Technical Report

UCAM-CL-TR-764
ISSN 1476-2986

Number 764

UNIVERSITY OF CAMBRIDGE

Computer Laboratory

Statistical anaphora resolution in biomedical texts

Caroline V. Gasperin

December 2009

15 JJ Thomson Avenue
Cambridge CB3 0FD
United Kingdom
phone +44 1223 763500

http://www.cl.cam.ac.uk/

This technical report is based on a dissertation submitted August 2008 by the author for the degree of Doctor of Philosophy to the University of Cambridge, Clare Hall.

Technical reports published by the University of Cambridge Computer Laboratory are freely available via the Internet:

http://www.cl.cam.ac.uk/techreports/

ISSN 1476-2986

Statistical anaphora resolution in biomedical texts
Caroline Varaschin Gasperin

Summary

This thesis presents a study of anaphora in biomedical scientific literature and focuses on tackling the problem of anaphora resolution in this domain. Biomedical literature has been the focus of many information extraction projects; there are, however, very few works on anaphora resolution in biomedical scientific full-text articles. Resolving anaphora is an important step in the identification of mentions of biomedical entities about which information could be extracted.

We have identified coreferent and associative anaphoric relations in biomedical texts. Among associative relations we were able to distinguish 3 main types: biotype, homolog and set-member relations. We have created a corpus of biomedical articles that are annotated with anaphoric links between noun phrases referring to biomedical entities of interest. Such noun phrases are typed according to a scheme that we have developed based on the Sequence Ontology; it distinguishes 7 types of entities: gene, part of gene, product of gene, part of product, subtype of gene, supertype of gene and gene variant.

We propose a probabilistic model for the resolution of anaphora in biomedical texts. The model seeks to find the antecedents of anaphoric expressions, both coreferent and associative, and also to identify discourse-new expressions. The model secures good performance despite being trained on a small corpus: it achieves 55-73% precision and 57-63% recall on coreferent cases, and reasonable performance on different classes of associative cases. We compare the performance of the model with a rule-based baseline system that we have also developed, a naive Bayes system and a decision trees system, showing that the ours outperforms the others. We have experimented with active learning in order to select training samples to improve the performance of our probabilistic model. It was not, however, more successful than random sampling.

Acknowledgements

I am very glad that I could count on the support of so many people during my PhD and I would like to thank all of them.

I thank my supervisor Ted Briscoe for his close guidance, encouragement and patience.

I thank the FlySlip project members, Ian Lewin, Nikiforos Karamanis, Andreas Vlachos and Ruth Seal, for the fruitful discussions that contributed to this thesis. Special thanks to Nikiforos and Ruth for their efforts on annotating the corpus.

I thank the CAPES foundation (Brazilian Ministry of Education) for funding my studies, making possible the great experience that the time in Cambridge has been.

I thank Clare Hall for the warm college atmosphere.

Special thanks go to my Brazilian friends in Cambridge, who made me feel home and helped me discover more about my country despite being so far: Juliano Iyoda, Leda Sampson, Pedro and Caroline Anselmo, Carlos Hotta and Paula Signorini, Andre Sartori, Cristiana Viegas and Yuri Sobral, you are special friends.

Thanks to my Cambridge friends, specially Pelin Akan for her sweetness, and my housemates, Richard Southern, Daniel Lackner, Susie Maidment and Anja Hagemann for their enjoyable company.

Thanks to my family for their constant encouragement and unconditional support; knowing that I can always count on them is invaluable.

Finally, thanks to Saar Drimer for his love and comfort.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis presents a study of anaphora on biomedical scientific literature and tackles the problem of anaphora resolution in this domain.

Anaphora is the relation between two linguistic expressions in the discourse where the reader is referred back to the first when reading the second later in the text. The referring expression is usually called anaphor, and the previous expression it is associated with is called antecedent. This reference process can be supported by several relations between the entities represented by the expressions. When both linguistic expressions refer to the same entity, the relation between them is called coreference.

The concepts of anaphora and coreference have been used in different ways in the literature, causing some confusion in the field. van Deemter & Kibble [2000] have sought to distinguish them. They define coreference simply as reference to the same entity, while anaphora implies a dependency between two expressions, when the one which occurs later in the discourse depends on the previous one to be correctly interpreted. Coreference and anaphora can occur together or separately. Anaphora can happen between expressions referring to distinct (but associated) entities; in such cases it is called associative anaphora.

In this work we deal with the union of the two concepts: expressions that are simply coreferent, expressions that are only anaphoric, and expressions that hold both relations. Throughout this thesis I shall refer to all these possible relations as anaphora, expect in cases where distinguishing both concepts is necessary.

Anaphora resolution can be understood as the process of identifying an anaphoric relation between two expressions in a text and consequently linking the two, one being the anaphor and the other the antecedent. Resolving anaphora is a very important step in the text processing pipeline for executing tasks that require a full picture of the elements involved in the discourse and their relevance. Examples of such tasks are information extraction [Gaizauskas and Humphreys, 2000], text summarisation [Boguraev and Kennedy, 1999], and question answering [Watson *et al.*, 2003].

Different kinds of noun phrases (NPs) present anaphoric behaviour: pronouns, definite descriptions (NPs introduced by the definite article 'the'), demonstrative NPs (NPs that start with a demonstrative pronoun such as 'this', 'these'), proper names, among others. Much of the work done on anaphora resolution deal only with pronouns [Lappin and Leass, 1994, Kennedy and Boguraev, 1996, Mitkov, 1998]. Strategies for resolution of pronouns differ from approaches for resolution of non-pronominal NPs because the scope in which to look for the antecedent of a pronoun is known to be considerably smaller than that of non-pronominal NPs, and consequently different types of clues need to be used to identify the correct antecedent.

Non-pronominal NPs vary considerably in the distance at which they can be found from the antecedent, and also in the frequency in which they are anaphoric or not. For example, demonstrative NPs are known to be anaphoric most of the time and have a small scope of search for their antecedents (but greater than for pronouns), while definite descriptions are frequently found not to be anaphoric, and when they are, they are usually used to recall an

entity that has been mentioned a few or several sentences earlier.

The methods for resolution of non-pronominal NPs have to be capable of distinguishing which of them are anaphoric and also selecting the correct antecedent from a broad scope of candidates. The information available for resolution of non-pronominal NPs is also different from that available for resolution of pronouns. For example, while pronoun resolution may rely on syntactic binding constraints given anaphor and antecedent proximity, these do not hold for resolution of other types of NPs. On the other hand, resolution of non-pronominal NPs can benefit from lexical information present in the NP, that is, the words that form the NP, which does not happen for pronouns. Different systems have been proposed for resolution of non-pronominal NPs: Vieira and Poesio [2000]'s system for resolution of definite descriptions only, and systems for treating coreference of all types of NPs, including pronouns, such as [Soon *et al.*, 2001, Ng and Cardie, 2002b, Strube *et al.*, 2002] and the systems participating in the Coreference Task of the Message Understanding Conferences (MUC-6 and MUC-7) [MUC, 1995, MUC, 1998].

Anaphora resolution systems have been developed and tested in different genres of text, e.g. news articles, technical manuals, literary texts and scientific papers. Each genre of text presents a different distribution of the types of anaphoric NPs. For example, technical manuals contain many more pronouns than scientific texts, which contain very few of them, while biomedical scientific texts have a larger proportion of proper names than do newspaper texts.

We decided to investigate anaphora in biomedical scientific articles. In the biomedical field, the constant growth in the number of scientific publications makes it difficult for researchers to keep track of information, even in very small subfields of biology, and there is a real need for automatic information extraction, in which anaphora resolution is an essential step. Currently, progress in the field often relies on the work of professional curators, typically postdoctoral scientists who are trained to identify important information in a scientific article and place it in a template in a database that will be accessed by the research community later on. This is obviously a very time-consuming task.

Our decision to focus on the biomedical domain is not only related to the growing demand for up-to-date biomedical information. We also consider that the availability of manually-built knowledge sources (e.g. databases, ontologies) for the biomedical domain can provide valuable semantic information about the entities mentioned in the text. Such information can be really valuable for anaphora resolution since it can provide semantic classification for the entities mentioned in the text. This allows us to explore resolution techniques that require semantic information. Besides, the great majority of the entities in biomedical texts are referred to using non-pronominal NPs; this suited our goal of exploring anaphora resolution methods for anaphoric NPs other than pronouns, which have proven more challenging and have been less researched into. Hence we focus on these NPs and do not investigate pronominal reference.

Our objective in this thesis is to reach a better understanding of the anaphoric relations present in full-text biomedical articles, to develop the resources that would enable us to propose a corpus-based anaphora resolution system for this domain, and finally to implement a system that is able to resolve anaphora in these texts.

To develop a system for anaphora resolution in biomedical texts, it was necessary first to accomplish named-entity recognition and semantic tagging, besides developing a corpus for training and/or evaluation of the system, since there was no corpus of full-text biomedical articles annotated with anaphoric links that could be used for training an anaphora resolution system.

These efforts were part of the FlySlip project[1], whose ultimate goal was to develop a tool for facilitating the curation of scientific articles, and for which it was necessary to develop the

[1] http://www.wiki.cl.cam.ac.uk/rowiki/NaturalLanguage/FlySlip

infrastructure to process the articles. The FlySlip project was linked to the FlyBase project[2] in the Department of Genetics of the University of Cambridge, whose focus is the molecular biology literature related to fruit fly genomics.

To benefit from the resources available through the FlySlip project (archive of articles, expert curators, tools developed in the scope of the project, such as gene-name recognizer), we have opted for restricting our study to this subset of biomedical literature, the molecular biology articles related to the fruit fly. We regard all mentions of biomedical entities as the anaphoric expressions of interest to our study. More precisely, we focus on genes and other entities related to genes such as proteins and parts of genes.

The very few works on anaphora resolution in the biomedical domain developed so far have used abstracts of scientific papers instead of full text. We consider, however, that abstracts represent a very restricted use of anaphora, since anaphora is a phenomenon that develops through the text.

We have developed a probabilistic system for anaphora resolution in full-text biomedical articles. Our probabilistic model collects statistics from the training corpus that we have built. The model is an adaptation of the work of Ge *et al.* [1998] for pronoun resolution for the resolution of non-pronominal NPs. It is based on the decomposition of a probability conditional to several features into the product of few probabilities conditional to fewer features.

Our model aims to discover both coreferent and associative anaphoric relations between biomedical entities, as well as identify which of them are not anaphoric, that is, should not be assigned an antecedent. This is the first work on anaphora resolution in the biomedical domain that also deals with associative anaphora.

In the following section, we outline the main contributions of this work.

1.1 Contributions of this thesis

We have developed as part of the work presented in this thesis:

- a strategy for identifying and classifying noun phrases referring to biomedical entities in the text: Given our focus on the molecular biology subdomain, we adopted the Sequence Ontology for use as part of a dictionary-based system for the recognition and typing of the NPs of interest in the text. We describe this strategy in Chapter 4.

- an evaluation and training corpus: We have developed guidelines for the annotation of anaphora relations in full-text biomedical scientific articles, and used these to create an annotated corpus for training and evaluation of an anaphora resolution system. The guidelines include the annotation of coreferent and associative anaphoric cases, including domain-related kinds of associative anaphora. The corpus annotation process and the corpus developed are described in Chapter 5. This is the first corpus of anaphoric relations in full-text biomedical articles that has been developed.

- an anaphora resolution system: We have initially developed a baseline rule-based system for resolving anaphora in biomedical texts, which is described in Chapter 6. As the main contribution of this thesis, we have developed a probabilistic anaphora resolution system, which aims to resolve coreferent and associative anaphora cases. This system is trained on the annotated corpus and, despite the small amount of training data, reaches better performance than the baseline system. It is described in Chapter 7.

- active learning for anaphora resolution: Aiming to enhance the performance of the probabilistic system, we developed a complementary active learning strategy. This strategy has

[2] http://www.gen.cam.ac.uk/Research/flybase.htm

not been successful but our experiments can contribute to future attempts to use active learning for anaphora resolution. These experiments are detailed in Chapter 8.

1.2 Thesis overview

In the next chapter we describe the research area of biomedical text mining, within which our work fits. In Chapter 3 we present an overview of the research on anaphora resolution. In Chapter 4 we describe the process of identifying and typing the NPs that refer to biomedical entities. In Chapter 5 we discuss the anaphoric relations that we identified in biomedical texts and describe the process of manually annotating a corpus with such relations. In Chapter 6 we present a rule-based baseline system for the resolution of the anaphoric relations present in our corpus. In Chapter 7 we describe our probabilistic model for anaphora resolution in biomedical texts. In Chapter 8 we present experiments on an active learning strategy in order to improve the performance of the probabilistic model. In Chapter 9 we present our conclusions of this work and suggest directions for future work.

Chapter 2

Biomedical information extraction

New findings in Biology research are built upon already discovered characteristics of biomedical entities and relations among them, and easy access to this information in specific databases is vital for researchers. However, according to Hirschman *et al.* [2002], new information relevant to Biology research is still recorded as free text in journal articles and in free-text fields of databases. The number of articles published in biomedical journals per year is increasing exponentially, making it difficult for researchers to keep track of information [Morgan *et al.*, 2003]; more than 600 000 biomedical journal articles were published in the year 2007 according to PubMed[1], and more than 2 800 in relation to the Drosophila fruit fly according to FlyBase[2][3].

Projects like FlyBase employ full-time curators to read the relevant recently published papers and record the useful information in a template form that can then be updated into the database. The curators are typically postdoctoral scientists, trained to identify important information in a scientific article. This is a very time-consuming task which requires identification of gene, allele and protein names and their synonyms, as well as several interactions and relations between them. The information extracted from each article is used to fill in a template per gene or allele mentioned in the article.

This demand for information from the biomedical field has encouraged many researchers to efforts in developing natural language processing (NLP) tools to extract information from biomedical scientific articles. Different levels of information have been targeted by NLP projects, for example, recognising the names of biomedical entities (e.g. genes and proteins), identifying relations between these entities (e.g. interaction between proteins), linking various expressions in the text that refer to the same or related entities, among others.

Biomedical texts impose additional challenges to the realisation of these tasks in comparison with newspaper texts, which have been more widely used for developing and testing NLP tools. On the other hand, NLP can benefit from the several sources of refined semantic knowledge that are not commonly available for other domains; these are biomedical resources such as databases, ontologies, and terminologies.

In the next section, we shall discuss the special features of biomedical texts. Section 2.2 describes some of the resources available. Section 2.3 discusses the tasks that have been dealt with so far in exploring biomedical texts using NLP.

2.1 Specificities of biomedical text

Biomedical texts differ significantly from other text genres such as newspapers and fiction writing.

In biomedical texts, much background knowledge is required for the reader to understand the relation between the entities mentioned in the text. This is a common aspect of scientific

[1]PubMed is a database of biomedical literature: http://www.ncbi.nlm.nih.gov/sites/entrez

[2]FlyBase is a database of genomic research on the fruit fly *Drosophila melanogaster*: http://www.flybase.org

[3]These numbers were collected by searching PubMed and Flybase, respectively, for journal articles published in 2007

papers in general. For example, if an expression in the text refers to gene x and later on an unnamed protein is mentioned, it is likely that the writer refers to the protein encoded by gene x, and the reader can only understand that if he/she knows that genes encode proteins. To "understand" the relation between the entities in the text automatically, a system would need a domain ontology that encodes the known relations. For example, the Gene Ontology relates genes with the cellular components (e.g. cytoplasm, X chromosome) within which they are located; the Sequence Ontology relates the gene to parts of its sequence (e.g. exon, intron) and to its products (e.g. protein). We shall describe these and other ontologies in more detail in Section 2.2.

Usually a gene and the protein it encodes share the same name, causing some ambiguity in the text when the context does not provide enough information to determine whether the writer is talking about the gene or the protein. To avoid this ambiguity, writing conventions have been proposed, such as writing gene names in lowercase italicised letters and protein names in non-italicised uppercase letters[4]. It is common, however, that authors do not follow these conventions strictly, and distinct entities end up being referred to by the same string. Besides, gene/protein names may coincide with common English words, e.g. *for* (symbol for *foraging*), a fruit fly gene; with parts of the body of the organism on which it has an effect, e.g. *giant fibre*, a fruit fly gene that influences the behaviour of the giant fibre in the brain of the fly; and with the name of the disease associated with the gene disorder, e.g. *Huntington Disease*, a human gene. These sources of ambiguity impose extra challenges to a system that aims to recognise gene and protein names automatically.

Biomedical texts also have a large quantity of acronyms and abbreviations, which may be gene symbols or refer to other biomedical concepts. Such concepts can be introduced in full form by the author, preceded or followed by its abbreviated form, e.g. `CT0 (circadian time 0)` and `DCC (dosage compensation complex)`, or common knowledge is assumed and the acronyms are used from the first reference, e.g. `PCR` (polymerase chain reaction), `UAS` (upstream activation sequence), `RNAi` (RNA interference). These acronyms make the task of identifying gene names even more challenging: a gene-name recogniser that relies on the morphological form of the words (for example, characterising gene/protein names as tokens that contain letters and numbers, upper and lowercase letters, other special characters) may mistag acronyms as gene names.

The distribution of different types of noun phrases in biomedical articles also differs from the distribution in other genres of text. For example, pronouns are very rare, accounting for about 3% of noun phrases[5]; on the other hand, proper names are very frequent, given the frequent mention of gene, allele and protein names and the names of other biomedical entities. This aspect of biomedical text is directly relevant when developing a system to link noun phrases related to a specific entity, because different types of noun phrases have distinctive ways of referring back to a previously-mentioned entity in the text. Such a system should focus on the features of the most common types of noun phrases, that is, non-pronominal ones. Section 2.3.3 introduces the role of such a system in exploring biomedical texts.

Unlike other scientific articles, biomedical articles include a considerable amount of information written as captions of figures rather than in the body of the paper, since figures play an important role in describing biological experiments. Some of this information can not be found anywhere else in the text. For this reason, captions should not be ignored when processing the

[4]FlyBase conventions: http://www.flybase.org/static_pages/docs/nomenclature/nomenclature3.html#10 ; WormBase conventions: http://www.wormbase.org/wiki/index.php/UserGuide:Nomenclature

[5]According to the corpus created as part of this thesis, presented in Chapter 5. Newspaper texts have a slightly higher percentage of pronouns – for example, in the Wall Street Journal corpus 4.5% of noun phrases are pronouns –; fictional texts have a much higher rate – in the portion of fiction writing of the Brown corpus, 22% of NPs are pronouns.

text in order to extract information about the biomedical entities.

Another particular characteristic of biomedical articles is their logical organisation, which is often the same. Most articles reporting experimental work have an introduction, followed by a results section, discussion and a material and methods section. This aspect of biomedical texts can guide information extraction efforts to look for specific portions of text where the required information is more likely to be found.

2.2 Available resources

NLP research benefits from work on the biomedical domain, given the availability of specialised knowledge sources such as terminologies, ontologies and databases, which are scarce or nonexistent in other domains. Such resources allow researchers to go a step further in their work, enabled to make use of techniques that require this kind of knowledge. These resources, despite not having been developed primarily for text processing, can provide knowledge for NLP tasks, from lexical (e.g. gene names, domain-specific terms) to semantic (e.g. domain-specific relations between entities). Below we shall describe some of the most popular resources that can be used for the processing of biomedical texts.

