We give some examples below.

For instance, event/2, which means "a special set of circumstances", is usually used in "in that event ...", "in the event of ...", and so on. The pattern is reflected by the relatively high probability for IN DT in the left context from the training data. The same pattern (IN DT) also has higher probability and mutual information in event/2 than in its competitor event/1, as shown below (p=probability of the flexi-bigram, I=mutual information[14]):

[14] See Church and Hanks (1990) for details. In their study, owing to the corpus size, significant

event/2

	f-2g	*p*	*I*
→	IN DT	0.2778	1.5850
	NIL DT	0.1667	1.8480
	⋮	⋮	⋮
	CC DT	0.0556	1.8480
	CC IN	0.0556	1.5850
	⋮	⋮	⋮

event/1

	f-2g	*p*	*I*
	DT JJ	0.0833	1.1508
	NN IN	0.0833	1.7840
	⋮	⋮	⋮
→	IN DT	0.0444	0.4366
	NN DT	0.0444	0.4366
	⋮	⋮	⋮

But on the contrary, for example, none of sun/1 and sun/2 demonstrate any distinct syntactic pattern. This is obvious if we compare some observed left contexts for the two competing senses and the associated probabilities:

sun/1

f-2g	*p*	*I*
IN DT	0.1839	1.5481
NN DT	0.1149	1.2571
DT JJ	0.0690	1.6356
IN JJ	0.0345	1.3555

sun/2

f-2g	*p*	*I*
IN DT	0.1667	2.0589
NN DT	0.1667	1.8365
DT JJ	0.0333	1.3219
IN JJ	0.0333	1.3219

It can be seen that not only were similar patterns observed for both candidate senses, but also the relative frequencies were comparable. Hence words like "sun", which has candidate senses used in similar syntactic patterns, are not predisposed to

associations often yield an $I \gg 1$. In our case, however, the corpus may be too small to give significant estimates, and the magnitude is expected to shrink considerably.

syntactic characterisation and senses of such words are therefore not susceptible to disambiguation with syntactic information.

Under normal circumstances, when there are sufficient training data, we can expect more reliable prior probabilities of individual senses. However, in the present situation, it is helpful to analyse more thoroughly the effect of the prior probabilities, i.e. the difference between *up* and *np*.[15] The outcomes in *up* may be different from those in *np* for various reasons. One is that the effect produced by the observed prior probabilities is an aftermath of our interference in sampling. Another is that the bigram itself is similarly likely to associate with more than one candidate sense (including the expected one), and it is left to the prior probabilities to pick one. A third situation is that prior probabilities dominate and the most frequent sense is frequent enough to allow itself to be chosen at any time.[16] Only the first case has genuinely to do with the training data. For instance, the frequency of figure/9 was altered a lot because its synonym number/1 really outnumbered others by a large margin. Nevertheless, we observed that in general the prior probabilities were not as disturbed as was expected even when samples of close relatives were introduced into the training data. The other two cases, however, are like the example of "sun" and only show that *f-2g* is not a strong evidence for disambiguating the senses concerned. For example, level/1 was correctly disambiguated under *np*, but not *up*. A context of "... the level of ..." was found to be similarly probable for level/4 and level/1.[17] Nevertheless, there were far more samples for level/1 than level/4 (69 vs 5), and level/1 was therefore chosen at *np*.

Aside 2: Results for *col*

The tests with the information *col* produced results as shown in Table 6.4 (numbers are the correct cases out of the 34 qualified test samples).

Context Size	**POS and Prior Probabilities**					
	N Only		**V Only**		**N + V**	
	up	*np*	*up*	*np*	*up*	*np*
± 5	8	22	10	19	15	20
± 50	23	23	19	19	23	23

Table 6.4 *Results with col as Disambiguation Information*

In terms of context size with *col*, it can be seen that ± 50 is in general better than ± 5. One reason is that there may not be any other noun (or verb) found within such a small window as ± 5 words. Another possibility is, as we have argued, that

[15] Note the difference in handling prior probabilities in our case and in Leacock et al.'s (1998). Since ours was in any way a small corpus, the intact prior probabilities often required smoothing in one way or another. It therefore seems that including the additionally acquired samples in the prior probability computation may not be any less (or more) reliable than the original estimates. This is supported by our observation above.

[16] The fourth would of course be something genuine, though it would be relatively hard to tell with our deficient environment.

[17] level/1: a position on a scale of intensity or amount or quality, e.g. a high level of care. level/4: height above ground, e.g. the level of the water.

related words are not necessarily near neighbours. A longer context may be required to uncover them.

Unlike in *f-2g* where only two probabilities (i.e. from left and right *f-2g*) were combined, in *col* there were many more. The influence of the prior probabilities of individual senses is therefore inevitably weakened in *col*, and in fact is often overwhelmed at ±50-word context, as observed below. At ±5, where we may not find enough evidence from the small window, *np* did produce a substantial increase in correctness, compared to *up*. Performance at *up* improved as a larger window was used and sufficient evidence was considered, but then the difference between *up* and *np* vanished at ±50-word context.

In terms of the range of co-occurrence considered, a difference was found only at context size ±5. Taking verbs into consideration as well at this size may on the one hand make up for not finding a second noun within the vicinity; and on the other hand capture some of the subject-verb or verb-object collocation relations (e.g. "attract attention", "remain a mystery", and so on). However, at context size of ±50, including words other than nouns ceases to show any advantage. This suggests that where co-occurrence information is good evidence, nouns only would be strong enough for the purpose, given a reasonably wide context.

Main Results

In Table 6.5, we summarise WSD performance at each test condition, measured as the number of samples (out of 60) correctly disambiguated under each condition.[18] The figures for *col* are based on the conditions of "nouns only" and "uniform prior probabilities" (*up*), and those for *f-2g* are based on "uniform prior probabilities" (*up*).

Realisation	**Size**			
	±3	±5	±50	Full
col	–	8	23	–
com	–	–	–	20
def	–	6	15	2
rog	–	7	20	18
wn	–	7	17	10
f-2g	13	–	–	–

Table 6.5 *Summary of Results in Cloze Test Experiments*

Since performance at context size ±5 is relatively weak, we shall leave that out in our subsequent investigation. Analysing the raw results (documented in Appendix D), we can now identify the terms in our previous "formula" to see "what can be achieved with the ILR":

[18] The numbers for *f-2g* and *col* are in fact out of a total of 34 samples. However, our primary objective is to see how much can be achieved with the ILR, and all the information obtained from the ILR (*com*, *def*, *rog*, and *wn*) had been tested on all 60 samples. So even though *f-2g* and *col* were not applied to all 60 samples, we may only under-estimate the effectiveness of these two types of information, but not the usefulness of the ILR.

Total no. of samples	=	60	
All wrong / No response	=	11	(18.33%)
Correct ONLY with *f-2g*	=	3	(5.00%)
Correct ONLY with *col*	=	4	(6.67%)
Correct ONLY with *f-2g* AND *col*	=	4	(6.67%)

From these we can work out that:

Correct with ANY method	=	49	(81.67%)
Correct with SEMANTIC information	=	46	(76.67%)
Correct with SEMANTIC information in ILR	=	38	(63.33%)

For baseline comparison, we get 32 out of 60 samples correct (i.e. 53.33%) if the first sense from WN16 was chosen; and given that the average number of senses was 5.4, random choice would be expected to give about 18.5% accuracy. We can see that if different types of information could be properly combined, the WSD performance would be much better than that which any single type alone can give. The results will be discussed in more detail in the next few sections.

6.5.3 Fitting the Puzzle

We can now try to fit the puzzle together. In Experiment NAT-2, we set out to study target nature, in particular the information susceptibility of senses, by clustering our test samples according to their responses to the various types of disambiguation information. We also wanted to see how much the ILR can be expected to contribute to WSD.

The total number of correct disambiguations by any method (in the test) may have been underestimated because *f-2g* and *col* could not be applied to all our test samples. However, the boundary imposed by the ILR is rather clear. The fitted puzzle up to this stage is like Figure 6.4 (the sizes of the blocks are not proportional to the coverage of the clusters).

Figure 6.4 should be understood in the following sense: block F is still detached but it can potentially be merged with others except block B (note the borders). The reason for this is that block B corresponds to *wn* in our test and covers solely paradigmatic associations with the single IS-A criterion. Also, *wn* is the most direct mechanism and does not depend on other processes, unlike *rog*, for example, which depends on the sense linking procedure (see Chapter 4), or *f-2g* and *col*, which can be affected by data sparseness. As a result, samples which failed at *wn* would mean that they were genuinely not susceptible to it, so blocks B and F should not have anything in common. Hence samples which are covered by block F could be of two types. One is that they should in principle be captured by other types of information in our test but this was not discovered by our experimental conditions. The other is that they actually call for completely different types of disambiguation information (e.g. morphology, world knowledge). Inserting the actual samples, the blocks can be paired with the sense clusters as in Figure 6.5.

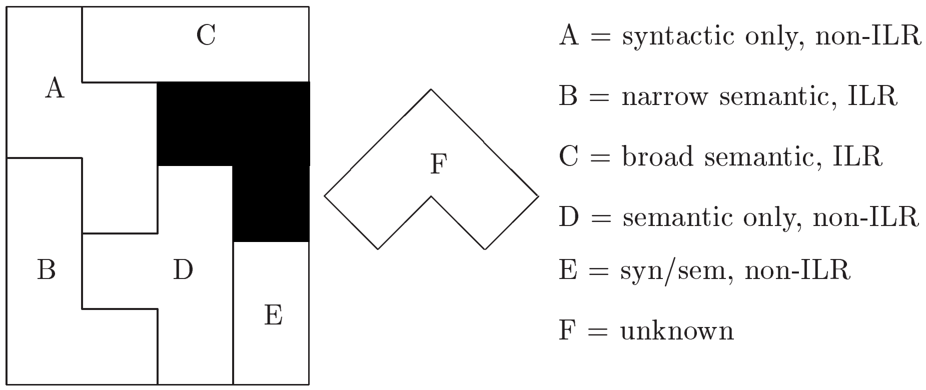


Figure 6.4 *Sense Clusters in terms of Information Susceptibility*

6.5.4 From Clusters to Sense Types

From what we learn in Experiment NAT-2, the prospect of WSD based on pre-defined, institutionalised senses is reasonably promising. Most of the test samples (over 80%, and possibly more), as naturally occurring as they can be, were successfully disambiguated in one way or another, as compared to the two baselines (53.33% with first sense and 18.5% with random choice).

The MFA in Experiment NAT-2 did not perform as well as it did in the last chapter. The major reason we think responsible for the difference is the size of the texts. The number of monosemous nouns may be much smaller in a 300-word text, rendering them less effective as drivers of the algorithm. In fact, with such short texts, the traditional WBA would be fine even for full-text context.

Figures 6.4 and 6.5 show that most noun senses in our samples can be disambiguated by one or more types of semantic relation with other noun senses (i.e. blocks B, C, D, and E in the figures). In particular, 21 out of 60 samples could be disambiguated with the narrowest paradigmatic association (i.e. block B). This does not necessarily mean that those cases are not susceptible to broader semantic relations, but for them the simple IS-A relation suffices. In fact it is quite interesting to note that all of our samples which were correctly disambiguated in *wn* were also correct in one or more other conditions. Seeing this, one might expect that we can simply do the job with the broadest semantic relation and do not need to bother with the finer distinctions within this "bag of relations". However, we believe the distinction is important because on the one hand, there remains the possibility that some (untested) senses may only work with *wn*. On the other hand, even if broader relations work, it is still useful to gather all plausible evidence so that senses can be more confidently selected.