2.2.1 Databases

For most model organisms[6], there is a dedicated genomic database where information about its genes is recorded, such as MGI[7] for the mouse, FlyBase for the fruit fly, WormBase[8] for the worm, and SGD[9] for yeast, among others. Each gene entry contains information including the gene name and symbol, synonyms for the gene name that are found in the literature, a brief summary describing its role, location, alleles, expression patterns, links to the Gene Ontology and to citations where the gene has been mentioned (there is a slight variation of these fields across databases).

The gene names, symbol and synonyms can be used in different ways to facilitate automatic recognition of gene mentions in the texts (see Section 2.3.1). The allele names can also be used for the same purpose.

Links to references in the literature allow the systems to place the genes back in their context in the text, and so use the context as a feature for recognising gene names and relations between genes/proteins.

The links to the Gene Ontology provide information about the cellular location, molecular function and biological processes of the gene products. This information can serve as training and evaluation resources for the automatic extraction of similar information from the text (for instance, evaluating a system for automatic prediction of the cellular location of a gene product).

FlyBase also includes links to the Sequence Ontology, with the intention of specifying the class of the gene.

2.2.2 Terminologies and ontologies

A biomedical terminology is a collection of names of entities (terms) employed in the biomedical domain, while a biomedical ontology is a collection of concepts representing the entities and focusing on the domain-related relations between the concepts. But in practice the two definitions get mixed up [Bodenreider, 2006]: terminologies usually disclose hierarchical (is-a) relations between terms, and ontologies include the various terms associated with the concepts.

[6]Model organisms are species that are extensively studied to understand particular biological phenomena, in the expectation that discoveries made in the organism model will provide insight into the workings of other organisms.

[7]http://www.informatics.jax.org/

[8]http://www.wormbase.org/

[9]http://www.yeastgenome.org/

2.2.2.1 Gene Ontology

The Gene Ontology (GO)[10] is in fact a set of three independent ontologies: one of cellular components containing 2 018 terms, a second of molecular functions containing 7 879 terms, and a third of biological processes containing 13 923 terms[11]. Each entry in these ontologies contains a definition of the term, synonyms if any, and is-a and/or part-of relations to other entries. Figure 2.1 shows simplified examples of portions of the three GOs.

```
%molecular_function ; GO:0003674, GO:0005554
  %antioxidant activity ; GO:0016209
   ...
  %auxiliary transport protein activity ; GO:0015457
   ...
  %binding ; GO:0005488
    ...
    %amine binding ; GO:0043176
      %2-aminoethylphosphonate binding ; GO:0033226
      %acetylcholine binding ; GO:0042166
        %acetylcholine receptor activity ; GO:0015464
        ...
      %amino acid binding ; GO:0016597
      ...
```

(a) Molecular functions

```
%biological_process ; GO:0008150, GO:0000004, GO:0007582
  ...
  %cellular process ; GO:0009987, GO:0008151, GO:0050875
    %absorption of light ; GO:0016037
    ...
    %cell communication ; GO:0007154
      ...
      %cell-cell signaling ; GO:0007267
        ...
        %transmission of nerve impulse ; GO:0019226
          %synaptic transmission ; GO:0007268
```

(b) Biological processes

```
%cellular_component ; GO:0005575, GO:0008372
  %cell ; GO:0005623
    <cell part ; GO:0044464
      %membrane ; GO:0016020
        ...
        %plasma membrane ; GO:0005886
        %postsynaptic membrane ; GO:0045211
        %presynaptic membrane ; GO:0042734
        ...
```

(c) Cellular components

Figure 2.1: Portions of the hierarchical view of Gene Ontology. '%' indicates an *is-a* relation; '<' indicates a *part-of* relation.

[10] http://www.geneontology.org/

[11] Term statistics dated from 7th October, 2007

The concepts expressed in these ontologies relate to the behaviour of gene products (instead of genes, as the ontology name might suggest). Gene products may be linked to one or more entries in these ontologies, and these links are called annotations, also available in the GO website. Most gene entries in the model organism databases have links to entries in each of the three GOs.

GO terms can serve to identify and classify expressions in the text, although the terms in the ontology usually do not map directly to terms in the text (e.g. GO entry: "activation of MAPK"; expression found in text: "MAP kinase activation" [Bodenreider, 2006]), so variations of these have to be considered to increase the number of mappings. The relations between the terms can be used to validate automatically extracted information against information contained in the GO annotations or model organism databases.

GO is less helpful, though, when handling molecular biology texts, since the information it carries starts at the gene product level.

2.2.2.2 Sequence Ontology

The Sequence Ontology (SO)[12] [Eilbeck and Lewis, 2004] is also part of the GO project but it is a completely independent ontology. While GO is a collection of terms used to describe gene products, SO is specialised in the molecular biology domain, describing the features and properties of biological sequences. The three basic kinds of relations between the terms in SO are is-a, part-of, and derived-from. For example, "transcript" is part-of "gene", a "processed transcript" is-a "transcript", and it derives-from a "primary transcript" that is also a transcript. Other kinds of relations are also present but are less frequent.

SO was created to provide a standardised set of terms and relationships with which to describe genomic annotations, but it can also be particularly useful for annotating scientific text in molecular biology, given SO's fine grainedness in relation to this subdomain and its precise relations, which can be mapped to relations between the entities in the text. A portion of SO can be seen in Figure 2.2 (SO is no longer provided in this format, but we have kept it here as an example because it shows the hierarchy of the concepts, while the current OBO – Open Biomedical Ontologies – format is flat).

2.2.2.3 MeSH

The Medical Subject Headings (MeSH) form a set of 16 hierarchies (trees) of terms, developed by the National Library of Medicine[13] to index, catalog and search for documents related to biomedicine and health in general. The scope of the terms is quite broad; hierarchies include root terms such as "Anatomy", "Diseases", "Chemicals and Drugs". The relation between the terms in any of the hierarchies can be understood as broader/narrower [Nelson *et al.*, 2001], in some cases corresponding to an is-a relation (e.g. "genes" - "pseudogenes"), in others it corresponds to a part-of relation (e.g. "genes" - "gene components"). A term can be found in more than one place in a hierarchy: for example, the term "glycomics" appears under "Biochemistry" and "Genetics" in the Natural Sciences hierarchy. Figure 2.3 shows a portion of MeSH's Biological Sciences hierarchy.

The only cross references between terms of independent branches of a hierarchy or between terms in distinct hierarchies are "see also" links to another term, but there is no specification of why or how the terms are related. MeSH's relations do not include any causal relation (e.g. "caused-by", "derived-from" or "product-of") between terms across the hierarchies. For example, the concepts of "gene" and "protein" are not related in MeSH (it is known that proteins are gene products); "gene" comes under "Genetic Structures" in the Biological Sciences hierarchy, while "protein" comes under "Amino Acids, Peptides, and Proteins" in the Chemicals

[12] http://www.sequenceontology.org/
[13] http://www.nlm.nih.gov/

```
 ...
@is_a@gene ; SO:0000704
  @part_of@non_transcribed_region ; SO:0000183
  @part_of@regulatory_region ; SO:0005836
    @is_a@attenuator ; SO:0000140
    @is_a@enhancer ; SO:0000165
    @is_a@insulator ; SO:0000627 ; synonym:insulator_element
    @is_a@locus_control_region ; SO:0000037
    @is_a@operator ; SO:0000057
    @is_a@polyA_signal_sequence ; SO:0000551
    @is_a@promoter ; SO:0000167
      ...
    @is_a@silencer ; SO:0000625
      ...
    @is_a@terminator ; SO:0000141
      ...
  @part_of@transcript ; SO:0000673
    ...
    @part_of@exon ; SO:0000147
      ...
    @is_a@processed_transcript ; SO:0000233
      ...
      @is_a@mRNA ; SO:0000234 ; synonym:messenger_RNA
        @part_of@CDS ; SO:0000316 ; synonym:coding_sequence
          @part_of@coding_end ; SO:0000327 ; synonym:translation_end
          @part_of@coding_start ; SO:0000323 ; synonym:translation_start
          @derived_from@polypeptide ; SO:0000104 @part_of@ protein ; SO:0000358
          ...
```

Figure 2.2: Portion of the hierarchical view of Sequence Ontology

and Drugs hierarchy. Another example is the term "Acanthamoeba Keratitis", found under "Eye diseases" in the Diseases hierarchy, which has no link to the term "Acanthamoeba", part of the Animals hierarchy and known cause of the disease.

2.2.2.4 UMLS

The Unified Medical Language System (UMLS)[14] is a set of three resources: a specialist lexicon, a metathesaurus and a semantic network.

The specialist lexicon is intended to be a general English lexicon that includes many biomedical terms. Each entry records the base form of a word (or multi-word term), its inflectional and possible spelling variants, its part of speech (words that function as more than one part of speech have one entry for each) and, for verbs, their subcategorisation patterns.

The metathesaurus is a collection of many existing terminologies/ontologies/thesauri that include biomedical information, such as those described in this section (e.g. MeSH, GO) and many more. Searching for a term in the metathesaurus results in a list of the definitions and synonyms for that term in each of the resources included in the metathesaurus, and the possibility of looking at other terms that are hierarchically related to that given in the several sources. The metathesaurus also provides a link to the concept in the semantic network to which the term is assigned.

The semantic network is divided into two independent hierarchies: one containing biomedical entities, and another biomedical events. There are several relations that link the concepts in a hierarchy and across both hierarchies. Such relations are, for example, "adjacent-to", "affects", "consists-of", "interacts-with", among others including the more common "is-a" and "part-of"

[14] http://www.nlm.nih.gov/research/umls/

```
Biological Sciences [G]
  Biological Sciences [G01]  +
  Health Occupations [G02]  +
  Environment and Public Health [G03]  +
  Biological Phenomena, Cell Phenomena, and Immunity [G04]  +
  Genetic Processes [G05]  +
  ...
  Genetic Structures [G14]
    Genome [G14.340]
      Genome Components [G14.340.024]
        Attachment Sites, Microbiological [G14.340.024.079]
        CpG Islands [G14.340.024.159]
        DNA Sequence, Unstable [G14.340.024.189]  +
        DNA, Intergenic [G14.340.024.220]  +
        Genes [G14.340.024.340]
          Alleles [G14.340.024.340.077]
          Gene Components [G14.340.024.340.137]  +
          Genes, Archaeal [G14.340.024.340.198]
          Genes, Bacterial [G14.340.024.340.201]
          Genes, cdc [G14.340.024.340.250]
          ...
        Insulator Elements [G14.340.024.420]
        Interspersed Repetitive Sequences [G14.340.024.425]  +
        Isochores [G14.340.024.430]
        Locus Control Region [G14.340.024.470]
        ...
```

Figure 2.3: Portion of MeSH hierarchy

relations. Figure 2.4 shows a portion of the entity hierarchy.

The relations represented in the hierarchy are "is-a" relations. If, for instance, we consider the concept "Gene or Genome", some examples of its relations across the Entity hierarchy are: "Gene or Genome" part-of "Cell", contains "Body Substance", produces "Amino Acid, Peptide, or Protein". Relations between concepts from the Entity hierarchy and those from the Events hierarchy are, for example, "Gene or Genome" affects "Physiologic Function", carries-out "Genetic Function", location-of "Molecular Function".

2.2.2.5 GENIA ontology

The GENIA ontology[15] is a small coarse ontology that contains concepts related to the biomedical domain in general. It was developed as the semantic classification used in the GENIA corpus. Figure 2.5 shows an example branch of the ontology.

In the GENIA corpus, a mention of a gene, for instance, is tagged as "domain or region of DNA", in the same way that sequences smaller or bigger than a gene would be tagged, making the distinction of gene parts impossible.

2.2.3 Corpora

The most popular source of biomedical text for natural language processing experiments are the abstracts provided by Medline[16]. Medline is a database of biomedical bibliographic information, and for each of its entries it provides the original abstract. Medline is indexed by MeSH terms and contain citations from 1950 to the present; currently it includes citations from 5 000 worldwide journals; in 2006 alone, 623 000 entries were added to Medline. Medline abstracts can be searched through PubMed.

[15] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/topics/Corpus/genia-ontology.html
[16] http://www.nlm.nih.gov/pubs/factsheets/medline.html

```
Entity
  Physical Object
    Organism
      ...
    Anatomical Structure
      Embryonic Structure
      Fully Formed Anatomical Structure
        Body Part, Organ, or Organ Component
        Tissue
        Cell
        Cell Component
        Gene or Genome
      ...
    Substance
      Body Substance
      Chemical
        Chemical Viewed Structurally
          Organic Chemical
            Nucleic Acid, Nucleoside, or Nucleotide
            Organophosphorus Compound
            Amino Acid, Peptide, or Protein
            Carbohydrate
            Lipid
          ...
        Chemical Viewed Functionally
          ...
```

Figure 2.4: Portion of UMLS Semantic Network

```
Substance
  Compound
    Organic
      Amino acid
        Protein
          Protein family or group
          Protein complex
          Individual protein molecule
          Subunit of protein complex
          Substructure of protein
          Domain or region of protein
        Peptide
        Amino acid monomer
      Nucleic acid
        DNA
          DNA family or group
          Individual DNA molecule
          Domain or region of DNA
        RNA
          RNA family or group
          Individual RNA molecule
          Domain or region of RNA
        Polynucleotide
        Nucleotide
```

Figure 2.5: Portion of GENIA Ontology

Unfortunately, most full-text articles are not freely available online due to copyright restric-

tions. However, in 2000 the Public Library of Science (PLoS)[17] was founded and it currently publishes eight open-access journals (such as PLoS Biology, PLoS Medicine, PLoS Genetics). The journal issues are available in XML format, which facilitates the use of the articles for NLP. PLoS articles can be searched through PubMed Central (PMC)[18]. PubMed Central is a recent initiative which digitally archives full-text articles from several journals that grant open access to the whole or part of its content (some journals impose a time limit after publication for articles to be freely available). PubMed Central is also supported by a new NIH (National Institutes of Health) policy from 2005[19], which aims to enhance public access to archived publications resulting from NIH-funded research.

Several projects have committed effort in annotating Medline abstracts with biomedical and/or linguistic information. Cohen *et al.* [2005] compare six corpora of biomedical abstracts that contain some kind of annotation; the authors compared them in terms of their design features, and related these features to the use rate of the corpora by researchers other than those who developed them. The corpora considered are: GENIA corpus [Collier *et al.*, 1999], Medstract corpus [Pustejovsky *et al.*, 2002], GENETAG corpus [Tanabe *et al.*, 2005], a corpus developed by Craven & Kumlein [1999] (referred by Cohen *et al.* as Wisconsin corpus), a corpus developed by Blaschke *et al.* [1999] (referred by Cohen *et al.* as PDG corpus), and a corpus developed by Franzen *et al.* [2002] (referred as Yapex corpus).

GENIA, Medstract, GENETAG and Yapex corpora have all biomedical entities (named and unnamed) annotated: GENIA classifies entities according to the GENIA Ontology, Medstract according to UMLS Semantic Network, while GENETAG and Yapex have only a single class that includes both genes and proteins. Wisconsin and PDG corpora, on the other hand, have only annotated the entities that take part in specific relations, and are the only corpora where domain relations are annotated: Wisconsin has protein-protein interactions, gene-disease associations and protein-cellular location associations annotated, where the entities taking part in the relation are classified as appropriate (protein, gene, disease or location); PDG has only protein-protein interactions annotated.

GENIA is the only corpus among these that has structural annotation, such as sentence boundary, tokenization and PoS tags. The Wisconsin corpus also contains the same information, but it has been automatically generated and not manually checked.

Medstract is the only corpus among these that contains annotation of anaphoric relations between entities (see Section 2.3.3).

GENIA, Yapex and Medstract are composed of abstracts, each having respectively 2 000, 200, and 46 abstracts. GENETAG, Wisconsin and PDG are composed of sentences instead of abstracts; GENETAG is composed by 20 000 sentences; Wisconsin has in total 67 201 sentences, a part consisting of positive samples of relations (5 457 for protein-protein interaction, 829 for gene-disease associations, and 769 for subcellular localisation) and the rest consisting of negative samples (42 015, 11 771, 6 360, respectively); and PDG is the smallest of all, having 283 "blocks" with one or a few more sentences that give evidence of a protein interaction.

GENIA, Yapex, Medstract and GENETAG are encoded in relatively standard formats: GENIA, Yapex and Medstract are distributed in XML, and GENETAG is distributed in the known token/TAG (e.g. smg/NEWGENE) format. On the other hand, Wisconsin and PDG are distributed in unique formats, where annotation is detached from the text and not easily mapped back. PDG has been refactored by Johnson *et al.* [2007] and encoded in XML; the new version is named PICorpus.

Cohen *et al.* show that the usage rate for these corpora varies considerably; GENIA is by

[17] http://www.plos.org/journals/index.html
[18] http://www.ncbi.nlm.nih.gov/sites/entrez?db=PMC
[19] http://grants.nih.gov/grants/guide/notice-files/NOT-OD-05-022.html

far the most widely used corpus, followed by GENETAG, Yapex, Medstract, Wisconsin and PDG. They conclude that what mostly favours the use of an annotated corpus by the research community is the format in which it is distributed—standard formats are preferred—and the presence of structural annotation, such as sentence boundaries, tokenization and PoS tags.

So far there exists no annotated corpus of full-text articles (rather than abstracts) with the kind of information annotated in the corpora mentioned above. That limits the scope of the research that can be undertaken since the text in abstracts represents different features from text in the articles' main body or even in figure or table captions.

2.3 Tasks

A number of subtasks can incrementally build up the structure of the texts in order to make information extraction more feasible and more precise. For instance, the PASTA system for extraction of information about the role of amino acid residues in proteins [Gaizauskas *et al.*, 2003] includes a module for terminology processing (identifying and classifying the NPs referring to entities of interest), a module for syntactic and semantic processing (where sentences are converted into semantic representations), a module for discourse processing (which identifies the instances from the semantic representation that refer to the same entity) and finally templates are created to organise the information gathered about the entities.

The following sections describe some of these subtasks that have been tackled so far by researchers.

2.3.1 Named-entity recognition (NER)

Named entities are those referred to in the text by a proper name rather than a common noun. Proper names can not be found in an ordinary lexicon and so need to be recognised as such in the text so that their grammatical and semantic role can be recovered. In biomedical texts the named entities of interest may be genes, proteins, drugs, chemical compounds, diseases, etc. Unlike in newswire text, where proper names usually refer to individual/unique entities (e.g. USA, Gordon Brown), in biomedical texts they refer to classes of entities, for example, a gene name refers to all instances of such gene in all DNA sequences of all organisms that contain that gene. Despite this conceptual difference, these names are usually treated in the same way as proper names; gene-name recognizers work on the same principles as general named-entity recognizers (which usually look for person, organisation and location names). The output of a named-entity recognition system usually consists of tags assigned to the words that are recognised as named entities, in the same way as PoS tagging.