The results from Experiment NAT-2 allow us to uncover more about the susceptibility of target words (or more precisely their senses) to various types of information, and thus the effectiveness of such information for disambiguating the words. The clusters in Figure 6.5 support the part of hypothesis H3 that senses can be grouped

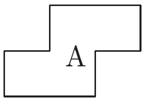
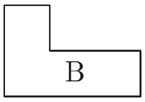
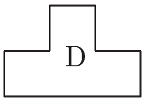
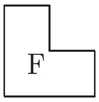
Cluster	**Test Samples**
 A	instant/1, way/2, place/1
 B	trick/5, performance/2, delay/1, rhythm/1, beam/4, sun/1, reflection/5, mirror/1, machine/1, aid/2, volunteer/2, precaution/1, figure/9, event/2, evidence/1, analysis/1, residue/1, diet/1, neglect/3, beauty/1, look/3
 C	spectator/1, space/5, lyric/1, panel/2, focus/6, defect/1, hands/1, injury/1, possibility/2, motive/1, strait/2, compensation/1, graft/1, custody/3, increase/3, drawing/2, step/1
 D	loss/2, substitute/1, level/1, attitude/1
 E	attention/1, competition/1, mystery/1, view/1
 F	safety/1, sense/3, exchange/3, assurance/2, notification/2, death/3, life/8, job/5, debt/1, security/7, flask/1

Figure 6.5 *Samples Clustered by Information Susceptibility*

according to their responses to different types of disambiguation information. However, given the fuzziness of these clusters, we can give no more than an intuitive account of the pattern. Also, unless we have more samples, it seems impossible to meaningfully "name" these clusters. Nevertheless, we can still analyse the patterns with respect to the different types of semantic relation associated with the clusters. One of the important points is that as we go up the specificity spectrum of semantic relations, the senses susceptible to those relations seem to involve more and more abstract thinking. Here are some examples from the various clusters:

- Block B: What is sun/1? One can readily point to that source of heat and light in the sky (assuming a sunny day). What is mirror/1? One understands it as the instrument which reflects the image of oneself when standing in front of it.

- Block C: What is space/5? One cannot possibly point to a concrete object, but would rather think of the region with planets, satellites, orbits, rockets, and so on. Similarly, what is injury/1? It can be a bleeding wound, a broken leg, a dislocated knee, among others.

- Blocks D and E: What is loss/2? Loss as such is difficult to describe. A loss can be anything but not everything is a loss. What is attention/1? It is even harder. If it is "the concentration of one's mind", then what is "concentration" and what is "mind"?

Hence we are moving towards examples where an abstract idea can only be expressed via further abstract ideas. As Wittgenstein (1958) claimed, one just needs to play the game differently with different kinds of meaning. Similarly, different kinds of sense are best characterised and disambiguated by different types of information. One could argue that abstract things might eventually be reducible to being expressed by more concrete things (e.g. a group of biological cells, etc.). However, this may involve a long chain of reduction, and in the gestalt sense, something as a unified whole may not be conceived as the same as the sum of its parts. The relation we found between the clusters and the semantic relation spectrum also coincides with human response in word association tests. For instance, in the Birkbeck word association norms (Moss and Older, 1996), "loss" triggers associations like "death" and "grief", none of which are related through a narrow IS-A relation. On the contrary, "sun" shows association with "moon", "diet" with "food", and "trick" with "magic". There are responses with syntagmatic associations, but at least these latter examples can easily be related by the WordNet hierarchy. Such association data therefore support the claim that different strategies are employed by people to conceptualise different groups of senses.

Nevertheless, as said in the previous chapters, the process of WSD is essentially finding a best match for the target word between the conventionalised context (CC), as in a lexical resource, and the triggering context (TC), as in the actual text. A sense can be concrete and can thus be understood via a narrow semantic relation with other words. However, when we are trying to disambiguate, the TC could be so variable that it may not always instantiate to the CC of the target word in the lexical resource. Moreover, as observed from the Denotative Approach (Experiment NAT-1) in Section 6.4, the typical TC for the same word can be very different according to text types. For example, one can expect an easy match of CC and TC for mirror/1 in a chapter about optics in a physics textbook, but not in a story like "The wicked queen asks the magic mirror ...". Here comes precisely the interaction among *Lexical Resource*, *Contextual Information* and *Nature of Target Words*. On the one hand, some words are inherently abstract and thus can only be explained abstractly in lexical resources, and disambiguation of them may subsequently require triggering contextual information to be evaluated against broader semantic relations. On the other hand, some words can be effectively explained in terms of their relations with other words in lexical resources, but if these words can practically occur in nearly any context, then whether such relations are sufficient for disambiguation would largely depend on the *actual* triggering context. So our investigation into the susceptibility of senses to different types of information does provide us an indication of how the utility of the information can differ in various situations. Hence we can try to capitalise on the various types of information accordingly during WSD. This will be further discussed in the next section.

While we see in Experiment NAT-2 that some senses are only identified with syntactic patterns, it may not at all be obvious that some other senses would only be identified with higher level information such as world knowledge and inference. Some examples in block F (unknown) may be of this kind. However, as there is an intrinsic distinction between lexical resources and encyclopaedia, we will not delve into the handling of those word senses here.

6.5.5 Prospect of the Connotative Approach

As suggested, the clusters of senses like those obtained in Experiment NAT-2 could be applied in reverse in WSD systems for tuning the relative importance of various types of disambiguation information. A scheme for doing this is proposed below.

First, to make the sense clusters useful in practice, we need to enrich them with more samples. Sampling target words in cloze test format would keep the annotation demand to a low level. However, when there are more samples, clustering may become tedious if done by hand, but could possibly be assisted with some statistical procedure like factor analysis to extract certain prominent or core clusters.[19]

Second, once we have obtained a group of larger sized clusters of reasonable generality, other senses could be incorporated into them in a more straightforward way, e.g. by using some similarity-based methods.

Third, the sense clusters could then form the sense types in the Sense Type Classifier in our Integrated Model (see Chapter 3), and such information can be used as a guideline to tune the weighting of different types of disambiguation information in the Information Filter. In other words, evidence which is likely to be strong in a given situation should receive more emphasis. Although how information is to be weighed would be another empirical issue, we expect that noise could be reduced and results improved by making the WSD process sensitive to target nature (and perhaps text types as well). This may also lead to a more realistic and cognitively sound computational model of WSD, as it has been shown that senses are best characterised by different types of information. In addition, our approach provides a more transparent

[19] We have tried feeding our data in Experiment NAT-2 to SPSS for factor analysis with the VARIMAX rotation. Six factors were extracted, and the first four correspond very closely to the results of our manual clustering.

picture of the contribution of individual disambiguation clues. Also, since the weight tuning is based on clusters of senses of similar nature (instead of individually based), it allows more room for handling senses even if they are not already in the clusters.

6.6 Summary

In this chapter, we have thoroughly studied an important but nevertheless overlooked intrinsic parameter of WSD, namely target nature. We discussed three aspects of target nature, namely polysemy, dispersion, and information susceptibility: the first two are word-based while the third is sense-based. Lacking any established paradigm for studying the parameter, we proceeded in two directions.

In the Denotative Approach, we defined the three aspects of target nature by some features in WN16, and further analysed the disambiguation results obtained in Chapter 5 with respect to these definitions. The difficulty of WSD, as expected, apparently correlated positively with polysemy but negatively with dispersion. More interestingly, we found that sense types as defined by WordNet top nodes were distributed in rather different patterns across text types. Although these sense types did not seem to relate to information susceptibility directly, analysing the disambiguation performance with respect to their average distribution patterns in different text types nevertheless allowed us to account for certain observations.

In the Connotative Approach, we started with no particular named sense types. Instead, using the cloze test format, we analysed the response of the target words (i.e. fillers of the cloze blanks), forming clusters of senses according to their susceptibility to various types of disambiguation information. From these clusters we inferred the possible interaction among the important parameters in WSD. Moreover, the contribution to WSD expected from our Integrated Lexical Resource (ILR) has been assessed, since we found about 63% of our test samples could be disambiguated just with the information available from the ILR. Suggestions were also made for expanding the sense clusters obtained by this approach and using such information to tune the relative importance of different disambiguation evidence.

In this chapter and the previous two, we have done a limited but systematic investigation of three major parameters that are likely to influence the difficulty of WSD. According to Sparck Jones (1986:p.20), sense labelling involves two related problems:

1. to discover the procedure we use for finding out what a text is about, for picking up its message; and
2. to discover the kind of information that the procedure depends on.

Past studies in WSD have addressed these questions independently, and only on external constraints. It is without doubt that external factors regarding lexical resource and contextual information are critical for any WSD system. However, we have also shown that target nature, which is intrinsic to the words, plays an important role in determining the effectiveness of WSD along with the extrinsic factors. We therefore believe that WSD performance in general could be upgraded if the disambiguation process was made sensitive to the interaction of all these factors. So in the next chapter, we shall conclude by revisiting our Integrated Model and summing up the contribution of this study to the understanding of WSD.

7

Conclusion

In the last three chapters, we have thoroughly investigated the three parameters in our WSD model. In this conclusion chapter, we will summarise our findings and contributions in Section 7.1, and close the loop by relating these results to our Integrated Model (introduced in Chapter 3) in Section 7.2. Finally, we will finish with suggestions for future directions in Section 7.3.

7.1 Results and Achievements

I Lexical Resource

- We have designed a structurally-based sense-mapping algorithm to combine various types of lexical resource. LDOCE was used as an example of conventional dictionaries and Roget's Thesaurus as one of conventional thesauri. Word senses in these resources were linked via WordNet which functioned as a mediator.
- A series of pilot tests for the effectiveness of the algorithm was performed. On average, about 70% of the direct mappings were judged to be appropriate. The accuracy was found to be dependent on the polysemy of the words. Most errors apparently arose from the inadequacy of individual resources, e.g. a sense occurring in one resource may not occur in another, and our experiments suggested the error scale. For the indirect mapping, the accuracy was slightly inferior as expected, but was still better than what one could get by chance. The results were therefore in favour of our algorithm.
- Using the second part of the algorithm, a full-scale linkage was done between WN16 and ROGET for the nouns shared by them. Over 30,000 senses from more than 18,000 nouns were mapped, and the accuracy was estimated to be similar to that in the pilot tests. The results of the combination formed the Integrated Lexical Resource (ILR) which provided most of the lexical information used in the rest of our work.

II Contextual Information

- Our work has contributed to the systematic study of the effect of information from the triggering context on WSD. Context was separated into size and realisation (i.e. the type of information and its computational handling). We first fixed the latter as a narrow semantic relation, namely one of IS-A, and experimented with various context sizes. We then fixed the context size at full-text and tested semantic relations of different specificity obtainable from the ILR.

- We have designed the Monosemous Filtering Algorithm (MFA) which expands context at lower cost but in a restricted way. This enabled us to test much wider context sizes, which would otherwise be impractical due to combinatorial explosion. We found that recall in general increased with context size, but at variable expense of precision depending on text types. Also, the effectiveness of the MFA for approximating large contexts was better on scientific writings than on fictional texts. This might be a result of some inherent differences among text types in the repetitive use of words, the tendency of words being used with topically neutral senses, the quality of the monosemous nouns for driving the MFA, and so on.

- Contrary to most past studies which only used as much information as available from a single lexical resource, we actually used information from various resources in combination from the ILR. The individual and combined effects of several types of semantic relation were tested. We found that multiple information often performed better than single information, although the increase was not monotonic. This shows that not all words are suitably disambiguated with the same type of lexical information. Nevertheless, the results also demonstrated the success of enriching lexical information for the benefit of WSD by combining resources.

III Nature of Target Words

- We have also investigated the role played by the nature of target words, which was never extensively researched in past studies. Three aspects of target nature were discussed, namely polysemy, dispersion, and the information susceptibility of senses. We approached the problem from two directions, attempting to arrive at some classification of word senses based on their response to various types of disambiguation information.

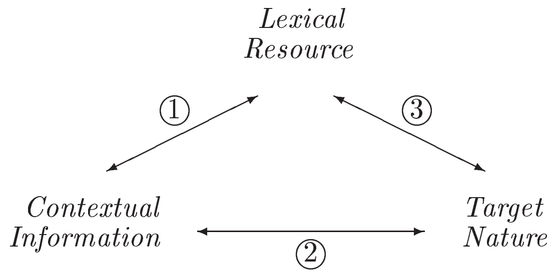
- In the Denotative Approach, we analysed the disambiguation outcomes obtained from a previous part of the study with respect to the three aspects of target nature defined by information in WordNet. Results supported the general hypotheses that WSD got more difficult as the number of senses increased and their dispersion decreased. Although the correctness and error of disambiguation did not seem to depend on the sense types as defined by WordNet top nodes, we observed interesting distribution patterns of such sense types across different text categories. The most notable feature was that fictional texts, when compared with press reportage and scientific writings, had almost twice the proportion of "entity" senses but about one-third fewer "abstraction" senses.

- In the Connotative Approach, we started with no particular named sense types, but derived plausible sense clusters from the responses of the test samples to different types of disambiguation information. The cloze test format was used in the experiment. Again, we tested disambiguation with semantic relations of different specificity available from the ILR. In addition, we informally used some other information obtained from the Brown Corpus to complete our information spectrum. For the 60 blanks in our test, we found that over 60% could be disambiguated solely with the semantic information from the ILR, while over 80% could be disambiguated with additional syntactic and co-occurrence information from the corpus. Moreover, we manually clustered the test samples according to their information susceptibility, and were able to identify certain underlying characteristics. We believe that results of this kind could be used to build the Sense Type Classifier (STC) in our model for tuning the weighting of different disambiguation clues in the Information Filter (IF).