Most of the work in biomedical NER has focused on recognising gene and protein names; recently two editions of the BioCreative evaluation workshops have paid attention to this task [Blaschke *et al.*, 2004, Krallinger and Hirschman, 2007]. These names, as described in Section 2.1, are usually ambiguous, which poses a challenge to classifying them as protein or gene names. The following approaches have been adopted to tackle biomedical NER: dictionary-based, rule-based, and machine learning/statistical approaches.

Dictionary-based approaches rely on a compiled list of gene/protein names that is used to find perfect or similar matches in the text. This list is derived from databases that record these names, as did for example Hanisch *et al.* [2003] and Krauthammer *et al.* [2000]. The main problem of dictionary-based approaches is their low precision, caused by the overlap between some gene names and common English words. They also become outdated quite quickly given that new gene names are constantly being created; this affects the recall of such systems.

Rule-based approaches rely on manually or automatically generated rules that indicate whether a word is or is not a gene/protein name. These approaches can consider beyond the morphological level and take into account the context of the word as well. One of the most successful rule-based systems [Cohen and Hersh, 2005] for gene and protein name recog-

nition is AbGene [Tanabe and Wilbur, 2002]. It has two phases: firstly an extended version of the Brill PoS tagger, where new tags for gene and protein names are added and hand-tagged sentences from biomedical text are used for training, was used to tag gene/protein names; and secondly post-processing rules were manually generated to help eliminate false positives and false negatives. The main disadvantage of rule-based approaches is the cost of hand-crafting rules and the difficulty of adapting them to other sub-domains, with different naming conventions [Park and Kim, 2006].

Several machine learning approaches make use of Hidden Markov Models (HMMs) as their base statistical framework, and differ on the set of features used. The main problem of machine learning approaches is building a big enough and representative training corpus. To overcome this problem, Morgan *et al.* [2003] proposed a strategy to generate a large amount of noisy training data automatically. Their strategy consists of using a dictionary-based system that makes use of gene names and bibliographic references from the FlyBase database: for each publication about the fruit fly, FlyBase records the genes that are mentioned in it; the authors collected the Medline abstracts for a set of these publications and tagged the gene names associated to them in the abstracts. With the generated corpus of abstracts, they have trained a HMM. Vlachos *et al.* [2006] have improved the Morgan *et al.* strategy by using an enlarged dataset and different software.

2.3.2 Semantic tagging

Besides identifying the names of biomedical entities in the text, it is also important to identify common nouns (rather than proper names) that refer to biomedical entities. It is also desirable to classify them according to their role in the domain of the text.

Having the semantic information about the words is relevant to further tasks that try to find relations between expressions in the text; for example, to find the relation between a gene and a disease, it is first necessary to know that a NP refers to a gene and another to a disease.

As the vocabulary used to refer to biomedical entities in general (common nouns such as "gene", "RNA" and "enzyme", instead of proper names) remains practically unchanged (in contrast with proper names), using a dictionary-based approach is usually a good enough strategy. However, the ambiguity problem is still present, with some words referring to more than one type of entity.

For instance, Castaño *et al.* [2002] make use of the UMLS Semantic Network concepts to type the entities found in the text (e.g. "protein", "cell", "organism"). Bodenreider [2006] shows examples of how GO can be used for the same purpose. In the GENIA corpus, all NPs referring to biomedical entities are tagged according to the GENIA ontology (e.g. "protein", "protein complex", "domain or region of DNA"). The PASTA system uses its own set of semantic classes (e.g. "protein", "non-protein compound", "species") to classify the terms in the text (the terms are identified by morphological analysis or by consulting a lexicon they have built from online resources).

2.3.3 Anaphora resolution

After identifying all NPs referring to biomedical entities in the text, it is important to know which NPs refer or are related to the same entity. Anaphora resolution is the process of linking these NPs. Anaphora is the linguistic phenomenon where an expression further in the text refers back to a previously-mentioned expression. For example, in the following passage, there are anaphoric relations between the highlighted mentions: the anaphoric relations between (a) and (c) and between (b) and (d) are coreferential, because both mentions refer to the same entity; the relation between (b) and (c) and between (c) and (d) are associative, because they are related but do not refer to the same entity.

(1) ``... is composed of **five proteins**(a) encoded by **the male-specific lethal genes**(b) ... **The MSL proteins**(c) colocalize to hundreds of sites ... male animals die when they are mutant for any one of **the five msl genes**(d).''

Resolving anaphora is essential for information extraction, that is, in order to recover all the information about an entity in the text, it is necessary to take into account even the sentences where the entity is not explicitly mentioned by its name. For the extraction of domain relations between biomedical entities, e.g. interaction between proteins, anaphora resolution can be crucial, as in the following example, where linking (b) to (a) is necessary to recover the relation between CED-3 and CED-4:

(2) ``**The CED-3 protein**(a) is one of a continuously growing family of caspases ... **this protein**(b) is activated by **CED-4** ...''

It is important to have semantic information about the entities in order to verify whether two expressions are anaphorically related; for example, if two NPs are tagged as genes, it is more likely that they are anaphorically related than if they had different tags. That means it is very important to have as input to an anaphora resolver the output of NER and semantic tagging systems. The lack of appropriate sources of semantic information in other domains limits the anaphora resolution techniques that can be adopted.

The large majority of entities in biomedical texts are referred to using non-pronominal noun phrases, like proper nouns, acronyms or definite descriptions. Hence focusing on these noun phrases should contribute more to the resolution process.

Very few systems for anaphora resolution have been developed for the biomedical domain. Castaño *et al.* [2002] developed a salience-based system for anaphora resolution that uses semantic information derived from the UMLS Semantic Network. They have developed the Medstract corpus (mentioned in Section 2.2.3) to evaluate their system. Gaizauskas *et al.* [2003] developed the PASTA system, which is an information extraction system that aims to extract relations between proteins. With that in mind, they implement an inference-based coreference resolution module which reasons on semantic representations of sentences: entities that have semantic predicates is common are considered coreferent. The authors also use the same mechanism to link representations of hypothetical entities that are part of an information extraction template to entities seen in the text. Yang *et al.* [2004] developed a supervised machine-learning approach for anaphora resolution and evaluated it on a portion of the GENIA corpus, which is tagged with semantic information based on the GENIA Ontology. They focus only on coreferent cases and do not attempt to resolve associative links.

Section 3.3 in the next chapter describes these systems in more detail. They have been developed based on abstracts of biomedical articles, which represent a very restricted use of anaphora. We believe full-text articles present a more realistic view of anaphora in biomedical texts, mainly when information extraction is considered the target application.

2.3.4 Relation extraction

It is important for Biology research to identify the relations between entities involved in biological processes. Such relations could, for instance, include the interaction between proteins, the association between a gene and a disease, or between a disease and drugs. The automatic extraction of relationships from text focuses usually on a prespecified kind of relationship. The most explored relation between biomedical entities has been protein-protein interaction, which had a task dedicated to it in the last BioCreative evaluation workshop[20].

There have so far been several approaches adopted for relation extraction. The simplest

[20] http://biocreative.sourceforge.net/biocreative_2_ppi.html

technique consists of looking for entities that occur together in a specific scope of text (e.g. sentence, paragraph, the whole abstract) with considerable frequency. Stapley and Benoit [2000] predicted the relation between two genes by checking how often they co-occur in the same Medline abstract. Ding *et al.* [2002] later tested the same approach considering sentence and paragraph as scope of co-occurrence, and compared it to considering the whole abstract.

Another approach consists of using template-like patterns (usually in the form of regular expressions) that should match the relationships in the text. An example of such a system is that presented in [Blaschke *et al.*, 1999], in which they use manually built patterns based on a set of verbs that denote the relations of interest (e.g. `protein <P1> <verb> protein <P2>`) in order to extract the relations. This type of patterns can also be learned automatically from a dataset where relations are annotated by considering the context of the entities taking part in the relations. Huang *et al.* [2004] have adopted a dynamic programming algorithm to compute patterns by aligning relevant sentences and key verbs that describe protein interactions.

In order to have a more flexible framework than pattern-matching, some works adopted syntactic parsers to recover relations between whole noun phrases. Park *et al.* [2001] used a parser based on a combinatory categorial grammar in order to extract relations between proteins; their system looks for the syntactic arguments of a set of verbs of interest, being able to recover even NPs that take part in coordination and apposition clauses. Fundel *et al.* [2007] have developed RelEx, a system for relation extraction that relies on dependency parse trees. RelEx creates candidate relations by extracting paths connecting pairs of mentions of proteins from dependency parse trees; these should also contain any of a list of relevant terms describing the relation. The relations are filtered using a small set of rules, and also the occurrence of negation, coordination and passive voice in the trees is treated accordingly.

Elaborate machine-learning techniques have also been adopted for relation extraction tasks. Bunescu and Mooney [2005] have applied kernel methods to the extraction of relations between proteins. They have used as training data the AIMed corpus, which contains 225 Medline abstracts where around 1000 protein-protein interactions have been annotated. They have used the words surrounding the protein mentions as features for the kernel model. Bundschus *et al.* [2008] have developed a probabilistic system for extracting relations between genes and diseases and between diseases and treatments using Conditional Random Fields, which treat the task as one of sequence labelling. For the extraction of disease-treatment relations they have used as training data 2001 Medline abstracts where these relations were annotated and classified as *cure, only disease, only treatment, prevents, side effect, vague, does not cure*. For extracting gene-disease relations, they have used as training data GeneRIF phrases associated with gene entries in a database in fields describing diseases caused by abnormal behavior of the gene.

The coverage of relation extraction systems is affected by the presence of anaphoric expressions in the text. Fundel *et al.* have perfomed an analysis of errors made by their RelEx system; they report that 12% of false negative errors are due to anaphora, that is, where one of the entities involved in the relation is referred to by an anaphoric expression (e.g. "this protein"), which was not inicially tagged as a valid mention of a protein.

2.4 Summary

In this chapter we have described what differentiates biomedical scientific articles from other genres of text and have presented the lexical and semantic resources available for the biomedical domain, which can be exploited by natural language processing tools. We have also described the tasks that are necessary to be performed on biomedical texts in order to be able to extract information from them. Each of these tasks incrementally builds up a layer of understanding of the information present in the text. The task of anaphora resolution takes advantage of the information accumulated from named-entity recognition and semantic tagging, and can contribute, for example, to the extraction of relations between entities. The next chapter describes

what anaphora resolution consists of and presents the approaches taken so far to accomplish it.

Chapter 3

Anaphora and anaphora resolution

3.1 Anaphora

Anaphora is the relation between two linguistic expressions in the discourse where the reader is referred back to the first when reading the second later in the text. According to Hirst [1981], anaphora is the linguistic device of making an abbreviated reference to some entity in the discourse in the expectation that the reader will be able to disabbreviate the reference and determine the identity of the entity. By abbreviate, Hirst means containing fewer bits of disambiguation information rather than lexically shorter. The following example of anaphora was extracted from a biomedical text:

(3) ``... is the use of **non-coding RNAs** transcribed from genes located on the X chromosome itself. **These RNAs** ...''

In this example, "these RNAs" is an abbreviated reference to "non-coding RNAs".

The referring expression is usually called the anaphor, while the expression it refers to is called its antecedent.

The reference process can be caused by several distinct relations between the entities represented by the textual expressions involved. When both expressions represent the same entity, the relation between them is called coreference.

The concepts of anaphora and coreference have been used in different ways in the literature, causing some confusion in the field. van Deemter and Kibble [2000] have sought to distinguish the two concepts. They define coreference as the relation holding between linguistic expressions that refer to the same extralinguistic entity. On the other hand, they define anaphora as a relation where interpretation of a referring expression is dependent on a previous expression (antecedent) within the same discourse. Thus an anaphoric relation may or may not be coreferent: an expression may be anaphoric in the strict sense that its interpretation relies on the preceding expression, although the expressions involved may refer to distinct entities. If an anaphoric relation is not coreferent, it is usually called bridging or associative. On the other hand, a relation might be just coreferent, in the sense that the entity has been mentioned earlier. Figure 3.1 represents the intersection between the concepts.

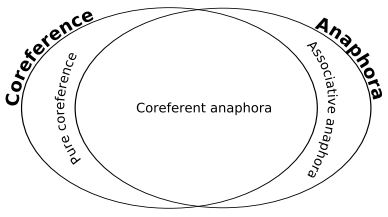


Figure 3.1: Coreference vs. anaphora

The confusion between coreference and anaphora arises mainly in cases that do not present the abbreviation mentioned by Hirst, where the referring expression is a repetition of a previous expression. In Example 4, the relation between the highlighted expressions is controversial: it can be seen as merely coreferent, since both expressions carry the same information, but on the other hand one can argue that the second mention would seem out of place if it were not for the presence of the previous one, revealing a dependency between the expressions.

(4) ``**Initiator caspases** are thought to be at the beginning of a proteolytic cascade that amplifies the cell death signal and results in the activation of the effector caspases. **Initiator caspases** usually have long pro-domains, while effector caspases have short pro-domains.''

It is clear when an anaphoric relation is not coreferent, since these are the cases where the expressions have different referents, as in Example 5.

(5) ``The expression of **reaper** has been shown to be regulated by distinct stimuli (...). Recently, a Drosophila p53 ortholog was identified by searching the genome database, and it was shown to bind a specific region of **the reaper promoter**''

Example 5 presents an associative anaphora case, where the referents of the expressions hold a semantic relation to each other. Associative anaphora is the phenomenon in which a referring expression is used to refer to an entity not previously mentioned in the text, but the existence of which can be inferred by virtue of some previously mentioned entity [Hawkins, 1978, Meyer and Dale, 2002].

Coreference is a symmetrical and transitive relation, while anaphora is not. Anaphora is dependent of context, coreference is not.

Coreference resolution can be understood as the process of linking all textual references to the same entity, forming coreference chains. Anaphora resolution, on the other hand, consists in linking an anaphoric expression to its antecedent, the previous textual entity that the anaphor is anchored to, forming anaphor-antecedent pairs.

Anaphors typically refer back to other constituents in the same sentence, or to constituents in earlier utterances in the discourse. Syntactic information plays a central role in establishing appropriate referents for the former case, intrasentential anaphora, while semantic and pragmatic information are crucial in the latter case, intersentential anaphora [Carbonell and Brown, 1988].

Different kinds of noun phrases can present anaphoric behaviour: pronouns, definite descriptions, proper names, demonstrative NPs, among others. Pronouns are the most reduced form of anaphoric expressions[1]; they are almost always anaphoric and coreferent. The scope within which the antecedent of a pronoun may be found is known to be smaller than for non-pronominal (lexical) NPs. Usually it can be found in the same sentence as the pronoun or one or two sentences earlier. Hobbs [1978] reports statistics from a corpus of three texts (from different genres) containing a thousand pronouns, where 90% of the antecedents are found in the same sentence of the pronoun or in the previous sentence when the pronoun occurs before the verb (and 98% of the antecedents are found in the current sentence or the previous sentence). This limits the number of antecedent candidates when trying to resolve the pronoun's anaphoric relation.

Demonstrative NPs (NPs that start with a demonstrative pronoun such as "this", "these") are also known to be anaphoric most of the time and have a small scope of search for their antecedents, but larger than for pronouns.

On the other hand, definite descriptions (understood as all NPs introduced by the definite article "the") behave differently. Many of them are not anaphoric (50% for newspaper texts

[1]In fact, zero anaphors are the most reduced form of anaphora, but they do not form a NP; they are gaps in a phrase or clause that have anaphoric function.

according to [Vieira, 1998]), and when they are, they are often used to recall an entity that has been mentioned some sentences earlier. This means the methods for definite description resolution have to be able to identify which are anaphoric, and for the ones that are, choose the candidate from a broader scope.

Proper names are the NPs which allow for the longest distances between the anaphor and antecedent, since there is no ambiguity when an entity is referred to by its name, even when such entity was mentioned paragraphs earlier.

Indefinite NPs (NPs beginning with the indefinite article "a") usually introduce new entities in the discourse and are rarely anaphoric.

There are several theoretical linguistic studies that aim to establish a theory of the use of one anaphoric expression rather than others in specific cases; Huang [2000] describes three models, the topic continuity, the discourse hierarchy, and the cognitive model, and proposes a pragmatic model to describe anaphoric distribution in discourse, that is, the choice of a particular referential/anaphoric form at a particular point in discourse.

The main premise of the topic continuity model (Givón in [Huang, 2000]), also called distance-interference model, is that anaphoric encoding in discourse is essentially determined by topic continuity, measured primarily by factors such as linear distance (the number of clauses/sentences between the two mentions of a referent), referential interference (the number of interfering referents), and thematic information (maintenance or change of the protagonist). Roughly, what the model predicts is this: the shorter the linear distance, the fewer the competing referents, and the more stable the thematic status of the protagonist, the more continuous a topic; the more continuous a topic, the more likely that it will be encoded in terms of a reduced anaphoric expression.

In the hierarchy model (Fox in [Huang, 2000]), it is assumed that the most important factor that influences anaphoric selection is the hierarchical structure of discourse; mentions at the beginning or peak of a new discourse structural unit (e.g. paragraph, turn, episode) tend to be made by a full NP, whereas subsequent mentions within the same discourse structural unit tend to be achieved by a reduced anaphoric expression.

The basic idea underlying the cognitive model (Tomlin, Gundel in [Huang, 2000]) is that anaphoric encoding in discourse is largely determined by cognitive processes such as activation and attention—activation of a referent in one's current short-term memory at moment t_n is a result of focusing one's attention on that referent at a previous moment t_{n-1}. With that in mind, the central empirical claim of the cognitive model is that full NPs are predicted to be used when the targeted referent is currently not activated, whereas reduced anaphoric expressions such as pronouns are predicted to be selected when such a referent is currently activated.

The basic idea of the pragmatic model is that anaphoric distribution can be predicted in terms of the systematic interaction of some general pragmatic strategies such as Levinson's Q-, I-, and M-principles, which are: Q-principle: do not say less than is required (bearing I in mind); the I-principle: do not say more than is required (bearing Q in mind); and the M-principle: do not use a marked (lexical) expression without reason. Huang suggests that such principles underlie anaphoric distribution in the following ways: (1) establishment of reference tends to be achieved through the use of an elaborated form, notably a lexical NP; (2) shift of reference tends to be achieved through the use of an elaborated form, notably a lexical NP; and (3) maintenance of reference tends to be achieved through the use of an attenuated form, notably a pronoun.

Computational models for anaphora resolution are inspired by theoretical linguistic models such as those mentioned above. However, natural language processing tools still perform poorly in automatic recovery of cognitive and pragmatic clues from the discourse, making models as the cognitive and pragmatic more difficult to account for in computational grounds than the topic continuity and discourse hierarchy models.

In the following section we discuss several issues related to the automatic resolution of anaphora and describe the systems that have been proposed so far.

3.2 Anaphora resolution

Anaphora resolution has been considered one of the most challenging problems in NLP. There has been prevailing consensus that the difficulty of the problem lies in its dependence on sophisticated semantic and world knowledge.

Anaphora resolution systems usually aim to resolve only anaphors which have noun phrases as their antecedents because resolving anaphors which have verb phrases, clauses, sentences or even paragraphs/discourse segments as antecedents, is a more complicated task [Mitkov, 1999].