7.2 Closing the Loop

In this section, we will close the loop by revisiting the Integrated Model introduced in Chapter 3 and consolidating the interaction among its components with our results.

According to the simplified diagram below relating the three parts in the model, our results illustrated the three forms of interaction shown:



① Lexical Resource ⟷ Contextual Information

The types of lexical information available from any particular lexical resource are limited, but WSD requires a wide range of information distributed across different existing resources. An obvious solution is to combine various resources (especially various types of resource) to enrich the sets of lexical information which can be used simultaneously and compatibly. Our sense-mapping algorithm in Chapter 4 has demonstrated the feasibility of such a combination, and our application of the ILR in the disambiguation experiments in Chapter 5 supports the claim that lexical information from various types of resource is needed for WSD.

② Contextual Information ⟷ Target Word

Different types of information from the triggering context are needed because they are useful for different target words. The effect of context size and the individual contribution of some types of information were studied in Chapter 5. We found that larger context sizes were in general beneficial to WSD, but to variable extents for different text types. Meanwhile, different types of information do not have uniform effects for all target words, and this is also closely related to text types. Our analysis in Chapter 6 showed that the distribution patterns of sense types (as defined by WordNet top nodes) varied considerably with text types. Also, our experiment with cloze blanks resulted in clusters of senses which suggested that more abstract senses required broader semantic relations for effective disambiguation. We have proposed a bidirectional relation between the STC and the IF in our model, such that the information susceptibility of target word senses should adjust the way different types of disambiguation evidence are combined, while the disambiguation outcomes could empirically refine the classification of sense types.

③ Target Word $\longleftrightarrow$ Lexical Resource

Following from the last point, word senses are most effectively disambiguated by different types of information, probably because senses are best characterised by different types of lexical knowledge. This also brings us back to the first point on the need to combine lexical resources. However, the lexical knowledge as represented in lexical resources is nonetheless limited, and our experiments in Chapter 6 (using only the various types of information available from the ILR) showed that the limited lexical semantic knowledge in our ILR could perform WSD satisfactorily slightly more than 60% of the time. Extra information might need to be extracted from actual uses of the words (e.g. from co-occurrence in corpora), which could upgrade the WSD accuracy to more than 80%, as our limited experiments suggested.

7.3 Future Directions

All in all, this work has contributed to the understanding of the interaction among the three major parameters in any WSD system from a holistic perspective. Nevertheless, the whole problem of WSD is much bigger than can be described here. In particular, there are still many unexplored areas in terms of the properties of the variables as well as the application of the results from various parts. Future development can thus cover several (related) aspects, which include:

- expanding the resource combination process to obtain even richer information from various places for WSD;
- testing the effect of a wider range of disambiguation information, such as incorporating lexical information which relates words across POSs (e.g. selectional restrictions);
- refining the word sense classification experiments by testing more samples, and also improving the qualities of the lexical information extracted from corpora, e.g. using larger corpora, so that more test words could be covered and the estimations could be more reliable;
- demonstrating the effect of making evidence combination sensitive to the nature of target words by doing experiments on weighting adjustment with the sense clusters taken into account; and
- extending the study to the disambiguation of words of other POSs.

A

The SEMCOR Files

A.1 Text Types and Letter Codes

The SEMCOR files are divided into various categories, each indicated by a letter code, as follows:

A	Press: Reportage	J	Learned and Scientific Writing
B	Press: Editorial	K	Fiction: General
C	Press: Reviews	L	Fiction: Mystery and Detective
D	Religion	M	Fiction: Science
E	Skills and Hobbies	N	Fiction: Adventure and Western
F	Popular Lore	P	Fiction: Romance and Love Story
G	Belles Lettres, Biography, etc.	R	Humour
H	Miscellaneous		

A.2 Statistics on Noun Distribution

Table A.1 shows some statistics on the noun distribution in various SEMCOR text categories. The percentage in bracket in the 4th column indicates the average proportion of noun tokens out of the average text length (Ave Len); and that in the 6th column indicates the average proportion of monosemous noun tokens out of the average number of all noun tokens.

Cat	# Files	Ave Len	Ave # Nouns		Ave # Mono Nouns	
			Token	Type	Token	Type
A	7	1850	415 (22.47%)	244	109 (25.99%)	75
B	2	1871	423 (22.58%)	324	113 (26.92%)	91
C	3	1912	406 (21.20%)	289	97 (23.89%)	76
D	4	1951	414 (21.21%)	209	75 (18.39%)	45
E	14	1909	501 (26.33%)	260	112 (21.88%)	64
F	19	1922	458 (23.82%)	264	116 (25.53%)	70
G	18	1939	425 (21.90%)	261	87 (20.52%)	62
H	12	1932	500 (25.91%)	212	112 (22.27%)	55
J	43	1941	484 (24.92%)	203	115 (23.60%)	53
K	29	1943	361 (18.60%)	241	83 (22.94%)	63
L	11	1947	308 (15.86%)	206	69 (22.36%)	51
M	2	1914	314 (16.40%)	224	80 (25.27%)	61
N	10	1979	346 (17.47%)	217	71 (20.34%)	52
P	6	1954	327 (16.73%)	214	76 (22.88%)	56
R	6	1934	366 (18.94%)	262	94 (25.32%)	77

Table A.1 *Noun Distributions in Various SEMCOR Categories*

A.3 Text Examples

In this section, we show an excerpt of one example file for each SEMCOR category, to illustrate the different text types.

A (Press: Reportage) – br-a01

The Fulton County Grand Jury said Friday an investigation of Atlanta's recent primary election produced "no evidence" that any irregularities took place.

The jury further said in term end presentments that the City Executive Committee, which had over-all charge of the election, "deserves the praise and thanks of the City of Atlanta" for the manner in which the election was conducted.

B (Press: Editorial) – br-b13

Sizzling temperatures and hot summer pavements are anything but kind to the feet. That is why it is important to invest in comfortable, airy types of shoes.

There are many soft and light shoe leathers available. Many styles have perforations and an almost weightlessness achieved via unlined leathers. Softness is found in crushed textures.

C (Press: Reviews) – br-c01

It is not news that Nathan Milstein is a wizard of the violin. Certainly not in Orchestra hall where he has played countless recitals, and where Thursday night he celebrated 20th season with the Chicago Symphony orchestra, playing the Brahms Concerto with own slashing, demon-ridden cadenza melting into the high, pale, pure and lovely song with which a violinist unlocks the heart of the music, or forever finds it closed.

D (Religion) – br-d01

One of the most widespread features of contemporary thought is the almost universal disbelief in the reality of spirit. Just a few centuries ago the world of spirits was as populous and real as the world of material entities. Not only in popular thought but in that of the highly educated as well was this true. Demons, fairies, angels, and a host of other spiritual beings were as much a part of the experiential world of western man as were rocks and trees and stars. In such a world the words "matter" and "spirit" both referred to directly known realities in the common experience of all. In it important elements of Christianity and of the Biblical view of reality in general, which now cause us much difficulty, could be responded to quite naturally and spontaneously.

E (Skills and Hobbies) – br-e02

The most beautiful bed of pansies I've seen was in a South Dakota yard on a sizzling day. Pansies are supposed to like it cool, but those great velvety flowers were healthy and perky in the glaring sun. I sought out the gardener and asked him what he did to produce such beauties in that weather. He seemed puzzled by question. "I just love them", he said.

F (Popular Lore) – br-f03

Most psychiatrists dismiss these instances of that weird feeling as the deja vu (already seen) illusion, just as they dismiss dream previsions as coincidences. In this manner they side-step the seemingly hopeless investigation of the greater depths of mystery in which all of us grope continually.

When a man recognizes a certain experience as the exact pattern of a previous dream, we have an instance of deja vu, except for the fact that he knows just why the experience seems familiar. Occasionally there are examples of prevision which cannot be pushed aside without confessing an unscientific attitude.

G (Belles Lettres, Biography, etc.) – br-g01

Northern liberals are the chief supporters of civil rights and of integration. They have also led the nation in the direction of a welfare state. And both in objectives of non-discrimination and of social progress they have had ranged against them the Southerners who are called Bourbons. The name presumably derives from the French royal house which never learned and never forgot; since Bourbon whiskey, though of Kentucky origin, is at least as much favored by liberals in the North as by conservatives in the South.

H (Miscellaneous) – br-h01

The Office of Business Economics (OBE) of the U.S. Department of Commerce provides basic measures of the national economy and current analysis of short-run changes in the economic situation and business outlook. It develops and analyzes the national income, balance of international payments, and many other business indicators. Such measures are essential to job of presenting business and Government with the facts required to meet the objective of expanding business and improving the operation of the economy.

J (Learned and Scientific Writing) – br-j01

It has recently become practical to use the radio emission of the moon and planets as a new source of information about these bodies and atmospheres. The results of present observations of the thermal radio emission of the moon are consistent with the very low thermal conductivity of the surface layer which was derived from the variation in the infrared emission during eclipses (e.g., Garstung, 1958). When sufficiently accurate and complete measurements are available, it will be possible to set limits on the thermal and electrical characteristics of the surface and subsurface materials of the moon.

K (Fiction: General) – br-k01

Scotty did not go back to school. parents talked seriously and lengthily to own doctor and to a specialist at the University Hospital - Mr. McKinley was entitled to a discount for members of family - and it was decided it would be best for him to take the remainder of the term off, spend a lot of time in bed and, for the rest, do pretty much as he chose - provided, of course, he chose to do nothing too exciting or too debilitating. teacher and school principal were conferred with and everyone agreed that, if he kept up with a certain amount of work at home, there was little danger of losing a term.

L (Fiction: Mystery and Detective) – br-l11

Perhaps if Felix had first come upon us when this boy was not cavorting so gaily up and down the hall outside the murdered woman's apartment, we might have had less trouble convincing Felix of seriousness. This, you will remember, was still New Year's Day. By the time Felix turned up it was early afternoon, which, one would think, would be late enough so that by then, except for small children and a few hardy souls who had not yet sobered up, it could have been expected that people would no longer be having any sort of active interest in the previous night's noisemakers and paper hats.

M (Fiction: Science) – br-m01

Now that he knew himself to be self he was free to grok ever closer to brothers, merge without let. Self's integrity was and is and ever had been. Mike stopped to cherish all brother selves, the many threes-fulfilled on Mars, corporate and discorporate, the precious few on Earth - the unknown powers of three on Earth that would be his to merge with and cherish now that at last long waiting he grokked and cherished himself.

N (Fiction: Adventure and Western) – br-n05

She was carrying a quirt, and she started to raise it, then let it fall again and dangle from wrist .

"I saw fire", she said, speaking slowly, making an effort to control anger. "You could burn down this whole mountainside with a fire that size. It wouldn't matter to a fool like you. It would to me".

"All right", Wilson said quickly. "The fire's too big. And I appreciate the advice".

P (Fiction: Romance and Love Story) – br-p01

The only thing unusual about the Old Man had long since happened. But the past was dead here as the present was dead. Once the Old Man had had a wife. And once she, too, ignored him. With a tiny fur-piece wrapped around shoulders, she wiggled satin covered buttocks down the street before him and didn't stop. In one hand she clutched a hundred dollar bill and in the other a straw suitcase. The way she strutted down the street, the Old Man would have been blind not to have noticed both. Without looking at him, without looking at anything except Drexel Street directly in front of, she climbed up into one of those orange streetcars, rode away in it, and never came back.

R (Humour) – br-r05

Nothing in English has been ridiculed as much as the ambiguous use of words, unless it be the ambiguous use of sentences. Ben Franklin said, "Clearly spoken, Mr. Fogg. You explain English by Greek". Richard Brinsley Sheridan said, "I think the interpreter is the hardest to be understood of the two". And a witty American journalist remarked over a century ago what is even more true today, "Many a writer seems to think he is never profound except when he can't understand own meaning".

B

Inspecting the Integrated Lexical Resource

The Integrated Lexical Resource (ILR) formed by mapping the noun senses between WN16 and ROGET in Chapter 4 was inspected with a sample of noun senses. The samples and the results of the inspection are shown in Sections B.1 and B.2 respectively.