Most anaphora resolution systems deal only with coreferencial cases, only a few systems aim also to resolve associative anaphora cases, which are considered more challenging and more dependent on semantic information.

Many sources of information play a role in determining the antecedent of an anaphoric expression. For instance, the distance between an anaphoric expression and the antecedent candidate, lexical information such as head-noun matches can be an indicator of coreference. Lexical constraints such as gender and number agreement can help eliminate some antecedent candidates; syntactic patterns can help determine whether an expression is indeed anaphoric; syntactic roles can indicate preference for particular antecedent candidates and semantic relations can describe the nature of the anaphoric relation, and so on. However, no single source of knowledge is a completely reliable factor. For example, matching head nouns can be modified by different modifiers that make the coreferent relation unlikely (as in e.g. "`ced-2 gene`" and "`egl-1 gene`"), while expressions that disagree in number can still be coreferent if one has a collective meaning, e.g. "`MSL family ... the MSLs`". Furthermore, the knowledge sources are combined differently depending on the type of NP to be resolved. For example, pronoun resolution can never count on head-noun matching but can limit the search for antecedents to a distance of few previous sentences, while definite descriptions resolution can rely on string matching but have to consider other factors to be able to select an antecedent among the NPs from a broader set of sentences.

Below we present the description of a generic anaphora resolution system, similar to that proposed by Ng [2003]:

Step 1: Identification and selection of noun phrases to be resolved: the NP selection can be based on linguistic information, for example the type of NPs, or based on domain information, when a system aims to resolve only NPs that are related to a specific domain.

Step 2: Extraction of features that describe the selected noun phrases: features may be lexical, syntactic, semantic, among others. Developers can opt for sophisticated features that require complex NLP tools to be extracted (which might not always be available or robust enough), or more superficial features, acquired through shallow processing.

Step 3: (optional) Determining if the noun phrase is new in the discourse, that is, has no antecedent: a system can include a module for determining whether a NP is anaphoric, before trying to find an antecedent for it. Such modules can be useful when the anaphora resolution model adopted by the system returns an antecedent in all cases.

Step 4: Creation of the set of antecedent candidates: systems consider as possible antecedents only the NPs that occur before the anaphor in the text. Some systems consider them all, while others impose a maximum number of previous sentences to be considered.

Step 5: (optional) Filtering of unreasonable candidates: some systems exclude candidates that do not conform to some basic constraints, for example number agreement (when aiming to resolve coreference).

Step 6: Scoring/ranking or searching candidates: this is the core part of an anaphora resolution system. It is the module that interprets the features extracted in Step 2 and determines whether two NPs are anaphorically related based on them. This module can be built, for example, by a set of hand-made heuristics, or a machine-learning algorithm. Most resolution models rank all antecedents according to a computed score or a set of rules (and return the first candidate as antecedent), while other systems search in a particular order for a candidate that conforms to a set of constraints (returning the first to succeed as antecedent).

Steps 2 to 6 are performed once for each NP selected in Step 1.

Keeping in mind the steps above, anaphora resolution systems can be compared according to their approaches to each step. Concerning Step 1, the selection of noun phrases to be resolved, some systems focus on one particular type of anaphoric expression, while others aim to cover several types. Most of the work done on anaphora resolution deals only with pronouns; well-known works for pronoun resolution in English are [Lappin and Leass, 1994, Kennedy and Boguraev, 1996, Mitkov, 1998, Ge *et al.*, 1998]. Definite descriptions were approached, for instance, in [Bean and Riloff, 1999, Vieira and Poesio, 2000]. [Strube *et al.*, 2002, Ng and Cardie, 2002c] address a broader range of NPs: pronouns, definite and demonstrative NPs and proper names. Given that pronoun resolution and non-pronominal anaphora resolution present different challenges, most systems focus on one or the other. The set of features used by a pronoun resolution system usually differs from the set of features used to resolve non-pronominal anaphora. Strube *et al.*, for instance, shows how a measure of string matching can improve the performance of a system on non-pronominal anaphora resolution, while it makes no difference for pronoun resolution. A system can also select the NPs to be resolved based on semantic information, instead of by type of NP. For instance, McCarthy and Lehnert [1995] select only NPs that refer to people, companies, governments and other entities involved in joint capital ventures, since this was the domain of the texts they were processing.

Concerning Step 2, related to the features used to describe NPs to be resolved, we can distinguish between systems which make use of discourse, semantic and deep syntactic knowledge, called knowledge-rich approaches, and systems which avoid the use of sophisticated knowledge and instead rely only on lexical and possibly shallow syntactic information, called knowledge-lean approaches. NLP tools for acquiring sophisticated linguistic knowledge, including semantic, have not been able to reach as high accuracy as tools for performing well-defined tasks, such as part-of-speech tagging. Accordingly, systems which rely on less sophisticated tools to derive their features from are considered to have broader coverage, but less precision, than systems which rely on complex (sometimes manually coded/corrected) features. For instance, the Lappin and Leass system for pronoun resolution [Lappin and Leass, 1994] is acclaimed for not relying on semantic or pragmatic constraints but, on the other hand, is criticised for relying on full parsing, which is also considered an expensive resource; Kennedy and Boguraev [Kennedy and Boguraev, 1996] modify the Lappin and Leass system by approximating the output of full parsing through a set of cheaper heuristics.

Step 3 is an optional part of an anaphora resolution system. Some systems opt for a module to decide whether a NP is anaphoric or not, before looking for antecedents for it; while other systems opt for going straight to looking for antecedents, and consider not anaphoric those NPs for which no antecedent was found. For NPs like definite descriptions, which according to Vieira and Poesio [2000] are not anaphoric 50% of the times they appear in newspaper texts, adopting

this step can considerably affect the system's overall performance. Lappin and Leass have implemented a module to detect pleonastic pronouns, more precisely the non-anaphoric "it", based on lexical and syntactic information. Vieira and Poesio [2000], Bean and Riloff [1999], and Uryupina [2003] have proposed strategies to detect discourse-new definite descriptions. Vieira and Poesio's discourse-new heuristics were concerned with appositive constructions, copular constructions and postmodification, among other clues. Bean and Riloff used basically the same heuristics as Vieira and Poesio, but additionally they verified whether the definite description was in the first sentence of the text and also whether it was a "definite only", i.e. its head always happens with the definite article in the text. Uryupina distinguishes discourse new and unique (e.g. "the USA") definite NPs; she trains two rule-learning classifiers, one with discourse-new vs. discourse-old instances, and another with unique vs. non-unique instances. Both classifiers are trained with the same syntactic features used by Vieira and Poesio, plus a measure of "definite probability" derived from internet counts (how many times the NP appears with the definite article, with the indefinite article ("a"), and independent of determiner); the author combines the output of both classifiers and finds that uniqueness information is relevant to determining anaphoricity. Ng and Cardie [Ng and Cardie, 2002b] distinguish anaphoric and non-anaphoric cases among all kinds of NPs by using a set of 37 features (lexical, grammatical, semantic and positional) for training a decision-tree and a rule-learning classifier.

Concerning Step 4, selection of antecedent candidates in most systems simply involves the construction of a set of noun phrases preceding the anaphor under consideration in the associated document, although some systems impose a maximum distance (usually in number of sentences) from the anaphoric expression within which to look for the antecedent, in order to reduce the computational overload and to avoid noise. Distance from the anaphor is a feature that plays a role in all anaphora resolution systems—it is understood that the further away a candidate is, the less likely that it is the correct antecedent, unless the distance is compensated by other factors. For instance, Mitkov's algorithm limits the search for pronoun antecedents to the two sentences preceding the pronoun; Vieira [1998] experiments with a maximum distance of 1, 4 and 8 sentences, verifying that precision drops and recall increases with distance. Her algorithm, however, allows some "special NPs" to ignore the distance limit; for example, NPs with same head noun as the anaphor. Lappin and Leass, instead of imposing a distance limit, impose a penalty weight according to distance, which in summary causes candidates at more than two sentences away to have their weight already below a threshold and consequently to be ignored. Ge [2000] considers distance through the Hobbs' algorithm [Hobbs, 1978], selecting at most 25 candidates which are ordered according to Hobbs' syntactic constraints.

Step 5, filtering "unreasonable" candidates, is another optional part of an anaphora resolution system. Some systems, in order to reduce the set of candidates, eliminate some based on simple heuristics that should point out unacceptable cases. For example, Strube *et al.* coreference resolution system discards candidates when: they are embedded in the same clause as the anaphor; they are not of the same semantic class as the anaphor; they do not agree in gender and number with the anaphor (only in case this is a pronoun). The system also ignores all antecedent candidates for anaphors that are indefinite NPs.

Step 6 is the core part of an anaphora resolution system, that is, the resolution model, which integrates the information built up in the previous steps, processes them, and returns the antecedents for the anaphors. We distinguish two basic types of resolution models, knowledge-based and corpus-based. In knowledge-based approaches [Lappin and Leass, 1994, Mitkov, 1998, Vieira *et al.*, 2002] the resolution procedure is based on a set of hand-crafted rules that specify whether two discourse entities are anaphorically related; some knowledge-based systems try to approximate theoretical discourse models to account for anaphora behaviour. In corpus-based approaches [Ge *et al.*, 1998, Strube *et al.*, 2002, Ng and Cardie, 2002c], on the other hand, the

knowledge is automatically obtained from corpora annotated with anaphora information, which have become available more recently. The main advantage of corpus-based approaches is that complex and unpredicted situations that indicate anaphora can still be captured, while knowledge-based approaches are more conservative, the developer being responsible for creating rules to account for predicted cases. An important aspect to be considered at this step is the types of anaphora to be resolved—coreferent and/or associative. Most systems developed so far focus on resolving only coreferent cases ([Strube *et al.*, 2002, Ng and Cardie, 2002c], and all pronoun resolution systems). Among the few systems that try to solve associative anaphora are those of Vieira and Poesio [2000], Meyer and Dale [2002], Poesio *et al.* [2002], Bunescu [2003]. Resolving associative anaphora is considered a more difficult task than resolving coreference, since the NPs involved in the associative relation do not refer to the same entity and require the system to be able to infer a semantic relation between them as a clue to supporting the anaphoric relation.

In the next subsections we describe extant systems for anaphora resolution in more detail, distinguishing between knowledge-based and corpus-based systems.

3.2.1 Knowledge-based systems

Knowledge-based approaches to anaphora resolution may be divided in four groups [Ng, 2003, Hoste, 2005]: discourse-oriented approaches, in which discourse structure is taken into account, as in that proposed by Grosz *et al.* [1995]; factor-based approaches, such as that of Lappin and Leass [1994]; syntax-based approaches such as Hobbs' [1978]; and heuristic-based approaches, such as that adopted by Vieira and Poesio [2000].

3.2.1.1 Discourse-oriented approaches

Discourse models, especially centering [Grosz *et al.*, 1995] and focusing theory [Grosz, 1978, Sidner, 1979] have been successfully used for anaphora resolution. Both theories assume that certain entities in the discourse are more central or in focus than others and this imposes certain constraints on the referential relations that occur in the text.

Centering is a theory for interpreting pronouns in a discourse. It models the local coherence of a discourse and is composed of a set of constraints governing center movement (the conditions under which the center of a discourse should move from one discourse entity to another) and center realisation (the conditions under which a discourse entity can be referred to by a pronoun). Such constraints consider morphosyntactic, binding and semantic criteria. The works by Tetreault [2001] and Strube and Hahn [1999] are examples of systems using the centering framework. Tetreault presents variations of a centering-based pronoun resolution algorithm; the best performing one reaches 80.4% accurary on newspaper texts and 81.1% accuracy on fictional texts.

Sidner's focusing framework keeps a set of data structures, including the current focus, a list of alternative candidate foci, and a focus stack to represent the current state of a discourse. For each sentence, the focusing algorithm uses a set of rules to determine whether there is a shift in focus and updates the data structures accordingly. For each anaphor encountered, another set of rules is used to rank candidate antecedents based on the focus-tracking data structures. The work of Rich and LuperFoy [1988] combines the principles of Discourse Representation Theory, centering, and focusing in different modules. Each module proposes candidate antecedents and evaluates other modules' proposals.

The main limitation associated with focus-based approaches is their complex and restricted nature. A discourse model is dependent on the genre of text (discourse) it represents, and modelling unrestricted text is a highly complex task. Grosz, for example, validated her focus work only on restricted task-oriented dialogs, where the structure of the sentences was very limited. Besides, when considering long discourses (such as full scientific articles) the antecedent for an anaphoric expression might be a long way back in the text, which would compromise the

structures available to track the focus of the discourse.

Gaizauskas and Humphreys [2000] propose a coreference resolution module as part of the LaSIE information extraction system. This module builds a discourse model based on predicate-argument representations of the elements in the sentences. For every sentence in the text, a parser produces predicate-argument representations and these are added as instances to a small generic ontology which represents the world model; the world model plus the instances is considered to be the discourse model. Once the discourse model is built, the system searches for instances that could be merged into one—coreferent instances—by comparing their attributes. The authors adopt specific comparison rules for proper names, common nouns and pronouns. Their system has reached 71.93% precision and 50.71% recall using the MUC scoring system. Azzam *et al.* [1998] extends the coreference resolution algorithm implemented in LaSIE with an improved version of Sidner's focusing approach, which is able to handle more complex sentences and intrasentential reference. Azzam *et al.* conclude, however, that there is no observable difference between the performance of the coreference algorithms with and without focusing. They report that the main limitation of the focus-based approach is its reliance on robust syntactic and semantic analysis in order to find the focus.

3.2.1.2 Factor-based approaches

Factor-based approaches combine various knowledge sources, including morphological, lexical, syntactic, semantic, and in some cases pragmatic information, in the form of constraints and preferences (factors). Constraints are applied in order to remove bad antecedents, and preferences are used to rank candidates that satisfy all constraints. In contrast to discourse-based, factor-based approaches do not rely on an elaborate discourse theory, although some discourse information can be formulated as preferences (rather than constraints).

Carbonell and Brown's [1988] work is an example of a factor-based algorithm for pronoun resolution. Various constraints are proposed: gender and number agreement, semantic (e.g. selectional constraints) and pragmatic constraints (e.g. considering whether some action that occurs between the antecedent candidate and the anaphor implies that they cannot take part in an anaphoric relation). As preferences, they have considered recency, topicalisation, syntactic parallelism and semantic parallelism (having the same thematic role as the anaphor) in order to select an antecedent. They tested their algorithm on a small test suite containing 27 pronouns, from which 23 (85%) were resolved correctly.

Lappin and Leass' [1994] pronoun resolution algorithm relies on a set of syntax-based constraints and salience-based preferences. In contrast to Carbonell and Brown, who make use of semantic and pragmatic constraints that are generally hard to encode with reasonable accuracy, Lappin and Leass instead employ only morphological constraints such as gender and number agreement, and syntactic constraints such as the requirement that the antecedent and the pronoun do not be arguments of the same head constituent. They assume, however, perfect output from a morphological analyser and a full syntactic parser. The salience factors are, for example, sentence recency, grammatical role, syntactic parallelism, among others; each salience factor is associated with an initial weight that indicates the contribution of the factor to overall salience. These weights are lowered once the distance between the anaphor and the antecedent candidate increases. An anaphoric NP is resolved to the most salient preceding entity. Once an anaphoric NP is resolved it is added to the antecedent's equivalence class. The salience of an entity is given by the salience of the equivalence class to which the candidate NP belongs, while the salience of the class is calculated on the factors applied to each of its members. Lappin and Leass' algorithm was able to correctly resolve 86% of the pronouns in their test set.

Due to the high error rate in case of full syntactic parsing, several alternatives to full parsing have been proposed ranging from partial parsing (e.g. [Kennedy and Boguraev, 1996]) to part-of-speech tagging (e.g. [Mitkov, 1998]). Kennedy and Boguraev modify the Lappin and Leass

algorithm in a way that it works on a flat syntactic analysis, provided by a part-of-speech tagger and a noun phrase grammar. Their system reaches 75% accuracy. Mitkov follows the same approach as both previous works, but instead uses only part-of-speech information to identify the noun phrases in a context of two sentences. Mitkov included additional factors to select the antecedent, for example giving preference to definite noun phases, counting the number of times the candidate NP is mentioned in the same paragraph, checking whether the candidate NP is in the heading of the section, etc. Mitkov's algorithm correctly resolves 86% of the pronouns in their evaluation data.

Meyer and Dale [2002] have created a factor-based algorithm, inspired by Lappin and Leass, but to handle definite descriptions. They have developed special factors to work as indicators of associative anaphora cases. They first extract "associative axioms" from the corpus. These are patterns that are evidence of association between two words (e.g. of-phrases, like "the leg of the giraffe", indicating a relation between leg and giraffe and forming the axiom `have(giraffe,leg)`). Secondly, they seek to generalise the axioms by searching for hyponym words in WordNet, so that ideally they can infer a more general pattern like `have(living thing, body part)`. The generalised axioms are used as a constraint in the resolution algorithm, so that candidates that do not fit any axiom can be eliminated. They have evaluated the performance of their system on resolving associative cases using different levels of generalization over WordNet: on the lowest level they reach 31-45% precision and 39-64% recall, and on the highest level they reach 8-11% precision and 79-91% recall.

A disadvantage of factor-based approaches is that the weights assigned to each factor have to be manually set by the developer. The works mentioned above have not presented an evaluation of the influence of variation in weight values.

3.2.1.3 Syntax-based approaches

Syntax-based approaches rely solely on syntactic and morphological information. For each potential anaphor, the search for an antecedent is performed via the traversal of parse trees.

One of the early approaches to coreference resolution which is still popular is Hobbs's syntax-based approach [Hobbs, 1978] for pronoun resolution. The algorithm considers the sentences in the text in reverse order, starting from the sentence in which the pronoun appears and searching for potential antecedents in the corresponding parse trees in a left-to-right, depth order that obey binding and agreement constraints. The algorithm's preferences for recency as well as for NPs in the subject position are generally believed to be the reason for its good performance on pronouns with intra-sentential antecedents [Lappin and Leass, 1994]. A match is found when the antecedent NP in question and the anaphoric pronoun agree in gender, number and person. Hobbs also uses selectional restrictions to rule out bad candidate antecedents. Hobbs did a hand-based evaluation of his algorithm on 100 pronouns from each of three different texts: a history chapter, a novel, and a news article. The algorithm performed successfully on 88.3% of the cases; accuracy increased to 91.7% with the inclusion of selectional constraints.

Syntax-based approaches are limited to pronoun resolution, since the resolution of other types of NPs is not as closely tied to syntactic structures.

3.2.1.4 Heuristic-based approaches

Heuristic-based approaches are composed by a set of hand-crafted heuristics for selecting an antecedent.