B.1 Sampling Noun Mappings in the ILR

Noun samples (underlined) were selected from the following excerpt, and the mappings for the senses of these nouns in the ILR were evaluated.

> "The <u>concert</u> opened with another big romantic <u>score</u>, Schumann's Overture to Manfred, which suffered <u>fate</u>, this <u>time</u> with orchestral <u>thrusts</u> to the Byronic <u>point</u> to keep it afloat. Hindemith's <u>joust</u> with Weber <u>tunes</u> was a considerably more serious <u>misfortune</u>, for it demands translucent <u>textures</u>, buoyant <u>rhythms</u>, and astringent <u>wit</u>. It got the <u>kind</u> of scrambled, coarsened <u>performance</u> that can happen to best of <u>orchestras</u> when the <u>man</u> with the <u>baton</u> lacks <u>technique</u> and <u>style</u>." (*Brown Corpus c01, 4th paragraph*)

Table B.1 shows the polysemy of the samples according to WN16 (Column W) and ROGET (Column R). Samples were classified into the LO, MED or HI group as in Experiment P2 in Chapter 4.

B.2 Judgement of the Mappings

Table B.2 shows our judgement on the appropriateness of mapping with respect to each test sense in WN16. In the **Source** column are the words and senses in WN16. In the **Target(s)** column are the mapped ROGET heads. The judgement was done according to the rules described in Chapter 4. The possible outcomes (column **Acc**) are Accurately Mapped (A), Incorrectly Mapped (I), Unmapped-a (Ua), Forced Error (F), and Unmapped-b (Ub).

LO			**MED**			**HI**		
Noun	**W**	**R**	**Noun**	**W**	**R**	**Noun**	**W**	**R**
concert	1	6	style	8	18	man	11	8
joust	1	1	time	10	9	score	11	14
kind	1	2				point	24	21
tune	1	2						
baton	2	2						
misfortune	2	4						
orchestra	2	4						
technique	2	4						
fate	3	7						
wit	3	8						
texture	4	4						
performance	5	11						
rhythm	5	10						
thrust	5	7						

Table B.1 *Polysemy of Nouns from the Excerpt*

Source		Target(s)	Words in common	Acc
concert				
1	*performance of music*	410 (Melody)	concert, turn	*A*
joust				
1	*combat between knights*	716 (Contention)	battle, duel, tilt, etc.	*A*
kind				
1	*category of things*	77 (Class)	category, sort, etc.	*A*
tune				
1	*succession of notes*	412 (Music)	harmony, melody, etc.	*A*
baton				
1	*conductor's stick*	218 (Support)	pole, rod, etc.	*A*
2	*instrument in relay*	218 (Support)	baton, pole	*A*
misfortune				
1	*unforeseen trouble*	616 (Evil), 731 (Adversity), 827 (Painfulness)	affliction, outrage, embarrassment, etc.	*A*
2	*unfortunate state*	731 (Adversity)	bad luck, failure, etc.	*A*
orchestra				
1	*instrumentalists*	412 (Music)	band, duet, trio, etc.	*A*
2	*main floor in theatre*	594 (Drama)	circle, stalls, etc.	*A*
technique				
1	*practical method*	624 (Way), 629 (Means)	method, technique	*A*
2	*skillfulness*	694 (Skill)	expertise, mastery, etc.	*A*
fate				
1	*future event*	69 (End)	destiny, ending, etc.	*A*
2	*ultimate agency*	156 (Cause)	agent, force, etc.	*A*
3	*condition in life*	159 (Chance)	fortune, luck, etc.	*A*
wit				
1	*humourous message*	839 (Wit)	humor, witticism, etc.	*A*
2	*mental ability*	498 (Intelligence)	cleverness, genius, etc.	*A*
3	*amusing person*	839 (Wit)	humorist, satirist, etc.	*A*
texture				
1	*feel of surface*	331 (Structure)	feel, texture	*A*
2	*essential quality*	–	–	*Ub*
3	*musical pattern*	–	–	*Ub*
4	*tactile surface appearance*	–	–	*Ub*
performance				
1	*dramatic or musical entertainment*	594 (Drama)	play, show, etc.	*A*
2	*act of presenting entertainment*	551 (Representation), 594 (Drama), 876 (Celebration)	presentation, show	*A*
3	*act of doing something*	676 (Action)	doing, execution, etc.	*A*
4	*process of functioning*	676 (Action)	labour, operation, etc.	*A*
5	*recognized accomplishment*	676 (Action)	completion, deed, etc.	*A*
rhythm				
1	*rhythmic unit in music*	410 (Melody)	beat, rhythm, tempo	*A*
2	*regular recurrence*	141 (Periodicity)	periodicity, regularity	*A*
3	*interval, cycle*	141 (Periodicity), 410 (Melody)	cycle, round, pause, suspension, etc.	*A*
4	*speech rhythm*	593 (Poetry)	accent, prosody, etc.	*A*
5	*birth control method*	–	–	*Ub*
thrust				
1	*force used in pushing*	279 (Impulse)	force, impetus, etc.	*A*
2	*thrusting blow with knife*	279 (Impulse)	blow, punch, etc.	*A*
3	*apply force to propel something*	287 (Propulsion)	drive, projection, etc.	*A*
4	*verbal criticism*	712 (Attack)	attack, fire, etc.	*F*
5	*sharp hand gesture*	279 (Impulse)	jab, thrust	*F*

Table B.2 *Appropriateness Judgement on Sense Mapping in ILR*

	Source	Target(s)	Words in common	Acc
style				
1	*particular kind of appearance*	77 (Class)	brand, sort, type, etc.	*A*
2	*manner of performance*	7 (State)	disposition, manner, etc.	*A*
3	*way of expressing something*	624 (Way), 848 (Fashion)	signal, show, style	*I*
4	*popular taste*	848 (Fashion)	trend, vogue, etc.	*A*
5	*stylish elegance*	841 (Beauty), 848 (Fashion)	chic, elegance, etc.	*A*
6	*elongated part of pistil*	366 (Vegetable life)	pistil, stigma, etc.	*A*
7	*pointed tool for engraving*	555 (Engraving)	punch, style	*A*
8	*tubular process*	624 (Way), 688 (Conduct)	process, style	*F*
time				
1	*an instance for some event*	–	–	*Ub*
2	*an indefinite period*	110 (Period)	day, month, period, etc.	*A*
3	*period of time under control*	110 (Period)	day, month, period, etc.	*A*
4	*a suitable moment*	116 (Instantaneity)	instant, moment, etc.	*A*
5	*continuum of experience*	410 (Melody)	measure, space, etc.	*F*
6	*clock time*	–	–	*Ub*
7	*the fourth coordinate*	410 (Melody)	measure, time	*F*
8	*experience on particular occasion*	116 (Instantaneity)	flash, time	*F*
9	*rhythm*	410 (Melody)	cadence, time	*A*
10	*period of imprisonment*	108 (Time), 110 (Period)	tenure, term, etc.	*I*
man				
1	*adult male person*	372 (Male)	boy, chap, male, etc.	*A*
2	*someone in armed forces*	686 (Agent)	artisan, craftsman, etc.	*A*
3	*any human being*	742 (Servant)	attendant, worker, etc.	*I*
4	*inhabitants of the earth*	371 (Humankind)	mankind, people, etc.	*A*
5	*member of the family Hominidae*	365 (Animality), 371 (Humankind)	primate, human being, etc.	*A*
6	*a male subordinate*	742 (Servant)	follower, subordinate, etc.	*A*
7	*adult male with manly character*	372 (Male)	fellow, male, etc.	*A*
8	*husband, lover, boyfriend*	372 (Male)	fellow, male, etc.	*I*
9	*personal attendant*	742 (Servant)	butler, valet, etc.	*A*
10	*Isle of Man*	365 (Animality)	man, newfoundland	*F*
11	*small object in board games*	365 (Animality), 372 (Male)	game, jack, etc.	*F*
score				
1	*number or letter indicating quality*	27 (Degree)	grade, mark, etc.	*A*
2	*musical composition*	412 (Music)	music, opus, etc.	*A*
3	*number expressing accomplishment*	85 (Number)	constant, decimal, etc.	*A*
4	*set of 20 members*	99 (Five and over)	octet, quintet, etc.	*A*
5	*grounds*	86 (Numeration), 87 (List), 802 (Credit), 803 (Debt), 808 (Accounts)	account, score	*F*
6	*facts*	27 (Degree), 547 (Indication)	point, score	*F*
7	*an amount due*	802 (Credit), 803 (Debt)	loan, bad debt, etc.	*A*
8	*notch to keep a tally*	547 (Indication)	imprint, notch, etc.	*A*
9	*strong resentment*	–	–	*Ub*
10	*act of scoring in game*	547 (Indication)	mark, pass, score	*F*
11	*seduction*	–	–	*Ub*

Table B.2 *(Cont.) Appropriateness Judgement on Sense Mapping in ILR*

	Source	Target(s)	Words in common	Acc
point				
1	*geometric element*	33 (Smallness), 452 (Topic)	part, section, etc.	*I*
2	*precise location*	547 (Indication)	base, line, etc.	*I*
3	*brief version of essential meaning*	452 (Topic)	gist, message, etc.	*A*
4	*an isolated fact*	475 (Reasoning)	case, reason, etc.	*A*
5	*specific identifiable position in series*	8 (Circumstance), 73 (Term)	position, stage, etc.	*A*
6	*short period of time*	69 (End)	peak, term, etc.	*I*
7	*object of an activity*	617 (Intention)	object, target, etc.	*A*
8	*a V shape*	69 (End), 213 (Summit)	peak, point, tip	*A*
9	*small circular shape*	33 (Smallness), 185 (Place), 547 (Indication), 586 (Writing)	dot, point	*A*
10	*unit of counting*	33 (Smallness)	bit, point	*F*
11	*promontory in water*	254 (Prominence)	headland, promontory, etc.	*A*
12	*a distinct part*	33 (Smallness)	detail, point	*F*
13	*style in speech*	–	–	*Ua*
14	*outstanding characteristic*	547 (Indication)	characteristic, trademark, etc.	*A*
15	*sharp end*	69 (End)	endpoint, tip, etc.	*A*
16	*compass point*	186 (Place)	direction, orientation, etc.	*A*
17	*linear unit*	69 (End), 213 (Summit), 587 (Print)	em, pica, etc.	*A*
18	*full stop*	547 (Indication)	comma, semicolon, etc.	*A*
19	*V-shaped mark of arrow pointer*	547 (Indication)	arrow, pointer, etc.	*A*
20	*property of a shape*	–	–	*Ua*
21	*distinguishing characteristic*	547 (Indication)	characteristic, trademark, etc.	*A*
22	*the muzzle's direction*	254 (Prominence), 475 (Reasoning)	muzzle, opening, etc.	*F*
23	*a wall socket*	185 (Place)	point, receptacle	*A*
24	*contact in the distributor*	27 (Degree), 73 (Term)	point, tread	*F*

Table B.2 *(Cont.) Appropriateness Judgement on Sense Mapping in ILR*

C

Detailed Results for Experiment REA-2

C.1 Combined Effect of Context Realisations

Table C.1 shows the combined effect of context realisations for individual text categories in Experiment REA-2 (refer to Chapter 5).

Cat		*wn+def*		*wn+rog*		*wn+rog+com*		*wn+rog+def+com*	
		−freq	*+freq*	*−freq*	*+freq*	*−freq*	*+freq*	*−freq*	*+freq*
A	Precision	53.11	72.71	55.95	72.40	50.52	70.10	48.28	65.24
	Recall	43.42	70.64	41.76	70.40	45.14	67.82	45.34	62.64
B	Precision	41.84	61.70	38.83	67.48	38.61	60.30	37.20	55.62
	Recall	35.31	59.50	30.38	65.43	33.12	58.32	33.52	53.49
C	Precision	38.87	62.34	43.10	65.19	40.49	62.85	38.57	55.21
	Recall	32.53	60.89	32.73	62.98	34.89	61.29	35.12	53.68
D	Precision	41.45	61.12	45.09	65.90	41.92	59.21	41.66	56.68
	Recall	34.35	59.97	33.00	64.45	36.92	57.25	37.43	55.27
E	Precision	47.79	68.61	46.97	70.58	46.40	65.96	45.71	62.70
	Recall	42.20	66.40	38.54	68.72	42.19	64.15	43.24	61.17
F	Precision	47.48	66.89	48.69	70.87	47.31	65.98	45.36	61.53
	Recall	41.12	65.20	38.99	68.73	42.48	63.53	42.25	59.71
G	Precision	41.95	60.63	42.31	65.77	39.92	59.78	39.66	54.96
	Recall	36.09	58.96	34.36	63.66	35.74	57.23	36.64	53.20
H	Precision	50.13	70.17	47.38	73.18	45.64	66.85	47.73	63.15
	Recall	43.15	68.18	38.59	71.28	40.95	64.22	44.35	61.36
J	Precision	45.56	70.09	45.73	72.23	44.84	66.31	43.87	63.24
	Recall	38.20	68.79	37.33	70.24	40.71	64.17	40.81	61.35
K	Precision	37.58	69.22	39.46	71.44	38.36	65.83	39.31	61.40
	Recall	28.69	67.42	26.86	69.61	31.80	63.68	34.66	59.51
L	Precision	34.66	67.38	38.53	70.14	36.01	63.71	35.67	60.53
	Recall	24.70	66.08	25.43	68.55	30.54	61.67	31.39	58.44
M	Precision	40.83	68.04	41.61	71.43	47.42	67.66	44.66	65.43
	Recall	31.15	66.01	29.53	67.42	39.55	64.42	38.73	62.80
N	Precision	37.95	69.11	42.74	73.60	39.57	64.44	37.92	60.90
	Recall	26.98	68.15	29.63	72.37	34.75	62.54	34.46	59.22
P	Precision	43.68	70.55	47.51	71.46	39.70	67.37	40.26	63.15
	Recall	32.43	69.57	29.20	70.56	33.95	65.82	35.59	61.49
R	Precision	37.24	66.87	39.46	69.55	37.00	63.19	37.40	57.99
	Recall	30.25	64.31	29.92	67.02	32.92	60.28	34.23	55.89

Table C.1 *Results for Individual Text Categories in Experiment REA-2*

D

Experiment NAT-2 and Cloze Texts

The cloze texts which were used as test materials[1] in Experiment NAT-2 (Chapter 6) are shown in Section D.1. Words in brackets are the preferred words for the corresponding blanks. The raw results on disambiguating the words for the blanks with various types of information are tabulated in Section D.2.

D.1 Cloze Texts and Test Blanks

Text-001: Vanished in Thin Air

Bob Clayton, a noted nightclub magician and entertainer waved his wand in the middle of a magic (trick), and was never seen again.

Bob has not been heard of since his (performance) at one of the clubs in Reno, Nevada, when he vanished from the stage. An all out police investigation proved futile as they could find no clues of the magician's whereabouts or why he disappeared so mysteriously. One of the (spectators) gave this explanation: 'He was performing before an audience of about 150 people, and everybody was watching intently. Suddenly, the magician cleared his magic wand, waved it and a cloud of bluish smoke appeared. Then in an (instant) he was gone. When the smoke raised, no one was on stage, no one at all.'

The police, however, were not as credulous as the spectators. They investigated the rumour that Clayton had staged his ingenious disappearing act just to escape from his creditors, since the magician was reported to be heavily in (debt). Despite the scepticism of the police, those who saw him go up in smoke maintain there is no (way) that the magician could have simply been doing a trick.

Text-002: Operation Weather Satellite

An unmanned Delta rocket carrying a weather observation satellite worth US$57 million blasted off today in the first US space launch of the year.

The smooth Delta launching came after two frustrating (delays). The first postponement was caused by a fuel leak and the second by intense crosswinds that, officials said, could have torn the rocket apart during its fiery ascent.

[1]Sources: http://www.gallaudet.edu/~etcweb/cloze{1-4}.html, Macabalitaw (1987), Nancarrow and Bradnack (1989a; 1989b)

The rocket placed in orbit an 840 kilogram Geostationary Operational Environmental Weather Satellite, which was identical to one that was destroyed, a few months ago, when a Delta rocket exploded after liftoff.

Since another weather satellite failed in orbit in August 1984, weather forecasters have had only a single satellite to monitor the entire United States.

The Delta was the first of six rockets scheduled to be launched this year by the National Aeronautics and Space Administration. This latest launch marked the third straight success this year for the US (space) program, which was plagued by a series of major failures last year. The most disastrous one was the loss of the space shuttle Challenger, which exploded only 73 seconds after liftoff on January 28, 1986, killing all seven crew members aboard.

Text-003: English through Song

RTV Hong Kong and the Chinese University's Extra Mural Department are encouraging young people to learn English by listening to popular Western music and imitating the pronunciation and the grammar.

Although reading a book or writing an essay can be encouraged, learning English by singing pop songs has been commended by some educationalists as a faster and more enjoyable route for Hong Kong's youngsters. It takes time to practise writing and to finish a book but a song can be played over and over again fairly quickly.

The (rhythm) of the music also helps people to remember the (lyrics) and thus gain new items of vocabulary, and, anyway, young people are naturally inclined towards music.

The English Through Song singing competition will be held as part of the RTV and Chinese University's on going programme to raise the standard of spoken English in Hong Kong. Twelve contestants have been nominated by a (panel) of judges drawn from members of the music and film industries. Primarily students, the amateur singers will battle it out with such classics as Tennessee Waltz and By the Time I get to Phoenix. In order to sing well, facial must match vocal expression.

Text-004: No More Glasses

Within a few years spectacles could become as old fashioned as ear trumpets or iron lungs. Even contact lenses could find themselves confined to museums of medical curiosities in the near future. The instrument that promises to make spectacles redundant is the laser. Scientist have now developed a new type of ultra violet laser which can correct the (focus) of the eyes in a small outpatient operation.

A computer can record the curvature of the eyes, then calculate the (level) of adjustment needed. It also directs the laser (beam). There is every possibility that this technique will be able to replace spectacles for thousands of people. The procedure uses a completely new type of laser called an excimea which has already proved itself in other areas such as putting out clots from arteries. It is basically intended for those who are born with shortsightedness, but there is no reason why it can't eventually be used for those who develop eye (defects) through age. In fact it's not too far fetched to predict that in time people could have their failing sight corrected every couple of years or so by laser beam.

Laser beams normalize eyesight by altering the lens of the eye itself and so correcting the focus. The operation is carried out under local anaesthetic and only takes ten to 12 minutes. But it involves very high degrees of precision and may be performed by an opthalmic expert only who is also a laser expert. In the wrong (hands) it could create a hole in the eye.

Text-005: Telescopes

People have always wondered about the universe. A remarkable instrument, the telescope, has allowed them to know more about it. The telescope seems to bring faraway objects right up to the viewer's eyes. For this reason, people sometimes say it has the power to shrink the universe. With a small telescope you can see features of certain planets, as well as many other details of our solar system. You can even see spots on the (sun) . These dark areas on our home star result from churning gases.

The word telescope is a blend of two Greek words meaning "far" and "see". Indeed, its job is to view things that are distant. There are several different types of telescope. The refracting telescope uses lenses like the ones found in eyeglasses to refract, or bend, light. The scientist Galileo invented the first practical refracting telescope in the early 1600's. Using it, he observed the rings of the planet Saturn and the Great Red Spot on the planet Jupiter. The scientist saw other planets as well.

Another kind of telescope is the reflector. It uses two curved mirrors, one to send light back to a point on the second. This, in turn, sends a (reflection) to the person using the device. For a long time scientists preferred the older refracting telescope because it provided a clearer, more reliable image, but then the making of glass improved. This made it possible to have better (mirrors) . Suddenly, the reflector was the telescope of choice among scientists.

In the last several decades, a new breed of telescope has emerged on the scene. Called the radio telescope, it is capable of searching much farther into the darkness of space than any other telescope. Radio telescopes do not explore through pictures but rather, through sound. They hear faraway stars and other distant objects. These objects give off signals that radio telescopes pick up.

Text-006: Newspapers in the U.S.

The first newspapers in this country appeared before the American Revolution. As the population of the United States grew throughout the 1800's, so did the newspaper business. In fact, it increased steadily.

A major change in the newspaper business took place in 1833, when Benjamin Day started a new paper called the New York Sun. Day charged only a penny for his paper at a time when most carried a price of around six cents. As a result, Day sold many more papers every day. Other penny papers soon sprang up, and by 1909 there were about 2,600 daily papers in the country. Some newspaper editors and owners were very powerful in those days. William Randolph Hearst, the owner of the Journal, had so much influence that he got the United States into a war!

During the twentieth century, the number of newspapers being published decreased. This was partly because the costs of putting out a newspaper were increasing and partly because of the popularity of radio and television. In the last few decades, many papers have disappeared. They were losing millions of dollars a year, so they had to shut down.

Modern newspapers use electronic computers for a great many tasks, including writing, editing, and page layout. These new (machines) have replaced the slower ones of days past. The typewriter and hand printing press are just two of these pieces of equipment that people no longer use. Yet, in some ways, the daily paper has not changed. Papers today still rely on the instincts and writing skills of human reporters to get the news out to readers. The (job) of the editor is the same, too. This person still has the responsibility to oversee the work of other employees and to write editorials which express views on major issues of the day. Newspapers today also include popular information and columns, such as movie schedules, reviews, and advice to readers. These attract (attention). Other features, such as cartoons and word games, also get people to buy more newspapers.

Text-007: The Red Cross

Usually life is routine, but sometimes disaster strikes. Floods, earthquakes, hurricanes, and plane crashes are just a few of the tragic possibilities that can occur. At such times the (loss) can be great. Whether natural, or manmade, disasters can claim high tolls in human life and property.

The American Red Cross, created in 1881, is an organization that offers relief when disaster strikes. The first goal of the organization is to provide (aid). This help may include giving food, clothing, and shelter, or it may be providing services, such as medical treatment or emotional support for victims. The founder of the Red Cross was the famous Civil War nurse, Clara Barton. She barely stood five feet tall, yet for someone so small, she had an enormous amount of energy and courage. During the war she often put in 18 hour days, sometimes right on the front lines, treating the injured, cooking hot cereal for the troops, and giving a kind word to battle weary soldiers. She worked hard.

Clara Barton decided to name her relief organization "Red Cross". She did not, however, originate the idea behind it. The concept began in Europe, where similar organizations had already long been at work helping soldiers wounded in battle. As in Barton's time, most of the people today who provide services on behalf of the Red Cross are (volunteers). In fact, the only people in the organization who do not work without pay are those involved in running its affairs. Part of the job of these salaried employees is to heighten the public's awareness of the Red Cross and to ask for contributions of money and time. Without these contributions, the Red Cross could not continue to do its good work.

In addition to its disaster relief services, the Red Cross also offers a wide range of public health and safety programs. A second goal of the organization nowadays is to educate. Local chapters may teach courses in first aid, swimming, or lifesaving techniques. Classes take place either at the Red Cross branch itself or at other locations in the community.

Text-008: The Ski Patrol

Do you know how to ski? You may think this leisure activity is new. In truth, scientists believe skiing to be quite ancient, dating back more than four thousand years. As evidence they offer primitive forms of skis that have been discovered. Skiing is one of the oldest and most important means of transportation in areas where there is a lot of snow. In fact, skiing became a sport only during the last half of the nineteenth century. As the popularity of skiing as a sport grew in the United States, people became more and more concerned about the (safety). The risks involved in traveling downhill at high speeds prompted this concern. To keep down the number of accidents and provide help for injured skiers, a National Ski Patrol System was created in 1938. Today, there are over 500 groups of ski safety experts nationwide, with a staff of more than 10000 volunteers. These volunteers patrol the slopes to offer assistance. They help both beginner and experienced skiers.

Suppose a skier wants to become a Ski Patrol volunteer. In order to join the Patrol, the person must learn first aid skills and exhibit his or her abilities on the slopes. Volunteers will be ranked by the skill they show. Ski Patrol members reviewing their performance will place each volunteer in a grade, or category.

The Ski Patrol concentrates on two main areas – first aid and rescue operations. Part of the training of volunteers is learning to recognize different (injuries). They must not only be able to know wounds when they see them but to treat these as well. In addition, volunteers receive instruction in methods for organizing search parties in the event a skier becomes lost. As part of its safety program, the Ski Patrol provides recreational skiers with information on the use of equipment and accident prevention. It emphasizes the importance of taking (precautions). These steps toward increased safety include skiing only in well marked areas and staying on slopes that are not beyond the ability of the skier.

Text-009: A Look at Arson

Government statistics indicate that the number of suspected arson cases has increased during the past three years. They amount to about three percent of all the fires in Hong Kong.