Vieira and Poesio [2000] have developed an heuristic-based approach to resolve definite descriptions. They have created three sets of heuristics: one for identifying direct anaphora (cases where the antecedent and the anaphor have the same head noun), another set for bridging anaphora (cases where the antecedent and anaphor have different head nouns; it includes associative anaphora), and a third set to identify discourse-new definite descriptions. They integrate

the three sets of heuristics by applying them in a particular order. They first apply the direct anaphora heuristics (basically, seeing if there is a previous NP with the same head noun as the anaphor, considering some restriction on pre- and post-modification). If these are unable to determine and antecedent, then the discourse-new heuristics are applied (e.g. considering the presence of special predicates such as "the first", "the best", restrictive postmodification, appositive or copular constructions). If the anaphor does not fit the discourse-new heuristics, then the bridging heuristics are applied (e.g. checking whether the anaphor's head noun match any of the antecedent's pre-modifiers, or whether anaphor and antecedent head nouns are part of the same WordNet synset, or whether they hold hyponymy/hypernymy, co-hyponymy or direct meronymy/holonymy relations in WordNet). Because the performance of their heuristics for bridging cases was considered poor, they evaluated their system with and without them. Using only the heuristics for direct anaphora and discourse-new cases, their overall performance on test data was 62% F-measure, 76% precision and 53% recall. With the inclusion of the bridging heuristics (bridging cases comprise 8% of the cases in their corpus), the overall performance became 62% F-measure, 70% precision and 57% recall. The use of WordNet for dealing with bridging anaphora was not very successful since (1) WordNet is a generic knowledge base, where all meanings of a word are included, resulting in false positive antecedents, and (2) WordNet is not complete enough and its organisation is not always clear (only 46% percent of the semantic relations present in their bridging cases could be found in WordNet).

Hand-crafting the heuristics is the main problem of this type of approach. It is a very complex task to create heuristics that cover all cases of anaphora and to prioritise the rules when their outcomes diverge.

Poesio *et al.* [2002] replaced the use of WordNet in Vieira's system with automatically acquired lexical knowledge in order to solve specifically the cases that involved meronymy. They have adopted a similar technique to that used by Hearst [1992] to extract hyponyms. They achieved 72.7% precision and 66.7% recall on resolving the bridging cases that involved meronymy, while WordNet could recover only 25% of the cases.

Bunescu [2003] also developed an heuristic-based system for resolution of definite descriptions, both coreferent and associative anaphora. For each anaphor-candidate pair, the author searches the Internet for the pattern "`<candidate's head noun>. The <anaphor's head noun> <verb>`" and from its frequency computes the mutual information between the anaphor and the antecedent candidate (a minimun frequency threshold is considered). The candidate that ranks highest is selected as antecedent. The author has experimented with several frequency threshold values: with a high threshold value, the system reached around 70% precision and 10% recall; with the lowest threshold the system reached its highest recall, around 42%, with 23% precision.

The main disadvantage of knowledge-based approaches in general is that they are conservative, usually only covering cases that are predicted by the developers. These approaches restrict the range of cases that can be resolved, since the framework at hand does not handle unpredicted types of cases. Besides, manually building and tuning rules and/or weights can be an expensive task, demanding great effort from the coder.

3.2.2 Corpus-based systems

Corpus-based approaches rely on manually annotated corpora as source of knowledge of a given task. Given the successful application of corpus-based approaches to several NLP tasks and the availability of corpora annotated with coreference information since the MUC efforts [Hirshman and Chinchor, 1997], researchers have attempted to apply corpus-based methods to anaphora and coreference resolution. Corpus-based algorithms are trained on real-world texts and hence are, in principle, more robust than knowledge-based systems. While a knowledge-

based system encodes its beliefs in the form of hard constraints, corpus-based systems learn soft constraints from annotated corpora and can therefore weight the available information and take into account exceptional cases [Ng, 2003]. The importance of each factor involved in the resolution process can be inferred by the distribution of cases in the corpus, provided that the corpus is representative.

Besides the resolution process, corpus-based approaches include a training process to extract from the corpus the information to be used by the resolution model. Corpus-based approaches interpret the anaphora resolution problem as a classification task: an anaphor-candidate pair is classified as coreferent/anaphoric or not; the probability of this relation is determined by the model according to what it has seen in the training corpus.

Each training instance (i.e. the anaphoric relation, or the absence of it, between two NPs) is described by a set of features which usually includes relational features (which test whether some property holds for the NP pair under consideration, e.g. head-noun matching) and non-relational features (which test some property of one of the NPs under consideration, e.g. the type of NP: pronoun, definite description, etc.). An instance is labelled as positive if the two NPs possess an anaphoric relation, and labelled as negative otherwise. Corpus-based approaches differ from each other in terms of how the model is learned and can further be divided into two classes: machine-learning and statistical approaches. In machine-learning approaches, the resolution model is induced from the training data according to a learning algorithm, while in statistical approaches, a probabilistic resolution model is built independently of the training data (although its development may be guided by a corpus) and the data is used solely to compute the statistics required by the model. While there are algorithms that can induce probabilistic models automatically from the training data, these would be classified here as machine-learning approaches.

3.2.2.1 Machine-learning approaches

Machine learning techniques have gained popularity in the research on coreference resolution. Some particular learners have been widely used, for example, the C4.5 decision tree learner [Quinlan, 1993] was used by Aone and Bennett [1995], McCarthy and Lehnert [1995], Soon *et al.* [2001], Strube *et al.* [2002], and the Ripper rule learner [Cohen, 1995] was used by Ng and Cardie [2002b, 2002c] and Uryupina [2003].

Aone and Bennett describe a system for resolving anaphora occurring in Japanese texts about joint ventures. They treat proper names, definite descriptions, zero pronouns and quasi-zero pronouns. The representation of each instance consists of 66 features, including lexical (e.g. part-of-speech), syntactic (e.g. grammatical role), semantic (e.g. semantic class), and positional features (e.g. distance between the potential antecedent and the anaphor). Two different methods are used to create positive training instances: transitive, where an instance is formed between a NP and each of its preceding NPs in the same anaphoric chain, and non-transitive, where an instance is formed between a NP and its closest preceding NP in the same anaphoric chain. Negative instances are generated by pairing a NP with each preceding NP that does not have an anaphoric relation with it. The system then uses the C4.5 decision tree induction system to train an anaphora classifier that determines whether two NPs possess an anaphoric relationship. Their best results using the transitive training strategy was 77.30% F-measure (86.73% precision and 69.73% recall). Using the non-transitive strategy, their precision increased but recall dropped: they reached 67.03% F-measure (89.74% precision and 53.49% recall).

McCarthy and Lehnert describe a coreference resolution system called RESOLVE, which also handles texts from the domain of joint ventures. 3 of the 8 features used are domain-specific; for example, there are features that test whether each of the NPs in the pair refers to a joint venture company. The domain-independent features can be characterised as lexical (e.g. check

whether the two NPs share a common phrase), semantic (e.g. check whether one NP is an alias of the other), and positional (e.g. check whether the two NPs are in the same sentence). No syntactic feature is used. To generate positive training instances from coreference chains, only the transitive method is used. Negative training instances are generated by pairing a NP with each of its preceding non-coreferent NPs. They also adopt the C4.5 decision tree algorithm as their classifier. Their best results were achieved using an unprunned tree: 86.5% F-measure, 87.6% precision and 85.4% recall.

Soon *et al.* adopt a knowledge-lean approach to a general-purpose coreference resolution system. They handle all NP types. They used the C5 decision tree learner (updated version of the C4.5), and it uses 12 surface-level features, which are all designed to be domain-independent: one lexical feature (string matching), eight grammatical features (gender and number agreement, apposition, and NP types), two semantic features (semantic class agreement and aliasing), and one positional feature (number of sentences between the two NPs). The non-transitive method is used to generate positive training instances from coreference chains. To reduce the ratio of negative to positive instances, only the negative instances where the anaphor is paired with NPs that are closer than the closest correct antecedent are considered. They have trained and tested their system on the MUC-6 and MUC-7 coreference data. They report 62.6% F-measure, 67.3% precision and 58.6% recall on the MUC-6 test data, and 60.4% F-measure, 65.5% precision and 56.1% recall on the MUC-7 test data. They also present the results of a feature selection experiment, where they trained the classifier with one feature at a time. This experiment indicated that string matching, aliasing and apposition are strong indicators of coreference.

Ng and Cardie have extended the work from Soon *et al.* They have largely expanded the feature set, using a total of 53 features, adding lexical (e.g. new features to account for more flexible string matching, such as head pre-modifier matching), semantic (e.g. measuring WordNet distance between head nouns), positional (including a distance measure in number of paragraphs), knowledge-based (adding the result of a knowledge-based algorithm for the NP pair as a feature) and mainly grammatical features (e.g. determining NP type, checking NP embedding, grammatical role, binding constraints) that include a variety of linguistic constraints and preferences. They have experimented with the C4.5 decision tree algorithm and Ripper rule induction algorithm. When using all the proposed features, they achieved 63.8%/61.6% F-measure (on MUC-6/MUC-7 test data, respectively), 58.3%/58.2% precision and 70.3%/65.5% recall using the C4.5 algorithm, and 64.5%/61.2% F-measure, 62.2%/60.6% precision and 67.0%/61.9% recall using the Ripper algorithm. These performance scores are lower than those achieved by their reimplementation of Soon *et al.*'s algorithm. They report that the poor performance on resolving common nouns was responsible for lowering the overall scores; for instance, they achieved 40.1%/45.2% precision on common nouns using C4.5. To overcome this, they have manually selected a high-precision subset of their features, which returned the expected improvement in precision (with smaller drops in recall). They reached 69.1%/63.4% F-measure, 74.9%/70.8% precision and 64.1%/57.4% recall using the C4.5 algorithm, and 70.4%/63.1% F-measure, 78.0%/72.8% precision and 64.2%/55.7% recall using the Ripper algorithm.

Machine learning techniques vary in terms of complexity and number of parameters that are required to be set by the developer. The more complex the learning algorithm used, the more training data are required for the system to induce a stable and reliable model.

3.2.2.2 Statistical approaches

Statistical approaches consist of a probabilistic model which uses the training corpus as source of the statistics required to estimate its probability terms. Statistical approaches for anaphora resolution aim to determine the probability that a NP is the antecedent of a given anaphor. The probabilistic model combines different sources of information as parameters (features) within probability equations.

Ge, Hale and Charniak [1998] proposed a probabilistic model for resolving third-person pronouns. The model consists of a probability equation, which is initially conditioned on a number of features and is then simplified to handle the sparseness of the training data. This approach consists of decomposing the probability equation for the model by discarding dependencies between features. The decomposition is done by making use of Bayes' rule, the chain rule and certain independence assumptions. The features used by their model encode positional information (the distance between the pronoun and the candidate antecedent), grammatical information (gender and animacity of the candidate antecedent), semantic information (selectional preferences based on the governing constituent of the pronoun), and a crude measure of salience (a mention count of the candidate antecedent). The authors show how the equation for the model is decomposed in factors that preserve only few dependencies among the features and each factor represents a source of information relevant for anaphora resolution. Statistics for each of the factors are collected from the training corpus. For a given anaphoric pronoun, the candidate antecedent that is assigned the highest probability by the model is selected as the antecedent. They have trained their system on a small corpus, and have reached 82.9% accuracy performing 10-fold cross validation. They also measure the importance of each information source in an incremental way, and conclude that gender and animacity information contributes the biggest improvement in performance.

The main advantage of statistical approaches like Ge *et al.*'s is their simplicity, and consequently the possibility of learning from a small amount of data. Since this type of model is non-parametric, all weights come from the distribution present in the training data.

Statistical approaches (as we define them here) are not induced from the training corpus: the corpus is used solely to provide the necessary statistics, so while the corpus still needs to be representative, it can, in principle, be smaller than the corpus needed to induce a machine-learning system.

The possibility of training an anaphora resolution system on a small corpus is particularly attractive to the biomedical domain, given that a corpus of biomedical scientific articles annotated with anaphora information is not available and one would need to start building a corpus from scratch.

3.3 Anaphora resolution in biomedical text

Biomedical text differs from that of other genres (e.g. newswire, fiction) in the aspects described in Section 2.1 from Chapter 2. Among these aspects, those which most influence anaphora are the NP-type distribution, the background knowledge assumed by the writer about the reader, and the writing conventions adopted in the domain to refer to biomedical entities.

Different types of NPs have a particular distribution in biomedical articles. For example, pronouns are very rare, accounting for a very small percentage of the noun phrases, while proper names occur very often, given the frequent mention of the names of biomedical entities. A system for anaphora resolution in the biomedical domain can benefit from focusing on the most common types of noun phrases, that is, non-pronominal.

Concerning background knowledge, the reader is required, in order to understand the text, to understand the underlying relation between the entities therein mentioned. For example, in the sentence below,

(6) ``The expression of **reaper** has been shown ... **the gene** encodes ...''

the reader has to be able to understand that reaper is a gene (given the context), so that he/she can capture the anaphoric relation and understand the content of the sentence. This aspect emphasises the need for semantic information as a feature in the anaphora resolution process. The biomedical domain is fortunately rich in resources that can provide semantic information,

like those described in Chapter 2 (e.g. databases, UMLS, GO, SO, etc.).

Another aspect affecting the anaphoric relations are the writing conventions adopted in the biomedical domain to distinguish between a gene name and a protein name. The most usual convention is writing gene names with lowercase italicised letters and protein names with non-italicised uppercase letters. The existence of such conventions allows for associative anaphora between proper names, which is not seen in other domains, as in the example:

(7) ``Drosophila has recently been shown also to have a CED-4/Apaf-1 homolog, named **Dark/HAC-1/Dapaf-1**. ... Like Apaf-1 and CED-4, loss of function mutations in ***dark/hac-1/dapaf-1*** result in a reduction in developmental programmed cell death.''

Very few systems for anaphora resolution have been developed for the biomedical domain. Castaño *et al.* [2002] developed a salience-based system for anaphora resolution (similar to the Lappin and Leass system for pronoun resolution). It seeks to resolve pronouns and nominal (which they call sortal) anaphora. The resolution process relies on lexical information (they compute a score of string similarity), grammatical features (e.g. number agreement), and semantic information (matching between semantic types derived from UMLS), which are used to compute a salience score for each antecedent candidate, and the most salient is selected. They have developed the Medstract corpus in order to evaluate their system. It is composed of a set of Medline abstracts where mentions of biomedical entities have been classified according to UMLS and anaphoric relations tagged. The system's best performance on pronouns was 80% precision and 71% recall and on sortals, 74% precision and 75% recall. The authors argue that UMLS is too coarse-grained, and assume that a finer-grained typing strategy would help to increase the precision of the anaphora resolution system.

Gaizauskas *et al.* [Gaizauskas *et al.*, 2003] developed the PASTA system, which is an adaptation of the general LaSIE information extraction system to the biomedical domain, more precisely to the extraction of the roles of specific amino acid residues in protein molecules. Their coreference resolution module, which works on top of an ontology-like representation of the discourse, populated by instances collected from the text, was presented in Section 3.2.1.1 above. For treating biomedical texts (rather than news articles used by LaSIE), they have changed the classes of named-entities considered, and the world model (which is instanciated with entities from the text to become the discourse model) had to be adapted to represent a domain model, containing as concepts "proteins", "residues" and "species" (instead of "persons", "organisations", "locations", etc.). They evaluated their information extraction system on a corpus of 1513 Medline abstracts, but have not reported on the performance of the coreference resolution module alone on the new domain.

Yang *et al.* [2004] evaluate a supervised machine-learning approach for anaphora resolution on a portion of the GENIA corpus, which is tagged with semantic information based on the GENIA Ontology. They focus only on coreferent cases and do not attempt to resolve associative links. Their system is similar to that of [Soon *et al.*, 2001]. It uses 18 features to describe the relationship between an anaphoric expression and its possible antecedent, and also adopts a decision tree algorithm. They achieved recall of 80.2% and precision of 77.4%. They also experiment with exploring the relationships between NPs and coreferential clusters (chains), which are formed during the resolution process: the first two NPs that are found to be coreferent start a cluster, and following NPs are checked against the cluster to verify whether they are coreferent. Thus selecting an antecedent is not based just on a single candidate but also on the cluster that the candidate is part of. For this they add 6 cluster-related features (e.g. string matching to any NP in the cluster, number of elements in the cluster) to the machine-learning process, and are able to improve their system performance, achieving 84.4% recall and 78.2% precision.

Kim and Park [2004] developed the BioAR system to resolve anaphoric mentions of proteins in order to link them to the protein record at the Swiss-Prot database. The anaphoric protein mentions to be resolved were extracted by an information extraction system, BioIE, which finds protein-protein interactions. They consider pronouns and all NPs with determiners as anaphoric expressions. For resolving pronouns they use a centering-like algorithm, and for resolving the other NPs, they use a similar system to Castaño *et al.* To filter out mentions that usually contain the article "the" (definite NPs) but are not anaphoric (e.g. "the nucleous", "the yeast Saccharomyces cerevisiae"), they have created a list of cellular component names, a list of species names, and a list of patterns which represent the internal structures of some non-anaphoric definite NPs (e.g. apposition). They achieve 75% precision and 56% recall on pronoun resolution and 75% precision and 52% recall on nominal anaphora resolution.

All these systems for anaphora resolution in the biomedical domain have been developed and tested on abstracts of biomedical articles, which represent a restricted use of anaphora. There is clearly a need to develop a system for tackling anaphora in full-text articles, since these contain the main source of data to be automatically extracted by any information extraction effort.

3.4 Evaluation of anaphora resolution systems

Anaphora resolution systems are usually evaluated against a gold-standard corpus where anaphoric relations have been manually annotated. The performance of anaphora resolution systems has been measured using Precision and Recall scores. There has been considerable discussion on how to calculate precision and recall when the output of the resolution system consists of coreference chains. The key issue when evaluating coreference chains is how to score chains that are partially correct (missing or exceeding some elements).

MUC-6 has proposed a scoring system that compares the coreference chains returned by a system with the coreference chains from a gold-standard corpus. The MUC-6 scoring scheme [Vilain *et al.*, 1995] compares equivalence classes defined by the coreference links, instead of comparing the links themselves. The recall score is obtained by determining the minimal number of links missing in the system response that are required to transform its corresponding equivalence classes into those formed by the gold-standard links. Assuming S as an equivalence class from the gold-standard, recall is computed as follows, for all i equivalence classes:

$$R = \frac{\sum i(c(S_i) - m(S_i))}{\sum ic(S_i)}$$

where $c(S)$ is the minimal number of links necessary to generate the equivalence class S—$c(S) = (|S| - 1)$. $m(S)$ is the number of missing links in the system response relative to S—$m(S) = (|p(S)| - 1)$; $p(S)$ is the number of subsets into which the system response partitions the gold-standard equivalence class. To compute precision, the roles for the gold-standard and the system response are reversed: S is assumed to be an equivalence class from the system response, and the missing links to turn the gold-standard equivalence classes into the system response are calculated.

The MUC-6 scoring algorithm, however, has two major shortcomings according to Bagga and Baldwin [1998]. The algorithm does not give any credit for separating out singletons (entities occurring in chains only consisting of one element). Nor does it distinguish between different types of errors. The authors argue that some errors do more damage than others; for example, they argue that a mistaken link between elements of two long coreference chains is more damaging than a mistaken link that merges shorter chains. Despite this, the MUC scoring system has continued to be used to evaluate coreference resolution systems.

In order to evaluate associative anaphora resolution rather than coreference relations, no specific scoring scheme has been proposed. Previous work have computed precision and recall in the usual away, comparing the associative links themselves in the system response and in the gold standard [Vieira and Poesio, 2000, Bunescu, 2003].