Fire Department (figures) show that last year alone, 19 people were killed in fires determined to be arson, representing 42 percent of the total of 45 deaths from fires. According to a Senior Division Officer of the Fire Services Department, fires that are started deliberately develop faster because the chemicals used accelerate the destructive process. And since there is a (possibility) of arson every time a fire is reported, firemen assume the responsibility of determining the cause of the fire and providing information to the police at the scene. In the (event) of suspicious circumstances, the police will begin an investigation.

The Department Official also explains that it is often not difficult to determine whether a fire started naturally or if it was premeditated. A decision can be made from (evidence) collected as well as by a chemical (analysis) of the debris. Firemen have to dig into the debris to determine the (place) where the fire began, and they even use their (sense) of smell in order to detect the use of chemicals. If, for instance, gasoline or any similar substance has been used to start a fire, some (residue) might be found in a partly burnt carpet.

When asked about the (motives) of arsonists, the Department Official says that they possibly include profit, a desire to hide a crime, revenge, or simply abnormal mental desires.

Text-010: Hazards of Pawning

Pawning valuables is commonly practised in Hong Kong. It can offer relief in times of financial (straits). However, with the recent burglaries involving jewelry shops and pawnshops in Hong Kong, people should think over the matter before deciding to pawn their property.

The practice is for pawnshops to offer loans well below the value of the pawned property. There is nothing objectionable about this so long as the property is in safekeeping and can be redeemed by the owner anytime he is able to pay back the loan with interest. Nevertheless, people who pawn their valuables in (exchange) for quick cash seem to be running a big risk.

In most cases, clients are not given any (assurance) of getting (compensation) in case their property is stolen. For instance, during the past Chinese New Year holiday, a pawnshop was broken into and burglars took off with pawned property worth over a million dollars. As a consequence, the pawnshop was forced to close down, but without giving due (notification) to its clients as to how they could contact the pawnbroker. The pawnshop owner just disappeared after closing his shop, and the borrowers' hopes of ever recovering their property vanished.

According to one police inspector, pawnshops in Hong Kong have very poor (security) systems. Some of the shops are over twenty years old, and the owners have not bothered to install modern and more effective alarm systems. As a result, these shops are often the targets of burglars. What is worst is that when a burglary occurs and pawned goods are taken by the burglars, the shop owners are not legally bound to compensate the owners of the property that was stolen.

Text-011: Dieters' Delight

David Mintz is an enterprising businessman who introduced a new product called Tofutti. It is selling briskly in the United States and is giving very stiff (competition) to ice cream. The advantage of Tofutti is that it is a non dairy product and is therefore free of cholesterol. People who are on a fat free (diet) can eat Tofutti without adversely affecting their health. In fact, this delicious product is being served to patients in many hospitals in the United States. Even those people who are allergic to ice cream and other dairy products can safely eat Tofutti, since it is butter free and does not contain any lactose.

The rapid success of this new product is quite amazing, considering the fact that it is made of a very simple and unimpressive ingredient called tofu, which is derived from soya beans. Tofu is a soft and gelatinous substance, but when it is mixed with sweeteners and other ingredients, it can really satisfy a person's sweet tooth. Its inventor David Mintz, is actually not surprised at the enormous success of his formula, since he had expected that Tofutti would sell well at supermarkets and sweet shops, as a very delicious ice cream (substitute) .

Currently, Mintz is working on how to produce tofu based soft drinks and vegetable salads. He is also toying with the idea of producing a non meat burger that would be made basically of tofu. It is predicted that American markets may soon be flooded with delicacies that are made of tofu but which taste and smell like real beefburgers. However, some people in the restaurant business maintain a wait and see (attitude) .

Text-012: The Wonders of Skin Culture

Today, skin can be cultured artificially by taking fragments of undamaged skin from any part of a burn victim's body. Thanks to this miracle of modern science, two very seriously burnt victims, the Selby brothers, are safe and well today.

Six-year-old Glen Selby and his five-year-old brother Jaime were almost burnt to (death) when they played with paint, and then tried to clean themselves with paint remover. Why one of them happened to strike a match remains a (mystery) to this day. As a result, the empty house where the brothers were playing burst into flames in an instant, and the boys sustained very serious burns all over their bodies. They were sent to a nearby hospital, where doctors worked continuously to save their lives. Then they were moved to a children's hospital where doctors tried hard to stabilize their condition.

Finally the brothers were transferred to an institute specialized in dealing with burns. Since there was no other way to save the boys' (lives), the doctors at the institute decided to take fragments of their skin to culture enough skin to cover the damaged areas of their bodies.

Undamaged skin was first taken from the boys' armpits and groins. It was then chemically processed and put in a (flask) containing a special solution. After ten days, considerable quantity of skin was produced in the flask. This greatly encouraged the doctors, who started to culture for more skin for their patients.

After five months, the Selbys were pronounced safe with half of their bodies covered by cultured skin. In the succeeding months, they still had to face more skin (grafts), but the most important thing was that their lives had been saved.

Text-013: Child Abuse in Hong Kong

Child abuse is defined as any non-accidental injury or harm caused to a child. It can also be considered to be a failure of parents or guardians to provide a child with his basic needs. Child abuse can take several forms: physical abuse, gross (neglect) and psychological or emotional abuse.

Physical abuse can be recognized easily from such signs as bruises, burns, and open wounds. Gross neglect takes the form of inadequate health care, intentional deprivation of parental care, and leaving the child alone for hours. As for children who suffer from psychological or emotional abuse, they become withdrawn and their academic performance is often poor.

It is not difficult to identify victims of child abuse. They are usually very sensitive and easily cry at school, and are not keen on taking part in normal physical activities.

Child abuse is often committed by working parents who are frustrated at work and vent their anger on their children. When parents are separated, the spouse who gets (custody) of the child sometimes maltreats the child as a way of releasing his anger or bitterness.

Recent surveys show that there has been an (increase) in child abuse in the sprawling housing estates in Tuen Mun and some other parts of the New Territories. In (view) of this, an organization called Against Child Abuse has decided to test its services in Tuen Mun and to extend them if the Tuen Mun experiment works out.

The problem of child abuse, so far has not become a major social problem in Hong Kong, but there are indications that it might soon grow worse. The government therefore should do something to check the growth of the problem.

Text-014: Writing in Chinese

The Chinese script has two great (beauties). The first is simply the (look) of the characters, as drawn or chiselled in different forms and calligraphic styles down the ages. The other is the singular conceptual beauty of silent (drawings) that enable speakers of mutually incomprehensible dialects to communicate.

But some of the script's users would deny that compared to the Latin, or indeed to any phonetic alphabet, characters are a pain in the neck to learn and use; or, more precisely, a pain somewhere behind the eyes. The first problem is the number of characters. Although some estimates for the total number of characters known to have existed hit the 100000 mark, a mere 1000 common characters will cover 90Unfortunately, without another 3000 or so to cover the next nine percent, the text won't make much sense, and without yet another 5000 or so the reader will lose the last percentile and its more literary flourishes.

There's a lot of characters for a schoolchild to learn and still leaves time for other activities. So one of the big 20th century political and cultural hot potatoes in China has been how to make life easier and thus boost literacy.

The principle behind the idea of substituting an alphabet for characters is very simple. If the sounds of a language make sense when spoken or broadcasted, then all you need is to represent them accurately on paper with a minimum of marks and you have a workable system. The stumbling (step) has always been the prior need to transcribe one standard sound system, that is, one dialect.

D.2 Raw Results of Experiment NAT-2

In Table D.1, we show the results in the study of the information susceptibility of words from the Connotative Approach, via the words in cloze blanks. In the table, we indicate correct disambiguation, incorrect disambiguation, and no response with a bullet (•) a cross (×), and a blank respectively. The results reported here for *f-2g* and *col* are based on uniform prior probabilities; and those for *col* are based on noun co-occurrences only.

Size	±3	±5				±50				Full (MFA)			
Realisation	*f-2g*	*wn*	*rog*	*def*	*col*	*wn*	*rog*	*def*	*col*	*wn*	*rog*	*def*	*com*
trick/5		•	•	•		•	•	•					×
performance/2	×					•	•		•		•		•
spectator/1							•	•					•
instant/1	•						×						
debt/1													
way/2	•								×				×
delay/1		•	•	•		•	•	•			•		•
space/5	×				×		•	×	×		×		×
rhythm/1	×				•	•	•	×	•		•		•
lyric/1	×							•	•				•
panel/2			•			×	•	×			×		×
focus/6							•	•					•
level/1	×				×	×		×	•				×
beam/4	×			•				•	•	•	•	•	×
defect/1											•		
hands/1	•							•	•				
sun/1	×				×	•	•	×	•	•	•	×	•
reflection/5						•	×	•			•		•
mirror/1		•				•	•	•		•			
machine/1	×				•	•	•	×	•	•			×
job/5		×		×		×	×	×		×			×
attention/1	•	×		•	•	×		×	•				×
loss/2	×	×				×		×	•		×		×
aid/2	•	•	•	•		•	×	•	×	•	×	×	•
volunteer/2								•		•			
safety/1	×					×	×		×		×		×
injury/1	×					×	•	×	•	×	•		•
precaution/1			•				•			•	•		×
figure/9	×				×	•	•	•	•		×		•
possibility/2	×				×	×		•	•		•		•
event/2	•	•	•	•	×	•	•	×	×		•		•
evidence/1	•				•	•	×		•		×		×
analysis/1	×					•	•	×	•		×		×
place/1	•				×			×	×		×		×
sense/3	×	•	×			×	×	×	×		×		×
residue/1						•		•		•		•	•
motive/1	×				•				×		×		•
strait/2											×		•
exchange/3								×			×		×
assurance/2											×		×
compensation/1								•			•		
notification/2													×
security/7							×	×			×	×	×
competition/1	•				×				•	×			
diet/1							•			•	•		×
substitute/1	×					×			•	×			×
attitude/1	×			×	•			×	•				×
death/3								×					×
mystery/1	•				•		×		•		×		
life/8				×		×	×	×			×		×
flask/1						×							×
graft/1											•		•
neglect/3		•				•	×	×		•	×		•
custody/3											•		×
increase/3	×						•	×	×		•		•
view/1	•				•	×	×	×	•		×		×
beauty/1	×					•			•				×
look/3	×			×	×	•	•	×			•		×
drawing/2	×							×	×		•		•
step/1	•		•			×	•	×	•				×

Table D.1 *Tabulation of Results in Experiment NAT-2*

References

Special References for Lexical Resources Used:

ROG11 1911. *Roget's Thesaurus.* Crowell Company. (Electronic version prepared by P. Cassidy, MICRA Inc., 1991.)

ROG87 Kirkpatrick, B. 1987. *Roget's Thesaurus of English Words and Phrases.* Penguin Books.

LDOCE Procter, P. 1978. *Longman Dictionary of Contemporary English.* Longman Group Ltd.

LLOCE McArthur, T. 1981. *Longman Lexicon of Contemporary English.* Longman Group Ltd.

WN15 1995. *WordNet 1.5* http://www.cogsci.princeton.edu/~wn/

WN16 1998. *WordNet 1.6* http://www.cogsci.princeton.edu/~wn/

Other References:

Adriaens, G. and S.L. Small. 1988. Word expert parsing revisited in a cognitive science perspective. In S.L. Small, G.W. Cottrell, and M.K. Tanenhaus, editors, *Lexical Ambiguity Resolution: Perspectives from Psycholinguistics, Neuropsychology, and Artificial Intelligence.* Morgan Kaufmann Publishers, Inc., San Mateo, CA, chapter 1.

Agirre, E. and G. Rigau. 1996. Word sense disambiguation using conceptual density. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96)*, pages 16–22, Copenhagen, Denmark.

Amsler, R. 1981. A taxonomy for English nouns and verbs. In *Proceedings of the 19th Annual Meeting of the Association for Computational Linguistics (ACL '81)*, pages 133–138, Stanford.

Binot, J-L. and K. Jensen. 1987. A semantic expert using an online standard dictionary. In *Proceedings of the Tenth International Joint Conference on Artificial Intelligence (IJCAI-87)*, pages 709–714, Milan.

Boggess, L. 1988. Two simple prediction algorithms to facilitate text production. In *Proceedings of the 2nd Conference on Applied Natural Language Processing*, pages 33–40, Austin, Texas.

Boguraev, B. and T. Briscoe. 1987. Large lexicons for natural language processing: Utilising the grammar coding system of LDOCE. *Computational Linguistics*, 13(3-4):203–218.

Boguraev, B. and T. Briscoe. 1989. *Computational Lexicography for Natural Language Processing*. Longman Group UK Limited, Essex, England.