3.5 Summary

In this chapter we have discussed the concepts of anaphora and coreference, and have described systems for anaphora resolution. We have discussed the general steps of an anaphora resolution system and have classified the systems according to their resolution approach: knowledge or corpus-based. Knowledge-based approaches rely on theoretical models or manually built rules and do not require any training data; these aspects characterise them as conservative models that have difficulty handling unusual/unforeseen cases.

Corpus-based approaches, on the other hand, learn from training data and are consequently more flexible. These approaches can combine different sources of information (features) in a soft way: the relevance of each feature is balanced by its prominence and frequency in the training instances. Among the corpus-based approaches that we presented, statistical approaches appear to be an interesting option when the training corpus available is small, since the corpus is used to collect statistics that will fit into a previously defined probabilistic model, instead of being used to induce a resolution model, as in machine-learning approaches. Although statistical approaches also require that the corpus be representative, it could, in principle, be smaller than the corpus required to induce a reliable model using a learning algorithm. Thus the statistical approach appeals to efforts in the biomedical domain where no corpus of scientific articles annotated with anaphora information is available.

Chapter 4

Biomedical entity recognition and classification

An essential step in information extraction is the identification of the NPs that refer to the entities about which one wants to extract information. In molecular biology texts, the central entity of interest is the gene, then entities related to the gene, like its products (e.g. proteins), its parts (e.g. codons), its variants (e.g. mutants), among others.

Among those there are named and unnamed entities. Genes and proteins have names; sometimes gene parts also take the gene's name, and gene variants receive variants of the gene's name. To identify these names in the text, we require a named-entity recogniser. Recognising gene/protein names is considered more challenging than recognising other named entities (e.g. city names, person names, company names), given the issues discussed in Chapter 2, which mainly concern the overlap with common English words and similarity to general acronyms.

To recognise unnamed biomedical entities, a simple approach is to have a list of the entities of interest and to mark them up in the text. The main challenge in this case is to compile a complete and coherent list and allow for inflectional and typographical variants.

Besides identifying the entities, it is also important to classify them according to a given set of classes of interest. The class information is useful for tasks that aim to find relations between the entities, which could be linguistic relations such as anaphora, or biological relations such as protein-protein interaction.

In this chapter we shall describe our strategy for recognising and classifying named and unnamed entities in molecular biology texts, more specifically in the fruit fly literature. For named-entity recognition (NER) we have adopted the system developed by Vlachos *et al.* [2006] (Section 4.1), whose goal is to identify and mark up gene names in the text. For the recognition of unnamed entities we have developed a dictionary-based approach based on the Sequence Ontology (SO) [Eilbeck and Lewis, 2004] (Section 4.2), which is responsible for identifying in the text noun phrases whose head nouns refer to biomedical entities and classifying them according to the relations present at SO. As a prerequisite, we require a syntactic parser that is able to indicate the noun phrase boundaries and its subconstituents (e.g. head noun, head modifiers) – for that we have adopted the RASP parser [Briscoe and Carroll, 2002]. Only after these steps, once we have identified all mentions to biomedical entities in the text, we can consider looking for relations among them, such as anaphora. Figure 4.1 summarises how the information from different levels of processing is combined. It shows (using XML mark up to illustrate) that each level adds up linguistic information to the text. This additional information is essential to accomplishing anaphora resolution.

4.1 Gene/protein name recognition

The NER system we use was developed by Vlachos *et al.*, and it is a replication and extension of the system developed by Morgan *et al.* [2004]: a different training set and software were used. The main characteristic of both systems is the generation of training data by automatically annotating Medline abstracts with the names, symbols and synonyms of the genes with which they were associated in FlyBase. As seen in Chapter 2, each fruit fly gene has an entry in

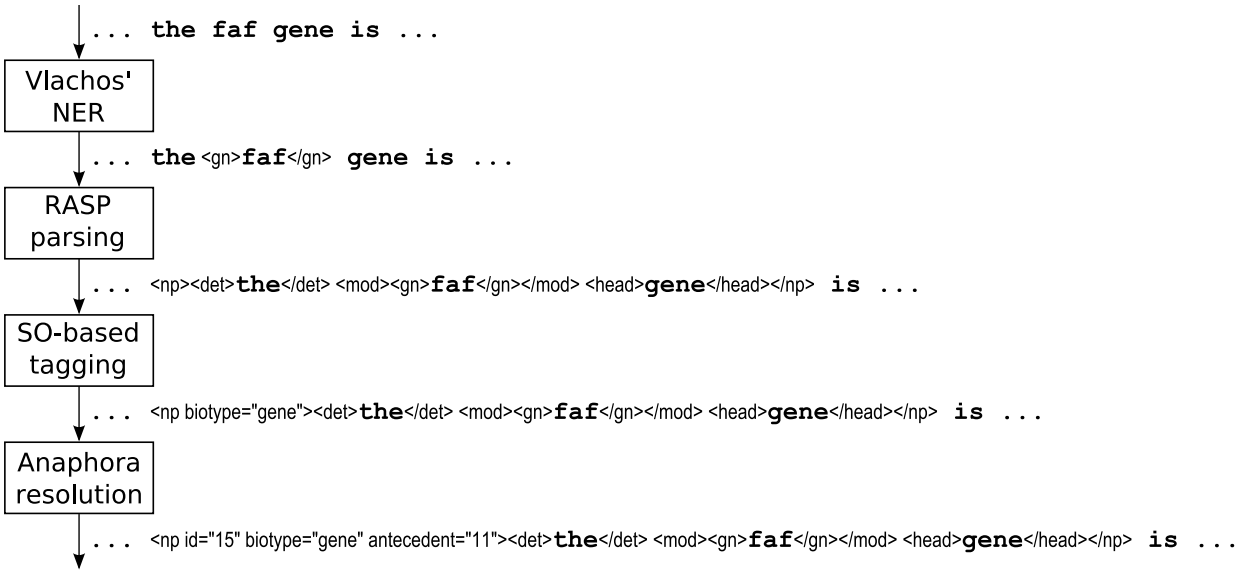


Figure 4.1: Pipeline for anaphora resolution

FlyBase, and each entry contains links to the publications where that gene is discussed. These links lead to the PubMed identifier for the abstract of each publication, so the abstracts can be recovered, the terms used to refer to the associated gene can be tagged, and the abstract can already be part of a training set. This strategy for generating training data automatically makes it possible to create a large training set, although it is not always accurate: as Morgan *et al.* note, the occurrence of gene synonyms that match common English words, such as "to" and "by", leads to the incorrect annotation of common words as gene names, resulting in precision errors in the training data; on the other hand, some genes that are mentioned in the abstracts might not be associated with the article in FlyBase, as FlyBase curators only consider some relevant sections of the article when curating, resulting in recall errors.

Vlachos *et al.* used a total of 16609 abstracts. These abstracts were split in sentences and tokenised using the RASP toolkit [Briscoe and Carroll, 2002], and were then automatically annotated as described. They were used to train a gene-name recogniser; the recogniser used was the open source toolkit LingPipe[1], implementing a 1st-order HMM model using Witten-Bell smoothing. To deal with gene names that had not been seen in the training data, a morphologically-based classifier was used. LingPipe achieves high precision by only generalising to unseen names in lexical contexts that are clearly indicative of gene names in the training data.

The recogniser was tested on a dataset developed and used by Morgan *et al.*; it consists of 86 abstracts containing about 7800 distinct gene names (referring to 5243 distinct genes) annotated by a biologist curator and a computational linguist. Its average performance was 82.54% recall and 79.84% precision.

4.2 Selecting and classifying biomedical entities

The first step towards identifying the NPs that refer to biomedical entities is to recognise all NPs (and their subconstituents) in a sentence. For that, we have parsed the sentences using RASP, which recognises the NP boundaries, its head and modifiers. After that, we tag all NPs that refer to biomedical entities according to our approach, which uses information from the NER

[1] http://www.alias-i.com/lingpipe/

module and the Sequence Ontology. Finally, we filter out all NPs that are not considered to refer to biomedical entities, and take those remaining to be considered for anaphora resolution.

4.2.1 Parsing and NP extraction

In the named-entity recognition step, RASP is used to detect sentence boundaries and to tokenise sentences. So, in this step, we continue using RASP to tag the tokens with their part-of-speech (PoS) and finally to parse PoS tag sequences. Before parsing though, we change the PoS tag of all tokens that had been recognised as gene names to the appropriate proper name tag—since RASP considers gene names to be unknown words, this improves parser performance as the accuracy of PoS tagging decreases for unknown words. RASP's tagger uses an unknown-word handling module which relies heavily on the similarity between unknown words and extant entries in its lexicon; this strategy works less well on gene names and other technical vocabulary from the biomedical domain, as almost no such material was included in the training data for the tagger.

The RASP parser outputs grammatical relations (GRs) for each sentence that is parsed [Briscoe *et al.*, 2006]. GRs are factored into binary lexical relations between a head and a dependent of the form `(GR-type head dependent)`. To find the NP head nouns, we consider the RASP GR types presented in Table 4.1, in which dependent slots are nominal; column 2 describes how the parser compiles these GRs. We also consider the same GRs when the noun slot is filled by a conjunction (e.g. `(ncsubj verb conj)`, in which case we look for complementary `(conj conj noun)` GRs, which encode relations between a coordinator and the heads of a conjunct. There will be as many such binary relations as there are conjuncts of a given coordinator; e.g. for "CED-9 and EGL-1 belong to a large family ..." we get (**`ncsubj`** `belong and`), (**`conj`** `and CED-9`) and (**`conj`** `and EGL-1`). To complete the NP, we look for GRs that contain determiners and pre-modifiers of the head nouns found, as shown in Table 4.2; we have adopted the concept of "base NPs", where we don't consider post-modifying clauses, so there are no overlapping base NPs [Lewin, 2007].

GR	Description
`(ncsubj verb noun)`	relation between non-clausal subjects and their verbal heads
`(dobj verb noun)`	relation between a verbal head and the NP to its immediate right
`(dobj prep noun)`	relation between a prepositional head and the NP to its immediate right
`(obj2 verb noun)`	relation between verbal heads and the head of the second NP in a double object construction
`(ta * noun)`	relation between the head of an NP or clause and the head of a text adjunct delimited by punctuation (quotes, brackets, dashes, commas, etc.), e.g. for "BIR-containing proteins (BIRPs)" we get `(ta proteins BIRPs)`.

Table 4.1: GRs used for NP extraction

Having done that, we extracted all NPs, with information about which elements are its head, modifiers and determiner, so we can start classifying the NPs according to the biomedical entities to which they refer.

The GR-based NP extraction strategy has recently been extended to take advantage of NER information for the ranking of n-best lists of GRs, derived from parsing alternatives for a sentence [Lewin, 2007].

GR	Description
`(det noun determiner)`	relation between articles, quantifiers, partitives and other single word forms which can begin NPs, and the NP head.
`(ncmod noun modifier)`	relations between non-clausal modifiers and the NP head, e.g `(ncmod genes msl)`.

Table 4.2: GRs used for finding head noun complements.

4.2.2 Typing biomedical NPs

After finding all NPs in the text, we would like to type them in order to be able to select those that refer to biomedical entities. For that to be possible, we have associated what we call "biotypes" to terms referring to biomedical entities. We have adopted the Sequence Ontology (SO) as our source of relevant terms and also as our source of relationships between the terms.

As described in Chapter 2, SO focuses on the molecular biology subdomain. It includes most vocabulary necessary to describe biological sequencing, from genes to proteins, and classifies the terms in a subsumption and relational network. However, as an ontology, SO can have several levels of relations linking two concepts; for example to find the relation between the concepts of gene and protein, there are several intermediate relations and concepts that constitute the path between the concepts.

In order to fit SO for the task of typing biomedical entities in the text, we have reorganised and simplified it in order to eliminate the intermediate levels between concepts of our interest. We have restructured SO's relations in order to give the gene a central role, so that we could divide the terms in classes according to their relation to the concept of a "gene"; these classes are our "biotypes". A gene may be defined as a sequence of DNA that encodes some biological function; specified sequences within genes are considered parts of the gene; and the units of function encoded by the gene are considered its products (intermediate products such as polypeptides or the final product, proteins). Different versions of a gene sequence are considered variants of a gene, and specific kinds of genes are seen as subtypes of genes (e.g. oncogenes). Portions of sequence that are broader than the gene are called "supertypes" of genes. Reorganising SO concepts in a limited set of classes helps us to consider indirect relations that would otherwise span several levels in the ontology.

The first step in the process of restructuring SO was to look for the path between a gene and its final product, a protein, through the *is-a*, *part-of*, and *derived-from* relations available. We got to the path shown in Figure 4.2.

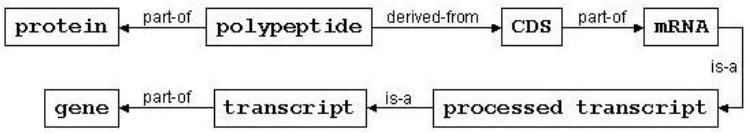


Figure 4.2: Sequence Ontology path from gene to protein

From this path, considering the gene as central, we made the following assumptions, that were reviewed and accepted by a biologist:

- whatever is-a transcript is also part-of a gene;
- whatever is-a processed transcript is also part-of a gene (consequently, mRNA is part-of a

gene);

- whatever is-a mRNA is also part-of a gene;
- whatever is part-of a mRNA is also part-of a gene (consequently, CDS is part-of a gene);
- whatever is derived-from a part-of a gene is a product of a gene (consequently, polypeptide is a product of a gene);
- whatever is-a polypeptide is also a product of a gene;
- whatever is part-of a polypeptide is also a product of a gene;
- whatever is composed by polypeptides is also a product of a gene (consequently, protein is a product of a gene).

On these assumptions, we decided to extract from SO all the entries related to the concepts in the path above, that is, all items related to "gene", "transcript", "mRNA", "CDS", "polypeptide", and "protein" by *part-of*, *is-a*, or *derived-from* relations, and organise them into three groups of terms: genes, parts of genes and gene products. For example, a "riboswitch" is an mRNA, so it is grouped together with mRNA as part of a gene; an "UTR" is a part of an mRNA, so it is also part of the same group (parts of genes). The group of gene products has been divided in two, proteins (final products) and parts of products (intermediate gene products), because we were interested in keeping the distinction between the two to be able to represent relations between mentions from these two groups. We also extracted entries referring to types of genes, which were included in the ontology under an entry called 'gene_class' (rather than an extra relation type), and entries referring to variants of genes, which were indicated by the variant-of relation in the ontology; this led to the creation of two more groups: subtypes of genes and gene variants. Finally, we created a group to represent DNA sequences that are greater than a gene, what we call supertypes. In summary, we have seven groups of entities, and consequently of terms referring to these entities, and each group represents a biotype. We have then the following biotypes: "gene", "product", "subtype", "part-of", "part-of-product", "supertype", and "variant". Figure 4.3 presents all the information extracted from SO. Each block correspond to one group of entities; inside the blocks we show the terms extracted from SO for each group – indented entries hold an *is-a* relation with the upper entry, and entries preceded by '*', a *part-of* relation. The arrows represent the relations between the blocks, from which the biotypes are derived.

During the corpus annotation phase (described in the next chapter) in which we used the above biotypes to type the entities, we encountered mentions of biomedical entities that could not be found among the terms that we extracted from SO. These mentions referred to entities that fit at least one of our biotypes, but which were referred to by alternative (more specific) terms. Because we aimed at typing all mentions of biomedical entities in the text, we felt the need to expand our groups of terms in order to include those that were not contemplated by SO. We observed that the missing terms referred essentially to: types of proteins (e.g. kinase, enzyme), types of parts of product (e.g. bipeptide, motif), terms related to homology between genes[2] (e.g. homolog, paralog, ortholog), the word "family" to account for families of genes, and other terms to refer to variants of genes (e.g. constructs, mutants).

To account for the types of proteins and parts of proteins, we have compiled a list of these based on the UMLS Metathesaurus. We have first selected all entries from the metathesaurus

[2]Two genes are homologs when they share a common ancestor, occurring within one species or in different organisms.

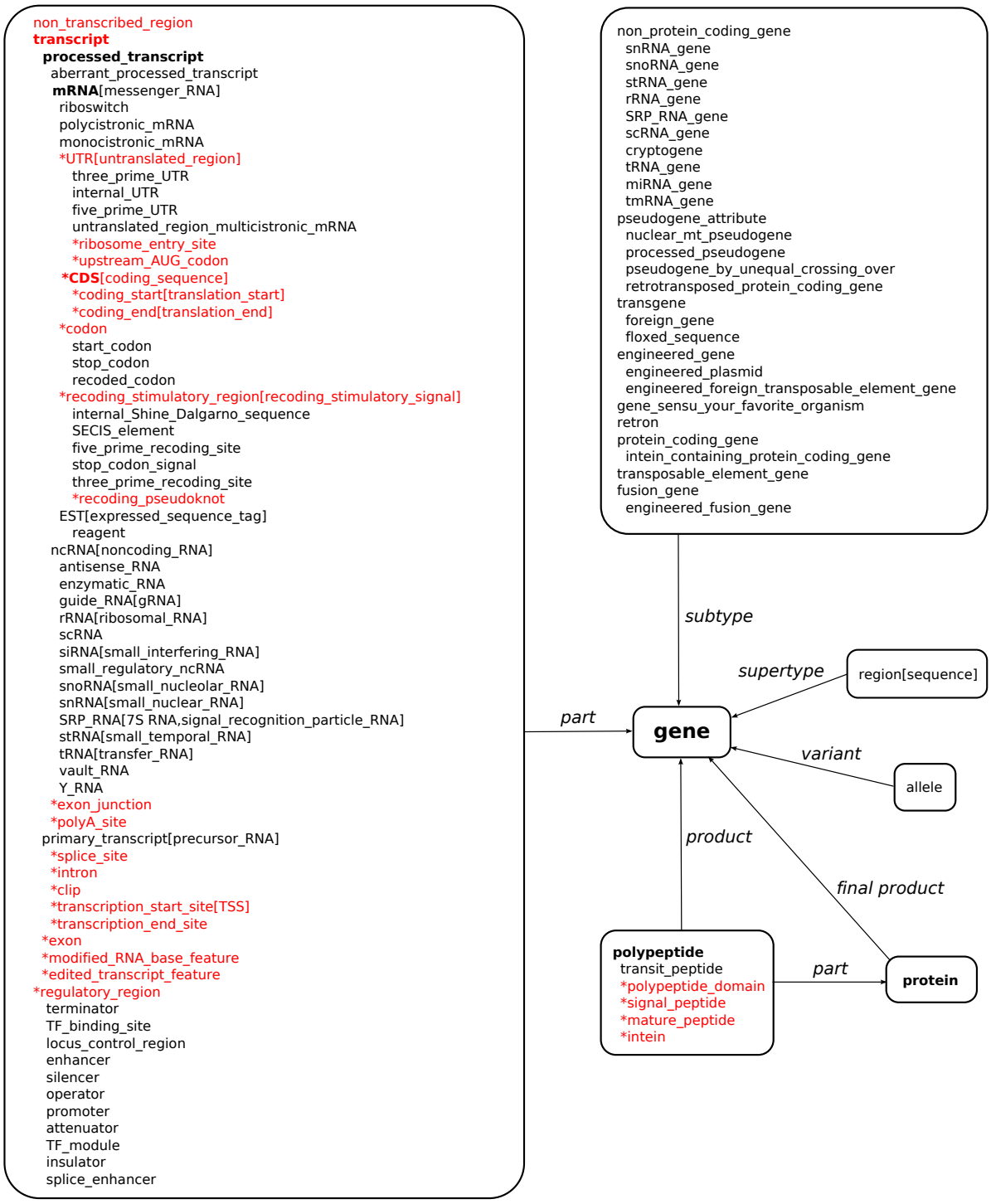
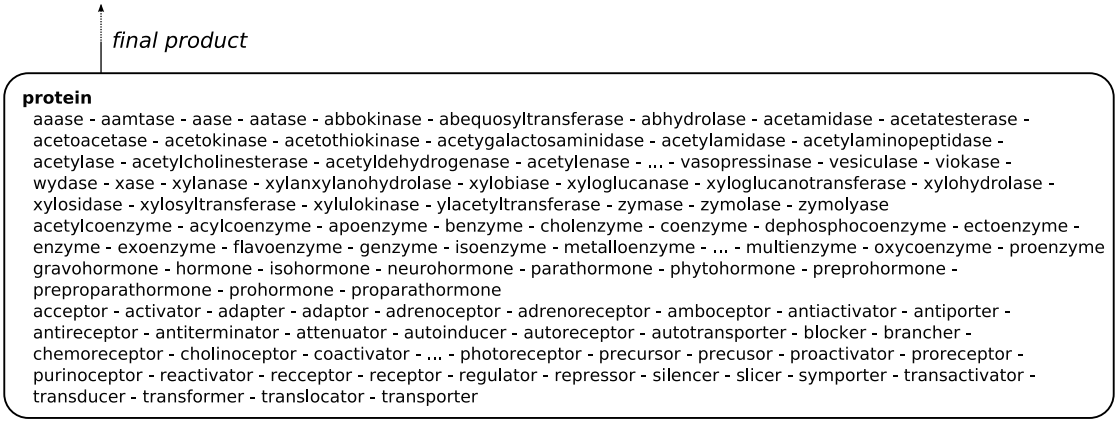
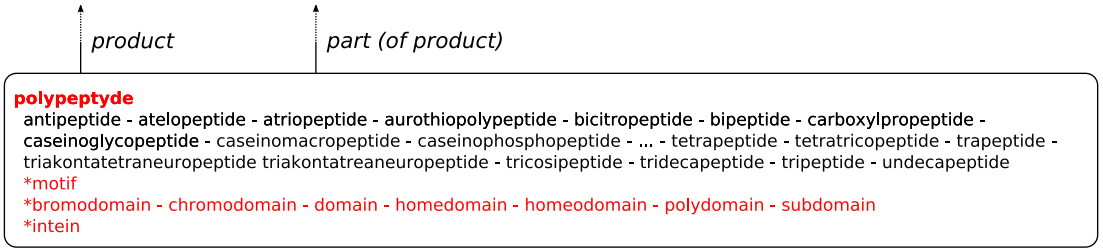


Figure 4.3: Structure derived from Sequence Ontology



(a) Extended product class



(b) Extended part-of-product class

(c) Extended variant class

(d) Extended gene class.

Figure 4.4: Additions to our ontology

whose semantic type is "Amino Acid, Peptide, or Protein", and have filtered these in order to eliminate named entities, that is, names of proteins or protein families, which are also present in the Metathesaurus. To be added to the group of terms referring to gene products, we have selected all words ending in "ase", "enzyme" and "hormone", and have manually selected terms that refer to proteins according to their function, e.g. "inhibitor", "receptor". To be added to the group of terms referring to parts of gene products, we have selected all words ending in "peptyde", "motif" and "domain". This process resulted in 2348 new terms.

To the variant class in our ontology, we added the terms "construct", "mutant" and "variant" as variations of genes; we have also moved the term "transgene" from subtype to variant. These changes have been suggested by a biologist curator from FlyBase, who participated in the corpus annotation task. Figure 4.4 shows the changes to our ontology.

Finally, to type the mentions to biomedical entities in the text according to their biotypes, we match the NP head noun against the terms associated to each of our biotypes. For multi-word terms we consider only the term's head noun, which we have indicated manually. For instance, we would tag "the third exon" with "part-of" biotype.

To classify the NPs whose head noun is a gene name tagged in the NER phase as opposed to a term from the ontology, we explored what is known about naming conventions in order to disambiguate between gene and protein names: if the name is uppercase or capitalised, it is tagged as "product"; if not, it is tagged as "gene".

Other NPs that still remain without biotype information are tagged as "other-bio" if any of its head pre-modifiers was recognised by NER as a gene name. These NPs refer mainly to events, e.g. "reaper transcription" or "Ser signaling".

This biotyping process achieves an overall accuracy of 65.35% when evaluated against the manually annotated corpus described in Chapter 5. Table 4.3 shows the number of occurences in our corpus and the performance of the typing strategy for each biotype.

Biotype	Occurrences	Precision (%)	Recall (%)
gene	249	71.9	68.1
subtype	156	94.3	64.7
variant	186	97.6	22.0
product	1189	84.9	76.5
part-of	194	57.5	33.5
part-of-product	241	73.7	62.3
supertype	36	100	22.2
other-bio	444	64.6	68.3
Total	2695	79.0	64.8

Table 4.3: Performance of biotyping strategy

The main cause of low recall for supertype and part-of tagging and low precision for part-of-product tagging is the word "sequence": it can be tagged as any of the three classes (referring to DNA sequences or protein sequences), so we opted for the class most frequently associated with the word in our annotated corpus, which is part-of-product. The recall of part-of tagging and precision of part-of-product tagging are also affected by the word "terminus" (which can also refer to a terminal part of a gene sequence or to an amino terminus, terminal part of a protein sequence), so we also adopted the part-of-product as the only class due to its higher frequency. The main source of mistakes concerning the variant and other-bio classes is the term "mutant": we have assumed it always to refer to an other-bio entity, the organism that carries a mutant gene, given the term's higher frequency with this meaning in our corpus.

The biotyped NPs are finally selected and considered for anaphora resolution. The biotype information is combined with other features to decide on an anaphora relation between two NPs, but basically it can be interpreted as follows: NPs with the same biotype may be coreferent; however, the anaphoric relation between NPs with different biotypes may be associative rather than coreferential. These assumptions are explored in our baseline system for anaphora resolution presented in Chapter 6, and in our probabilistic system presented in Chapter 7.

4.3 Limitations

The main limitation of our biotyping strategy is the lack of a disambiguation mechanism to be used when a word can be tagged with more that one biotype. To solve this problem, the context of the words would have to be analysed and used for disambiguation. For example, considering the NP "DNA sequence", the word "DNA" could be used to identify sequences that fit the part-of or supertype biotypes, while in "protein sequence" the word "protein" could be used to indicate part-of-product biotype. Further study would be necessary to identify words that are able to distinguish the senses of ambiguous words, and determine whether such words should be part of the NP or in the adjacent context.

Another limitation is the vocabulary coverage: the extensions we have made to the Sequence Ontology seem adequate and sufficient for our corpus, but future work with different scientific articles may reveal a need for further extension of the ontology.

The other-bio biotype could be refined if there is interest in identifying and classifying events related to the biomedical entities present in the text. The Gene Ontology, for example, could be used to identify which of the NPs classified as other-bio refer to molecular function or biological processes.

4.4 Related work

Castaño *et al.* [2002] makes use of the UMLS Semantic Network concepts to type the entities found in the text. Their corpus is composed of abstracts of no specific biomedical subdomain, so the types used are much more coarse-grained (in terms of biological entities) than those we used in our ontology, which is focused on the molecular biology subdomain. The types used by Castaño *et al.* are: "Amino acid, peptide or protein", "Embryonic structure", "Cell", "Bio-active substance", "Organism", "Functional chemical", "Bacterium", "Molecular Sequence", "Chemical", "Nucleotide", "Cell component", "Enzyme", "Gene or Genome", "Structural chemical", "Nucleotide sequence", "Substance", "Organic Chemical", "Pharmacologic substance", "Organism attribute", "Nucleic acid", and "Nucleotide". They consider the type matching as part of their anaphora resolution algorithm: they use a salience-based approach, where entity pairs with matching types are rewarded by 2 points; and no-matching pairs are punished by 1 point—this setting discourages the discovery of associative anaphoric relations between entities of different type.

Gaizauskas *et al.* [2000] have created their own set of semantic classes used to classify the terms in the text. They identify the terms by morphological clues (e.g. words ending in 'ase' refer to proteins) and by consulting a lexicon that they have built based on publicly available databases and corpora. They classify the terms according to the subdomain from which they want to extract information. For their PASTA system, which aims to extract information about the role of amino acid residues in proteins, they classify the terms as "atom", "base", "chain", "interaction", "protein", "non-protein compound", "region", "residue", "quarternary structure", "secondary structure", "supersecondary structure", "site" and "species". On the other hand, for their EMPathIE system, which aims to extract information about enzyme and metabolic pathways, the classes are "compound", "element", "enzyme", "location", "measure", "organization", "pathway", "person" and "organism".

4.5 Summary

In this chapter we have described our strategy for identifying and classifying the mentions of biomedical entities in a text. To identify gene/protein names we have adopted the Vlachos *et al.* named-entity recognition system. To identify NPs referring to biomedical entities of interest we have adopted the Sequence Ontology as our main source of terminology, and have enriched it by using parts of UMLS Metathesaurus. We have used the RASP parser to identify the NP boundaries and its subconstituents. We have used the relations present in the Se-

quence Ontology to classify the mentions according to 7 classes (biotypes): “gene”, “product”, “subtype”, “part-of-gene”, “part-of-product”, “supertype”, and “variant”. Once the biomedical entities were identified and classified we could then annotate the anaphoric relations between them. The annotation process is described in the next chapter.

Chapter 5

Anaphora annotation in biomedical texts[1]

In order to be able to train, test and evaluate our anaphora resolution system for the biomedical domain, it was necessary to have a gold-standard corpus, which should contain anaphora relations between biomedical entities. However, there was no corpus of full-text biomedical articles annotated with anaphoric links [Cohen *et al.*, 2005]. The lack of such data significantly impedes scientific progress in this area. For instance, although anaphora resolution was identified as one of the "new frontiers" in biomedical text mining in the call for papers of a recent conference, there were no papers on this topic published in the proceedings; the organisers attribute this to the lack of publicly available data [Zweigenbaum *et al.*, 2007]. We aimed at filling this gap by developing annotations that made our research possible and would facilitate future research on anaphora resolution in the biomedical domain.

Work has been done on annotating abstracts of research papers from Medline instead of full papers [Kulick *et al.*, 2004, Yang *et al.*, 2004, Castaño *et al.*, 2002]. However, as anaphora is a phenomenon that develops through the text, we believe that short abstracts are not the best source to study it and decided to concentrate on full papers instead. Sanchez *et al.* [2006] annotated full papers but were only interested in pronoun coreference and their data contain 18 pronouns only.

Annotating anaphora is a difficult task, given that the relation between two expressions can sometimes be subjective and subtle, and different annotators may disagree about it. It is not easy to explain precisely to an annotator the complex relation between expressions that he/she should be looking for, or to establish an exact procedure to be followed, so annotation guidelines usually employ several examples to describe the relations and impose a set of restrictions to make the task more consistent [Hirshman and Chinchor, 1997, Poesio, 2000, ACE, 2004]. Restrictions can include, for example, instructions to (do not) link expressions that take part in a particular syntactic relation (e.g. apposition), and to mark the closest antecedent. Guidelines vary in how they approach specific cases; van Deemter and Kibble [2000] discuss some aspects of the MUC guidelines [Hirshman and Chinchor, 1997] which according to them damage the quality and consistency of the annotation. Section 5.1.1 describes the main differences between some existing guidelines for anaphora annotation.

We have annotated both coreferent and associative anaphoric relations. As mentioned in Chapter 3, distinguishing coreference from anaphora, in particular coreferent anaphora (anaphora cases where the NPs involved are coreferent), can in some cases be difficult. While coreference simply describes expressions that refer to the same entity, coreferent anaphora consists in the linguistic dependency between two coreferent expressions, where the one which comes later in the text, the anaphor, depends on the earlier one, the antecedent, for it to be understood. Given that in some cases the distinguishing dependency is very subtle, we decided to consider both coreference and coreferent anaphora as one single class of relations.

Concerning associative anaphora, the association between the anaphor and the antecedent

[1]Part of the work presented in this chapter has been published in [Gasperin *et al.*, 2007].

may be due to diverse relations between the entities they refer to. These relations may, for example, be part-of or set-member relations, but also less well-defined relations, such as the relation between "the horses" and "the race" in the sentence `I watched the race, the horses were impressive`. In biomedical texts, the domain relations between the entities usually support the associative anaphoric relations between the expressions which refer to them. We took that into account and defined three types of associative relations that should be considered for the annotation. Limiting the types of relations to be considered makes the annotation more consistent, since unspecified relations can turn out to be too subjective and controversial. Section 5.1.2 details both coreferent and associative relations that we focused on.

In summary, we have developed (1) an anaphora annotation scheme tuned to the biomedical domain, which integrates linguistic and domain-specific knowledge, and (2) a corpus of full-text biomedical articles that has been annotated conforming to the proposed scheme. The resulting corpus is described in Section 5.3.

5.1 Anaphora annotation scheme

We consider as possible anaphoric expressions of interest all types of non-pronominal NPs referring to biomedical entities (which have a biotype assigned to them). We classify the NPs as: proper names (pn), definite NPs (defnp; e.g. "the gene"), demonstrative NPs (demnp; e.g. "this gene"), indefinite NPs (indefnp; "a protein"), quantified NPs (quantnp; e.g. "all genes", "four proteins"), and other NPs (np). We only annotate anaphoric relations where the antecedents are NPs; that is, we do not consider cases where the anaphor may refer back to a clause, sentence or even paragraph.

We have developed guidelines to describe the anaphoric relations that should be annotated and how to identify them. In the next subsection we present some aspects of existing guidelines for anaphora annotation and subsequently we describe our annotation scheme for the biomedical domain.

5.1.1 Existing schemes for anaphora annotation

Other schemes have been developed for the annotation of anaphora. The MUC-7 guidelines [Hirshman and Chinchor, 1997] instruct annotators to mark only the coreference (identity) relation between entities and do not deal with associative links. The GNOME project guidelines [Poesio, 2004] propose the annotation of coreference and the following kinds of associative links: 'element' (when the anaphor is an element of a set of objects), 'subset' , 'poss' (when the anaphor is owned by or is part of an entity), and the inverse version of these relations. The ACE guidelines[2] also focus on the coreference relation, just adding what they call "attributive relations" that essentially link appositive and predicative phrases to the anaphor.

All such guidelines provide a brief description of the relations of interest, and define some restrictions that should be applied to the annotation. The guidelines diverge in how to deal with some particular linguistic constructions, such as apposition, predicates and relative clauses. MUC-7 guidelines recommend appositive clauses to be annotated as coreferent, while GNOME and ACE guidelines recommend the opposite. MUC-7 also recommends that predicates be annotated as coreferent, unless they are introduced by a negative or modal clause, while GNOME recommends no relation should be annotated in these cases, and ACE recommends the use of attributive relations.

The guidelines also instruct the annotator to look for the closest antecedent. GNOME guidelines recommend the annotators mark at most one identity and one associative relation per anaphor.

In the biomedical domain, Castaño *et al.* [2002] present the Medstract corpus, where they annotated coreferent and set-member relations between biomedical entities in a set of Medline

[2]http://projects.ldc.upenn.edu/ace/annotation/2005Tasks.html

abstracts. They annotate pronominal and nominal (which they call "sortal") coreference cases. Sortal cases are phrases which refer to more than one entity, for example, "both enzymes", for which multiple antecedents are annotated, and the relation between them and the anaphor can be seen as 'set-member'.

The MedCo project [Yang *et al.*, 2004] used the MUC-7 scheme to annotate their data (a portion of the GENIA corpus), but have distinguished some special cases of the identity relation based on linguistic features: 'appos' (appositive relation), 'pron' (pronominal anaphora) and 'relat' (relative clause)[3].

We have opted not to distinguish the appositive relation from a usual coreference relation, and annotate main NP and apposition as coreferent. Since we are using base NPs as our annotation units, which do not include the apposition as part of the NP, we decided that was the most appropriate practice. For example, in the expression **"the remaining protein, MSL3,..."**, the annotator should link the apposition "MSL3" to the main NP "the remaining protein". The same was adopted for predicative mentions such as **"ced-4** is **a pro-apoptotic gene"**. Concerning relative clauses, as we have decided not to treat pronoun anaphora (nor, consequently, relative pronouns), we do not link them.

None of the existing annotation schemes takes into account the domain of the text when classifying their anaphoric links, and we believe that the record of which domain relation backed the anaphoric relation is an important piece of information for anaphora resolution, mainly when aiming to help automatic information extraction. We have considered this in our classification of associative relations between biomedical entities, as described in the next section.

5.1.2 A domain-relevant annotation scheme

Since we are interested in anaphoric relations between biomedical entities, we have focused on the domain relations between these, besides linguistics relations between their mentions, in order to classify the anaphoric relations.

We annotate the following anaphoric relations between two noun phrases:

- coreferent: when both mentions refer to the same entity, having the same biotype (e.g. two mentions of a same gene or protein, etc.)

- associative: when the mentions are related but do not refer to the same entity. We are interested in three types of associative relation:

  - biotype relation: when related mentions have different biotypes (e.g. a gene and one of its products)
  - homolog relation: when the related mentions are homologs[4], having the same biotype (e.g. a gene and its homolog from another organism)
  - set-member relation: when one of the related mentions refers to a set that contains the referent of the other mention (e.g. plural or coordinated mentions)

These anaphoric relations are detailed in the following subsections.

Since biomedical texts have a considerable amount of text placed in captions of tables and figures, we assume that biomedical NPs in such captions may have an anaphoric relation to an NP in the body of the text; however, the converse is not allowed, that is, an anaphor in the main body of the text cannot be linked to an NP in a caption.

[3]The MedCo guidelines are not publicly available, but some samples of their data can be found on their website.

[4]Two genes or gene products are homologs when they share a common ancestor, occurring within one species or in different organisms.

5.1.3 Coreferent mentions

We consider as coreferent the relation between two mentions that refer to the same biomedical entity. The annotator looks for the closest mention that is coreferent to the current mention and, if one is found, links them. In our annotation, we do not distinguish between coreferent relations that are anaphoric or not; for example, we annotate as coreferent both the expressions of Example 8 (not clearly anaphoric) and Example 9 (anaphoric).

(8) <np id="10" biotype="product">
Initiator caspases
</np>
are thought to be at the beginning of a proteolytic cascade...
<np id="15" biotype="product" ante="10" rel="coref">
Initiator caspases
</np>
usually have long pro-domains ...

(9) The expression of
<np id="20" biotype="gene">
reaper
</np>
has been shown...
<np id="25" biotype="gene" ante="20" rel="coref">
the gene
</np>
encodes ...