Brown, P.F., S.A. Della Pietra, V.J. Della Pietra, and R.L. Mercer. 1991. Word-sense disambiguation using statistical methods. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL '91)*, pages 264–270, Berkeley, CA.

Bruce, R. and L. Guthrie. 1992. Genus disambiguation: A study in weighted preference. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)*, pages 1187–1191, Nantes, France.

Bruce, R. and J. Wiebe. 1994a. A new approach to word sense disambiguation. In *Proceedings of the 1994 ARPA Human Language Technology Workshop*, pages 236–241.

Bruce, R. and J. Wiebe. 1994b. Word-sense disambiguation using decomposable models. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL '94)*, pages 139–145, Las Cruces, New Mexico.

Byrd, R.J., N. Calzolari, M.S. Chodorow, J.L. Klavans, M.S. Neff, and O.A. Rizk. 1987. Tools and methods for computational lexicology. *Computational Linguistics*, 13(3-4):219–240.

Calzolari, N. 1982. Towards the organization of lexical definitions on a database structure. In E. Hajičovà, editor, *COLING '82 Abstracts*, pages 61–64, Charles University, Prague.

Calzolari, N. 1984. Detecting patterns in a lexical data base. In *Proceedings of the 10th International Conference on Computational Linguistics (COLING 84)*, pages 170–173, Stanford University, California.

Calzolari, N. 1988. The dictionary and the thesaurus can be combined. In M.W. Evens, editor, *Relational Models of the Lexicon: Representing Knowledge in Semantic Networks*. Cambridge University Press.

Calzolari, N. and E. Picchi. 1988. Acquisition of semantic information from an on-line dictionary. In *Proceedings of the 12th International Conference on Computational Linguistics (COLING '88)*, pages 87–92, Budapest, Hungary.

Caraballo, S.A. 1999. Automatic construction of a hypernym-labeled noun hierarchy. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL '99)*, pages 120–126, College Park, Maryland.

Chen, J.N. and J.S. Chang. 1998. Topical clustering of MRD senses based on information retrieval techniques. *Computational Linguistics*, 24(1):61–95.

Chodorow, M.S., R.J. Byrd, and G.E. Heidorn. 1985. Extracting semantic hierarchies from a large on-line dictionary. In *Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics (ACL '85)*, pages 299–304, Chicago.

Chodorow, M.S., Y. Ravin, and H.E. Sachar. 1988. A tool for investigating the synonymy relation in a sense disambiguated thesaurus. In *Proceedings of the Second Conference on Applied Natural Language Processing*, Austin, Texas.

Choueka, Y. and S. Lusignan. 1985. Disambiguation by short contexts. *Computers and the Humanities*, 19:147–157.

Church, K.W. and P. Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Clark, E.V. and H.H. Clark. 1979. When nouns surface as verbs. *Language*, 55(4):767–811.

Collins, A.M. and E.F. Loftus. 1975. A spreading-activation theory of semantic processing. *Psychological Review*, 82(6):407–428.

Copestake, A. and T. Briscoe. 1995. Semi-productive polysemy and sense extension. *Journal of Semantics*, 12:15–67.

Cottrell, G.W. 1988. A model of lexical access of ambiguous words. In S.L. Small, G.W. Cottrell, and M.K. Tanenhaus, editors, *Lexical Ambiguity Resolution: Perspectives from Psycholinguistics, Neuropsychology, and Artificial Intelligence.* Morgan Kaufmann Publishers, Inc., San Mateo, CA, chapter 7.

Covington, M.A. 1994. *Natural Language Processing for Prolog Programmers.* Prentice Hall, Englewood Cliffs, NJ.

Cowie, J., J. Guthrie, and L. Guthrie. 1992. Lexical disambiguation using simulated annealing. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)*, pages 359–365, Nantes, France.

Cruse, D.A. 1986. *Lexical Semantics.* Cambridge University Press.

Dagan, I. and A. Itai. 1994. Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, 20(4):563–596.

Dagan, I., L. Lee, and F. Pereira. 1997. Similarity-based methods for word sense disambiguation. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics (ACL/EACL '97)*, pages 56–63, Madrid, Spain.

Dutch, R.A. 1975. *Roget's Thesaurus of English Words and Phrases (New Ed.)*. Longman, London, U.K.

Even-Zohar, Y. and D. Roth. 2000. A classification approach to word prediction. In *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2000)*, pages 124–131, Seattle, Washington.

Evens, M.W. 1988. *Relational Models of the Lexicon: Representing Knowledge in Semantic Networks*. Cambridge University Press.

Fass, D. 1991. met*: A method of discriminating metonymy and metaphor by computer. *Computational Linguistics*, 17(1):49–90.

Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA.

Fontenelle, T. 1990. Automatic extraction of lexical-semantic relations from dictionary definitions. In *Proceedings of the 4th International Congress on Lexicography (EURALEX '90)*, pages 89–103, Benalmádena, Spain.

Fontenelle, T. 1991. Lexical-semantic representation in lexical data bases. In *Proceedings of the 5th Conference of the European Chapter of the Association for Computational Linguistics (EACL '91)*, Berlin.

Fontenelle, T. and J. Vanandroye. 1989. Retrieving ergative verbs from a lexical database. *Dictionaries: Journal of the Dictionary Society of North America*, 11:11–39.

Fujii, A., K. Inui, T. Tokunaga, and H Tanaka. 1998. Selective sampling for example-based word sense disambiguation. *Computational Linguistics*, 24(4):573–598.

Gale, W., K.W. Church, and D. Yarowsky. 1992a. One sense per discourse. In *Proceedings of the Speech and Natural Language Workshop*, pages 233–237, San Francisco, CA. Morgan Kaufmann.

Gale, W., K.W. Church, and D. Yarowsky. 1992b. Estimating upper and lower bounds on the performance of word-sense disambiguation programs. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics (ACL '92)*, pages 249–256, University of Delaware, Newark, DE.

Guthrie, J.A., L. Guthrie, Y. Wilks, and H. Aidinejad. 1991. Subject-dependent co-occurrence and word sense disambiguation. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL '91)*, Berkeley, CA.

Guthrie, L., B.M. Slator, Y. Wilks, and R. Bruce. 1990. Is there content in empty heads? In *Proceedings of the 13th International Conference on Computational Linguistics (COLING-90)*, volume 3, pages 138–143, Helsinki, Finland.

Harley, A. and D. Glennon. 1997. Sense tagging in action: Combining different tests with additive weightings. In *Proceedings of SIGLEX '97 Workshop: Tagging Text with Lexical Semantics: Why, What, and How?*, pages 74–78, Washington, D.C.

Harris, D.P. 1985. Some forerunners of cloze procedure. *Modern Language Journal*, 69(4):367–376.

Hayes, P.J. 1977. On semantic nets, frames and associations. In *Proceedings of the 5th International Joint Conference on Artificial Intelligence (IJCAI-77)*, pages 99–107, Cambridge, MA.

Hirst, G. 1987. *Semantic Interpretation and the Resolution of Ambiguity*. Cambridge University Press.

Hirst, G. 1988. Resolving lexical ambiguity computationally with spreading activation and Polaroid words. In S.L. Small, G.W. Cottrell, and M.K. Tanenhaus, editors, *Lexical Ambiguity Resolution: Perspectives from Psycholinguistics, Neuropsychology, and Artificial Intelligence*. Morgan Kaufmann Publishers, Inc., San Mateo, CA, chapter 3.

Hirst, G. and D. St-Onge. 1998. Lexical chains as representation of context for the detection and correction of malapropisms. In C. Fellbaum, editor, *WordNet: An Electronic Lexical Database*. The MIT Press.

Hurford, J.R. and B. Heasley. 1983. *Semantics: A Coursebook*. Cambridge University Press.

Ide, N. and J. Veronis. 1998. Introduction to the special issue on word sense disambiguation: The state of the art. *Computational Linguistics*, 24(1):1–40.

Jonz, J. 1976. Improving on the basic egg: The M-C cloze. *Language Learning*, 26(2):255–265.

Jorgensen, J. 1990. The psychological reality of word senses. *Journal of Psycholinguistic Research*, 19:167–190.

Kakigahara, K. and T. Aizawa. 1988. Completion of Japanese sentences by inferring function words from content words. In *Proceedings of the 12th International Conference on Computational Linguistics (COLING '88)*, pages 291–296, Budapest.

Karov, Y. and S. Edelman. 1998. Similarity-based word sense disambiguation. *Computational Linguistics*, 24(1):41–59.

Kawamoto, A.H. 1988. Distributed representations of ambiguous words and their resolution in a connectionist network. In S.L. Small, G.W. Cottrell, and M.K. Tanenhaus, editors, *Lexical Ambiguity Resolution: Perspectives from Psycholinguistics, Neuropsychology, and Artificial Intelligence*. Morgan Kaufmann Publishers, Inc., San Mateo, CA, chapter 8.

Kilgarriff, A. 1992. *Polysemy*. Ph.D. thesis, University of Sussex.

Kilgarriff, A. 1997a. "I don't believe in word senses". *Computers and the Humanities*, 31(2):91–113.

Kilgarriff, A. 1997b. What is word sense good for? In *Proceedings of the 3rd Natural Language Processing Pacific Rim Symposium (NLPRS '97)*, pages 209–214, Phuket, Thailand.

Kilgarriff, A. 1998a. SENSEVAL: An exercise in evaluating word sense disambiguation programs. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC)*, pages 581–588, Granada, Spain.

Kilgarriff, A. 1998b. Gold standard datasets for evaluating word sense disambiguation programs. *Computer Speech and Language, Special Issue on Evaluation*, 12(3):453–472.

Kilgarriff, A. 1999. 95% replicability for manual word sense tagging. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL '99)*, pages 277–278, Bergen, Norway.

Kilgarriff, A. and J. Rosenzweig. 1999. English SENSEVAL: Reports and Results. In *Proceedings of the 5th Natural Language Processing Pacific Rim Symposium (NLPRS '99)*, Beijing, China.

Klavans, J. 1988. COMPLEX: A computational lexicon for natural language processing. In *Proceedings of the 12th International Conference on Computational Linguistics (COLING '88)*, pages 815–823, Budapest, Hungary.

Klavans, J., M. Chodorow, and N. Wacholder. 1990. From dictionary to knowledge base via taxonomy. In *Proceedings of the Sixth Conference of the University of Waterloo*, pages 110–132, Canada.

Klavans, J. and E. Tzoukermann. 1995. Combining corpus and machine-readable dictionary data for building bilingual lexicons. *Machine Translation*, 10:185–218.

Knight, K. and S.K. Luk. 1994. Building a large-scale knowledge base for machine translation. In *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94)*, pages 773–778, Seattle, Washington.

Krovetz, R. 1992. Sense-linking in a machine readable dictionary. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics (ACL '92)*, pages 330–332, University of Delaware, Newark, DE.

Krovetz, R. 1997. Homonymy and polysemy in information retrieval. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics (ACL/EACL '97)*, pages 72–79, Madrid, Spain.

Krovetz, R. and W.B. Croft. 1992. Lexical ambiguity and information retrieval. *ACM Transactions on Information Systems*, 10(2):115–141.

Kwong, O.Y. 1998. Aligning WordNet with additional lexical resources. In *Proceedings of the Workshop on the Usage of WordNet in Natural Language Processing Systems*, pages 73–79, Montréal, Canada. COLING-ACL '98.

Landes, S., C. Leacock, and R.I. Tengi. 1998. Building semantic concordances. In C. Fellbaum, editor, *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA, chapter 8.

Leacock, C., M. Chodorow, and M.A. Miller. 1998. Using corpus statistics and WordNet relations for sense identification. *Computational Linguistics*, 24(1):147–165.

Leacock, C., G. Towell, and E. Voorhees. 1993. Corpus-based statistical sense resolution. In *Proceedings of the ARPA Human Language Technology Workshop*, pages 260–265, Princeton, NJ.

Lesk, M.E. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 1986 SIGDOC Conference*, pages 24–26, Toronto, Canada.

Lin, D. 1997. Using syntactic dependency as local context to resolve word sense ambiguity. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics (ACL/EACL '97)*, pages 64–71, Madrid, Spain.

Lyons, J. 1981. *Language and Linguistics: An Introduction*. Cambridge University Press.

Lytinen, S.L. 1988. "Are vague words ambiguous?". In S.L. Small, G.W. Cottrell, and M.K. Tanenhaus, editors, *Lexical Ambiguity Resolution: Perspectives from Psycholinguistics, Neuropsychology, and Artificial Intelligence*. Morgan Kaufmann Publishers, Inc., San Mateo, CA, chapter 4.

Macabalitaw, L.D. 1987. *Use of English: Reading and Language Systems*. Witman Publishing Co. (H.K.) Ltd.

Mandala, R., T. Tokunaga, and H. Tanaka. 1999. Complementing WordNet with Roget's and corpus-based thesauri for information retrieval. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL '99)*, pages 94–101, Bergen, Norway.

Markowitz, J., T. Ahlswede, and M. Evens. 1986. Semantically significant patterns in dictionary definitions. In *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics (ACL '86)*, pages 112–119, New York.

McCarthy, D. 1997. Word sense disambiguation for acquisition of selectional preferences. In P. Vossen, G. Adriaens, N. Calzolari, A. Sanfilippo, and Y. Wilks, editors, *Proceedings of the Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 52–60, Madrid, Spain. ACL/EACL '97.

McHale, M.L., 1995. *Combining Machine-Readable Lexical Resources with a Principle-Based Parser*, chapter 6. Ph.D. Dissertation, Syracuse University, New York.

McHale, M.L. 1998. A comparison of WordNet and Roget's taxonomy for measuring semantic similarity. In *Proceedings of the Workshop on the Usage of WordNet in Natural Language Processing Systems*, pages 115–120, Montréal, Canada. COLING-ACL '98.

McHale, M.L. and J.J. Crowter. 1994. Constructing a lexicon from a machine readable dictionary. Technical Report RL-TR-94-178, Rome Laboratory, Griffiss Air Force Base, New York.

McRoy, S.W. 1992. Using multiple knowledge sources for word sense disambiguation. *Computational Linguistics*, 18(1):1–30.

Michiels, A. and J. Noël. 1982. Approaches to thesaurus production. In J. Horecky, editor, *Proceedings of the 9th International Conference on Computational Linguistics (COLING '82)*, pages 227–232. Amsterdam: North-Holland.

Mihalcea, R. and D.I. Moldovan. 1999a. A method for word sense disambiguation of unrestricted text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL '99)*, pages 152–158, College Park, Maryland.

Mihalcea, R. and D.I. Moldovan. 1999b. An automatic method for generating sense tagged corpora. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99)*, pages 461–466, Orlando, Florida.

Miller, G.A. 1990. WordNet: An on-line lexical database. *Special Issue of International Journal of Lexicography*, 3(4).

Miller, G.A. 1998. Nouns in WordNet. In C. Fellbaum, editor, *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA, chapter 1.

Miller, G.A., R. Beckwith, C. Fellbaum, D. Gross, and K.J. Miller. 1990. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–244.

Miller, G.A. and W.G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

Miller, G.A., R. Tengi, and S. Landes. 1999. Matching the tagging to the task. *Natural Language Engineering*, 5(2):135–145.

Morris, J. and G. Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21–48.

Moss, H. and L. Older. 1996. *Birkbeck Word Association Norms*. Psychology Press, Hove, U.K.

Nakamura, J. and M. Nagao. 1988. Extraction of semantic information from an ordinary English dictionary and its evaluation. In *Proceedings of the 12th International Conference on Computational Linguistics (COLING '88)*, pages 459–464, Budapest, Hungary.

Ng, H.T. 1997. Getting serious about word sense disambiguation. In *Proceedings of SIGLEX '97 Workshop: Tagging Text with Lexical Semantics: Why, What, and How?*, Washington, D.C.

Ng, H.T. and H.B. Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL '96)*, pages 40–47, Santa Cruz, CA.

Oller, J.W. 1994. *Cloze and Coherence.* Associated University Presses, Cranbury, NJ.

Pedersen, T. 2000. A simple approach to building ensembles of naive Bayesian classifiers for word sense disambiguation. In *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2000)*, pages 63–69, Seattle, Washington.

Porter, M.F. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137, July.

Pustejovsky, J. 1991. The generative lexicon. *Computational Linguistics*, 17(4):409–441.

Pustejovsky, J. and B. Boguraev. 1993. Lexical knowledge representation and natural language processing. *Artificial Intelligence*, 63:193–223.

Quillian, M.R. 1968. Semantic memory. In M. Minsky, editor, *Semantic Information Processing.* MIT Press, Cambridge, MA.

Rada, R., H. Mili, E. Bicknell, and M. Blettner. 1989. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(1):17–30.

Rapp, R. and M. Wettler. 1991. Prediction of free word associations based on Hebbian learning. In *Proceedings of the International Joint Conference on Neural Networks*, volume 1, pages 25–29, Singapore.

Resnik, P. 1993. *Selection and Information: A Class-Based Approach to Lexical Relationships.* Doctoral Dissertation, Department of Computer and Information Science, University of Pennsylvania.

Resnik, P. 1995a. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI-95)*, pages 448–453, Montréal, Canada.

Resnik, P. 1995b. Disambiguating noun groupings with respect to WordNet senses. In *Proceedings of the Third Workshop on Very Large Corpora*, pages 54–68, Cambridge, MA.

Resnik, P. 1997. Selectional preference and sense disambiguation. In *Proceedings of SIGLEX '97 Workshop: Tagging Text with Lexical Semantics: Why, What, and How?*, Washington, D.C.

Resnik, P. and D. Yarowsky. 1997. A perspective on word sense disambiguation methods and their evaluation. In *Proceedings of SIGLEX '97 Workshop: Tagging Text with Lexical Semantics: Why, What, and How?*, pages 79–86, Washington, D.C.

Resnik, P. and D. Yarowsky. 1999. Distinguishing systems and distinguishing senses: new evaluation methods for Word Sense Disambiguation. *Natural Language Engineering*, 5(2):113–133.

Ribas, F. 1994. An experiment on learning appropriate selectional restrictions from a parsed corpus. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING 94)*, pages 769–774, Kyoto, Japan.

Ribas, F. 1995. On learning more appropriate selectional restrictions. In *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics (EACL '95)*, Dublin, Ireland.

Richardson, R., A.F. Smeaton, and J. Murphy. 1994. Using WordNet as a knowledge base for measuring semantic similarity between words. Technical Report Working Paper CA-1294, Dublin City University, Ireland.

Rigau, G., J. Atserias, and E. Agirre. 1997. Combining unsupervised lexical knowledge methods for word sense disambiguation. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics (ACL/EACL '97)*, pages 48–55, Madrid, Spain.

Riloff, E. and R. Jones. 1999. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99)*, pages 474–479, Orlando, Florida.

Riloff, E. and J. Shepherd. 1997. A corpus-based approach for building semantic lexicons. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, pages 117–124, Providence, RI.

Riloff, E. and J. Shepherd. 1999. A corpus-based bootstrapping algorithm for Semi-Automated semantic lexicon construction. *Natural Language Engineering*, 5(2):147–156.

Roark, B. and E. Charniak. 1998. Noun-phrase co-occurrence statistics for semi-automatic semantic lexicon construction. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL '98)*, pages 1110–1116, Montréal, Canada.

Roget, P.M. 1852. *Roget's Thesaurus of English Words and Phrases*. Penguin Books.

Rye, J. 1982. *Cloze Procedure and the Teaching of Reading*. Heinemann Educational, London, U.K.

Sanderson, M. 1994. Word sense disambiguation and information retrieval. In *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR '94)*, pages 142–151, Dublin, Ireland.

Sanfilippo, A. 1997. Using semantic similarity to acquire co-occurrence restrictions from corpora. In P. Vossen, G. Adriaens, N. Calzolari, A. Sanfilippo, and Y. Wilks, editors, *Proceedings of the Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 82–89, Madrid, Spain. ACL/EACL '97.

Sanfilippo, A. and V. Poznański. 1992. The acquisition of lexical knowledge from combined machine-readable dictionary sources. In *Proceedings of the 3rd Conference on Applied Natural Language Processing*, pages 80–87, Trento, Italy.

Schütze, H. 1992. Dimensions of meaning. In *Proceedings of Supercomputing*, pages 787–796, Minneapolis, MN.

Schütze, H. 1997. *Ambiguity Resolution in Language Learning: Computational and Cognitive Models*. Center for the Study of Language and Information, Stanford, CA.

Schütze, H. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Sinclair, J. 1987. *Collins COBUILD English Language Dictionary*. HarperCollins, London, U.K.

Small, S. and C. Rieger. 1982. Parsing and comprehending with word experts (A theory and its realization). In W.G. Lehnert and M.H. Ringle, editors, *Strategies for Natural Language Processing*. Lawrence Erlbaum Associates, New Jersey.

Sparck Jones, K. 1986. *Synonymy and Semantic Classification*. Edinburgh University Press.

Stevenson, M. and Y. Wilks. 1999. Combining weak knowledge sources for sense disambiguation. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI-99)*, Stockholm, Sweden.

Tabossi, P. 1989. What is a context? In D.S. Gorfein, editor, *Resolving Semantic Ambiguity*. New York: Springer-Verlag.

Tengi, R.I. 1998. Design and Implementation of the WordNet Lexical Database and Searching Software. In C. Fellbaum, editor, *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA, chapter 4.

Towell, G. and E.M. Voorhees. 1998. Disambiguating highly ambiguous words. *Computational Linguistics*, 24(1):125–145.

van Rijsbergen, C.J. 1979. *Information Retrieval*. Butterworth & Co. (Publishers) Ltd., London, U.K., second edition.

Veronis, J. and N. Ide. 1990. Word sense disambiguation with very large neural networks extracted from machine readable dictionaries. In *Proceedings of the 13th International Conference on Computational Linguistics (COLING-90)*, pages 389–394, Helsinki, Finland.

Vossen, P. and A. Copestake. 1993. Untangling definition structure into knowledge representation. In T. Briscoe, A. Copestake, and V. de Paiva, editors, *Inheritance, Defaults and the Lexicon*. Cambridge University Press, chapter 13.

Vossen, P., W. Meijs, and M. den Broeder. 1989. Meaning and structure in dictionary definitions. In B. Boguraev and T. Briscoe, editors, *Computational Lexicography for Natural Language Processing*. Longman Group, Essex, England, chapter 8.

Wettler, M. and R. Rapp. 1993. Computation of word associations based on the co-occurrences of words in large corpora. In *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*, pages 84–93, Columbus, Ohio.

Wilks, Y. 1975a. An intelligent analyzer and understander of English. *Communications of the ACM*, 18(5):264–274.

Wilks, Y. 1975b. A preferential, pattern-seeking, semantics for natural language inference. *Artificial Intelligence*, 6:53–74.

Wilks, Y. 1978. Making preferences more active. *Artificial Intelligence*, 11:197–223.

Wilks, Y. 1998. Senses and texts. *Computational Linguistics and Chinese Language Processing*, 3(2):1–16, August.

Wilks, Y., D. Fass, C-M. Guo, J. McDonald, T. Plate, and B. Slator. 1989. A tractable machine dictionary as a resource for computational semantics. In B. Boguraev and T. Briscoe, editors, *Computational Lexicography for Natural Language Processing*. Longman Group, Essex, England, chapter 9.

Wilks, Y. and M. Stevenson. 1996. The grammar of sense: Is word-sense tagging much more than part-of-speech tagging? Technical Report CS-96-05, University of Sheffield.

Wilks, Y. and M. Stevenson. 1997. Sense tagging: Semantic tagging with a lexicon. In *Proceedings of SIGLEX '97 Workshop: Tagging Text with Lexical Semantics: Why, What, and How?*, Washington, D.C.

Wilks, Y. and M. Stevenson. 1998. Word sense disambiguation using optimised combinations of knowledge sources. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL '98)*, pages 1398–1402, Montréal, Canada.

Wittgenstein, L. 1958. *Philosophical Investigations*. Blackwell Publishers. Translated by G.E.M. Anscombe.

Yarowsky, D. 1992. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)*, pages 454–460, Nantes, France.

Yarowsky, D. 1993. One sense per collocation. In *Proceedings of the ARPA Human Language Technology Workshop*, pages 266–271, Princeton, NJ.

Yarowsky, D. 1994. Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL '94)*, pages 88–95, Las Cruces, New Mexico.

Yarowsky, D. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL '95)*, pages 189–196, Cambridge, MA.

Kirkpatrick, B. 1987. *Roget's Thesaurus of English Words and Phrases*. Penguin Books.

McArthur, T. 1981. *Longman Lexicon of Contemporary English*. Longman Group Ltd., Essex, England.

Procter, P. 1978. *Longman Dictionary of Contemporary English*. Longman Group Ltd.