5.1.4 Associative mentions

Associative anaphoric relations rely on ontological i.e. "world" relations between the entities referred to in the text. These relations are assumed by the writer to be known by the reader. We annotate as associative cases those instances in which these relations imply a dependency between the anaphor and its antecedent, that is, the meaning of the anaphor could not be fully understood if it were not for its relation with the antecedent. In the biomedical domain, these world relations are the actual relations between the biomedical entities, independent of the text, for example, the fact that a gene encodes a protein, or that a gene is composed by DNA sequences.

Given that associative relations are more subtle than the identity relations present in coreferent cases, the span of associative anaphoric links is usually shorter, that is, associative antecedents are usually close to the anaphor, while coreferent antecedents may be further away. According to Hawkins' [Hawkins, 1978] definition of associative anaphora, the anaphor in such a relation should be an entity not previously mentioned in the discourse, which is introduced based on its relation with a previously mentioned entity. However, we noticed that in long discourses like scientific papers, entities are introduced more than once, usually in different sections of the paper. With this in view, the annotator is encouraged to look for associative antecedents mainly within the same section of the paper as the anaphor.

Here we describe the main types of associative relations that we found in our corpus of

biomedical articles.

5.1.4.1 Biotype relation

The associative relation between two entities with different biotypes, as in examples 33 and 34, is marked as 'biotype' associative relation. The biotype relation may represent, for example, the link between a gene and its product, or between a gene and a DNA sequence that is part of it.

(10)
```
There was considerable excitement in the field when potential mammalian and
Drosophila homologs for
<np id="20" biotype="gene">
  ced-3
</np>
were discovered.
<np id="25" biotype="product" ante="20" rel="biotype">
  The CED-3 protein
</np>
is one of ...
```

(11)
```
...the role of
<np id="30" biotype="gene">
  the roX genes
</np>
in this process...interact with
<np id="35" biotype="partof" ante="30" rel="biotype">
  the roX RNAs
</np>
```

If we take into account the specific biotype of the entities that are involved in the 'biotype' relation, it is possible to determine a WordNet-like semantic relation behind the anaphora relation. For example, a biotype relation between a 'gene' and a 'subtype' of gene may be considered an hyponymy relation, the relation between a 'gene' and a transcript (biotype 'part-of') can be seen as a meronymy relation.

5.1.4.2 Homolog relation

Another type of associative relation is the homolog. In this case, the related entities have the same biotype but refer to entities in different organisms; see Example 38, where the gene named Bok is referred to as its instance in mammals and its instance in Drosophila flies.

(12) ...is most closely related to
<np id="40" biotype="gene">
mammalian Bok
</np>
.
<np id="45" biotype="gene" ante="40" rel="homolog">
The Drosophila Bok homolog
</np>
...

The homolog relation is quite interesting, with no obvious counterpart in other domains. Normally, any property that is assigned to a gene is also assigned to its homolog, so in the same paper the author can alternately talk about one or the other, since these are "equivalent", yet not identical, entities in different organisms. Homolog mentions are usually surrounded by species names, such as "mammalian" and "Drosophila".

However, homolog relations are often less obvious than in Example 38, and very much resemble a coreference relation, as shown in Example 39.

(13) ''...searches of the sea urchin sequences against all GenBank proteins detected only **the ring finger domain** of the sea urchin sequences. Based on the same approach, our study found that the starlet sea anemone and hydra genomes also encode several families of the N-terminal RAG1 domain. The only exception was the already mentioned sea anemone RAG1 core-like sequence. The approximately 90-aa N-terminus of the latter sequence is **the ring finger**.''

The ring finger domain is a specific piece of protein sequence. In this example, its first mention refers to the domain of a protein found in sea urchins, while the second mention refers to an homolog instance in sea anemones and hydras.

5.1.4.3 Set-member relation

The third type of associative relation, common to other domains as well, we call the set-member relation, which occurs when an entity is related to a set of which it is a part of, or vice-versa. The single entity and the entities in the set have the same biotype. It occurs mostly in the presence of noun phrases referring to coordinated NPs, plural NPs, and families of bio-entities. Below we describe situations in which set-member relations occur.

Coordination

It is common to find in a text mentions such as the genes reaper, hid, and grim. These mentions, which contain coordination, can have multiple antecedents. When this is the case, the relation between the mentions is marked as associative of the type 'set-member'.

(14) ...

```
<np id="50" biotype="gene">
  reaper
</np>,
<np id="51" biotype="gene">
  hid
</np>, and
<np id="52" biotype="gene">
  grim
</np>
are regulators of apoptosis...
<np id="55" biotype="gene" ante="50,51,52" rel="set-member">
  the genes reaper, hid, and grim
</np>
```

The same is true for the opposite case, when a simple mention refers to a coordinated one.

List

When a set of entities is mentioned and its mention is followed by a list of its members, as in an apposition construction, the members should be linked to the set by a 'set-member' relation. Members can be listed between commas, as in Example 41 or in brackets, as in Example 42.

(15) ...

```
<np id="40" biotype="product">
  two proteins
</np>
encoded by the recombination-activating genes,
<np id="41" biotype="product" ante="40" rel="set-member">
  approximately 1040-aa RAG1
</np>
and
<np id="42" biotype="product" ante="40" rel="set-member">
  approximately 530-aa RAG2
</np>
,...
```

(16) ...
<np id="50" biotype="product">
surface receptors
</np>
of vertebrate B and T immune cells (
<np id="51" biotype="product" ante="50" rel="set-member">
BCRs
</np> and
<np id="52" biotype="product" ante="50" rel="set-member">
TCRs
</np>
).

Plural

Plural mentions are treated in the same way as coordinated mentions, as they may also have multiple antecedents and be the antecedent of multiple mentions, as shown in Example 43.

(17) ...
<np id="60" biotype="gene">
ced-4
</np>
and
<np id="61" biotype="gene">
ced-9
</np> ...
<np id="65" biotype="gene" ante="60,61" rel="set-member">
the genes
</np> ...

Family

In the biomedical domain, an entity mention may be related to a mention of its family, and we consider this a case of set-member associative relation.

(18) ...
<np id="70" biotype="product">
the mammalian anti-apoptotic protein Bcl-2
</np> ...
<np id="75" biotype="product" ante="70" rel="set-member">
Bcl-2 family
</np> ...

(19) ...
<np id="80" biotype="product">
the MSLs
</np> ...
<np id="85" biotype="product" ante="80" rel="set-member">
MSL-1
</np> ...

Subset

We also consider 'set-member' relation that between a set and a subset of it, as in the example below.

(20) <np id="90" biotype="otherbio">
D-mib mutant discs
</np>
have no wing pouch ... The complete loss of D-mib activity in
<np id="92" biotype="otherbio" ante="90" rel="set-member">
D-mib1 mutant discs
</np>
...

Other

This is a special case of set-member relations, which includes mentions that contain the word 'other' (or similar words, like 'remaining'), as in Example 47.

(21) ... distribution in females ectopically expressing
<np id="5" biotype="product">
MSL2
</np>
but lacking
<np id="6" biotype="product" ante="5" rel="set-member">
other MSL proteins
</np>
.

In these cases, the 'other' mentions should be linked to their complements, that it, their antecedents are the mentions referring to the item excluded from the set.

5.1.4.4 Mixed relations

There are cases where the type of relation between two mentions is mixed, that is, it could be interpreted as a combination of the above types of associative relation. In Example 50, the relation between mentions 12 and 10 can be seen as biotype (gene-otherbio relation) and set-member.

(22)
```
While
<np id="10" biotype="gene">
  the neur and mib genes
</np>
are evolutionarily conserved, ...  events requiring
<np id="12" biotype="otherbio">
  neur activity
</np>
.
```

In such cases, the annotator should select the type of relation that he/she finds to be more prominent.

5.1.5 Other relations

GNOME guidelines include possessive relations as a class of anaphoric relations, for example, in the expression "**ingredients** of **the cream**", "the cream" is linked to "ingredients" by a possessive relation, or in the expression "your cream", "cream" is linked to "your".

We do not consider these relations anaphoric, because the relevant semantic relations are determined syntactically. However, we decided to annotate of-phrases like the one in the first example and mark them as possessive relations. We did not annotate cases like the second example, since our minimal annotation unit is an NP (we do not link separated constituents of an NP). Examples 23 and 24 present cases of possessive relations in our corpus.

(23)
```
...
<np id="50" biotype="partof-product">
  the approximately 600-amino acid core region
</np>
of
<np id="51" biotype="product" ante="50" rel="poss">
  RAG1
</np>
...
```

(24)
```
...
<np id="60" biotype="subtype">
  11 additional new families
</np>
of
<np id="61" biotype="subtype" ante="50" rel="poss">
  Transib transposons
</np>
...
```

5.2 Corpus annotation

We selected five biomedical papers[5] to be hand-annotated with anaphoric and coreferent links. The selected papers were chosen according to the following criteria: they were part of relevant journals in the biomedical field, were freely available on the internet, and focused on fruit fly genomics. We assumed that 5 papers would be the minimum corpus size for it to be useful for training a corpus-based anaphora resolution system. Given the difficulty of the task and time constraints, we have not annotated more papers.

Before starting the manual annotation process, we preprocessed the corpus automatically, following the steps presented in the previous chapter. First we applied the gene name recogniser described in [Vlachos *et al.*, 2006] to recognise gene names; secondly we identified the noun phrase boundaries and sub-constituents using the RASP parser [Briscoe and Carroll, 2002], and lastly we tagged all noun phrases with their biotypes according to the Sequence Ontology. We filtered out all noun phrases for which we could not define a biotype, keeping only those that referred to biomedical entities.

We then asked two annotators (a domain expert and a linguist) to review and correct the automatically defined biotypes, gene names and noun phrase boundaries. Finally the same two annotators were asked to insert the coreferent links, and I, a third annotator (computer scientist), and the domain expert annotator inserted the associative links[6]. We used the MMAX annotation tool [Müller and Strube, 2001]. The annotation task was divided into four phases to minimise the number of decisions that the annotator had to make at a time; these phases were:

1. Annotating noun phrases that contain a gene name: in this phase, the annotators were asked, for each sentence: (1) to look at the noun phrases that had been automatically tagged in the preprocessing phase, check if they were correct—if they did indeed contain a gene name, if the NP boundaries were precise, and if the assigned biotype was correct—and correct it (which might mean deleting it in the case it contains a mistakenly recognised gene name); and (2) to check if any noun phrase containing a gene name was missed by the preprocessing, and annotate it, assigning the appropriate biotype.

2. Annotating noun phrases that refer to an entity of interest but do not contain a gene name: in the same way as in the previous phase, in this phase the annotators are asked to correct the automatically tagged and include missed noun phrases that do not contain a gene name but which refer to an entity of interest (e.g. "the X-linked genes", "this protein"). We decided to separate phases 1 and 2 because phase 2 requires more attention than phase 1, where the gene names are quite obvious and facilitate the task. When this phase is finished, all entities of interest should have been tagged and have a biotype assigned to them.

3. Coreference linking: in this phase the annotators should create the links between the noun phrases (tagged in the previous phases) that are coreferent. The MMAX tool provides a mechanism for grouping the noun phrases in sets, which can be seen as coreference chains. No new noun phrases should be added in this phase.

4. Associative linking: in this phase the annotators should create the associative links between noun phrases. The annotators should look for the closest antecedent and the type of the relation should be indicated. MMAX has a pointing mechanism which links the anaphor to the annotated antecedent. No new noun phrases should be added in this phase.

[5]The FlyBase identifiers for these papers are: FBrf0132215, FBrf0134664, FBrf0184230, FBrf0188209, FBrf0188423.

[6]Due to time constraints, the domain expert annotated associative links in only two of the selected papers, FBrf0132215 and FBrf0188423.

Our annotation guidelines for phases 3 and 4 can be seen in Appendix A.

The annotation provided by the two annotators for phases 1 and 2 was automatically compared, and their discrepancies were discussed and harmonised. The annotation was compared at all decision levels taken by the annotator: mention selection and its boundaries (looking at mentions that one annotator had selected but not the other, or cases in which the mention boundaries differ), and biotype (checking if a difference on biotyping was conscious).

Besides helping us to find false disagreements and mistakes made by the annotators, these comparisons generated very fruitful discussions that enabled us to refine our guidelines. For example, we could observe the need to expand the Sequence Ontology: we decided to add some new entries to it, in order to be able to optimise our automated detection of relevant noun phrases. We were able to identify classes of words that were missing from the Sequence Ontology, for instance, different types of proteins, like `caspase`, `kinase`, `enzyme`. We obtained a set of these words from the UMLS Metathesaurus to complement SO, as described in the previous chapter.

After correction of the mistakes found through the comparison of both annotations for phase 1 and 2, we compared the annotation of coreferent links. We calculated the Kappa agreement coefficient for the annotation of the coreferent links; the first column of Table 5.1 presents the results. Kappa scores above 0.8 are considered a good level of agreement [Carletta, 1996]. Most true disagreements were due to the non-expert annotator's lack of domain knowledge and understanding of what is biologically relevant. In order to harmonise disagreement cases, we have also compared and discussed the annotation of the coreferent links. This process was able to identify some inconsistencies in the annotation (e.g. one annotator might have chosen a coreferent mention as antecedent but not the closest one, as indicated in the annotation guidelines). After this comparison and consequent revision of the annotation, we reached the Kappa scores presented in the second column of Table 5.1. A gold standard annotation was developed based on the domain expert results, and the annotation of the associative links was performed on top of it.

Annotating associative anaphora is known to have higher disagreement rates than annotating coreference [Vieira, 1998]. Only two of the papers were annotated with associative links by two annotators (computer scientist and domain expert), so these were used to compute the inter-annotator agreement for associative cases. Table 5.1 presents the Kappa scores for biotype, homolog and set-member cases for the two papers that were annotated by more than one annotator. We have revised this annotation correcting cases in which one annotator or the other had not chosen the closest antecedent (but instead a more distant mention to an equivalent entity). The Kappa scores for the revised annotation are also shown in Table 5.1.

	Coreferent		Biotype		Homolog		Set-Member	
	O	R	O	R	O	R	O	R
Paper 1	0.82	0.84	0.62	0.81	0.51	0.67	0.56	0.56
Paper 2	0.81	0.84	-	-	-	-	-	-
Paper 3	0.94	0.98	-	-	-	-	-	-
Paper 4	0.87	0.97	-	-	-	-	-	-
Paper 5	0.80	0.93	0.49	0.52	0.60	0.60	0.61	0.62

Table 5.1: Kappa scores for each paper per anaphoric class. (O) corresponds to the original, (R) to the revised annotations.

The low rates of agreement on associative cases reflect the difficulty of the task. Most cases

of disagreement on associative cases are related to one of the following issues:

- Mixed relations: cases where the antecedent does not fit a single associative relation, but more than one at the same time. In the example below, mention (d) appears to have both a biotype and a set-member relation with (b) and (c).

  ``The antigens can be identified after they are specifically bound by **surface receptors**(a) of vertebrate B and T immune cells (BCRs and TCRs, respectively). Because the vast repertoire of **BCRs**(b) and **TCRs**(c) cannot be encoded genetically, ancestors of jawed vertebrates adopted an elegant combinatorial solution. The variable portions of **the BCR and TCR genes**(d) are composed of ...''

  In such cases, annotators were instructed to choose the most prominent relation and annotate it. In this example, one of the annotators chose to create a biotype relation between (d) and one of the previous mentions (c, the closest). The other annotator felt compelled to find an antecedent that fit perfectly one relation or the other, and has chosen the mention of "surface receptors"(a) in the first sentence of the example as biotype antecedent of (d).

- Syntactic relations: the annotator may be misled by syntactic relations into annotating anaphoric relations between syntactically related NPs. For instance, one of the annotators chose to annotate a biotype relation between mentions (a) and (b) in the examples below:

  ``...families are represented by **transposons**(a) flanked by **TIRs**(b) ...''

  ``...a part of **a motif**(a) that is conserved in **the Transib TPases**(b) ...''

- Recent coreferent relations: the annotation guidelines explain that it is unlikely that an associative relation between two mentions exists when the current mention refers to an entity that has recently been mentioned in the text. This is because an entity that is salient in the readers mind does not need an indirect (associative) relation to introduce it. However, there were cases in which one of the annotators found that there was room for an associative relation while the other did not. That was the case in the following example, where one of the annotators linked (c) and (b) in a biotype relation, and the other did not (given the presence of (a) as a coreferent mention).

  ``The approximately 600-amino acid core region of RAG1 is significantly similar to **the transposase**(a) encoded by DNA transposons that belong to the Transib superfamily. (...) **Transib transposons**(b) also are present in the genomes of sea urchin, yellow fever mosquito, silkworm, dog hookworm, hydra, and soybean rust. (...) Furthermore, the critical DDE catalytic triad of RAG1 is shared with **the Transib transposase**(c) as part of conserved motifs.''

These sources of disagreement could be reduced by refining our guidelines and specifying more objectively which procedure to be adopted in each situation. Although this can greatly contribute to consistency in the annotation, it can also undermine the annotators natural reasoning when resolving anaphora. Due to time constraints we could not rerun the annotation with improved guidelines, and have opted to run our experiments on the current data. We have used the annotation provided by the computer scientist annotator for our anaphora resolution experiments since it contains annotations for all five papers.

5.3 The resulting corpus

Following the annotation process described above, we created our corpus. For the five papers that we have annotated, we obtained a total of 2720 noun phrases of interest. Table 5.2 shows the distribution of the NPs according to the biotypes.

	gene	subtype	variant	supertype	partof	partof-product	product	otherbio
Paper 1	59	0	0	11	4	17	216	8
Paper 2	30	0	1	0	2	17	138	10
Paper 3	99	8	54	0	9	21	345	244
Paper 4	41	1	131	2	25	3	135	146
Paper 5	20	147	0	23	154	183	355	36
Total	249	156	186	36	194	241	1189	444

Table 5.2: Biotype distribution

We can see some variation in the distribution of each biotype across the papers, based on the subject of each paper. For example, Paper 5 discusses the similarity between proteins based on the comparison of parts of the protein sequence, and so the higher number of partof-product NPs in comparison to other papers. Paper 4 discusses several mutants of a particular gene, so the high number of variant NPs.

Table 5.3 shows the distribution of the different types of NPs among the different anaphoric classes.

Class/NPs	pn	defnp	demnp	indefnp	quantnp	other np	Total
coreferent	681	416	59	36	52	380	1624
biotype	46	105	3	8	4	119	285
homolog	7	8	0	3	0	6	24
set-member	152	125	26	14	68	158	543
poss	29	12	3	6	5	38	93
discourse new	67	115	0	76	41	169	468
Total*	878	689	74	135	149	771	2696 NPs \ 3037 relations

Table 5.3: Anaphoric class distribution according to NP form. *Last row 'Total' does not correspond to the sum of the values of the previous rows: it shows the total number of NPs of a type, which can have more that one anaphoric relation annotated.

We can see that around 80% of the definite NPs are anaphoric in our corpus, compared to the 50% presented in [Poesio and Vieira, 1998] for newspaper texts. Concerning demonstrative NPs, all of them are anaphoric. We can also observe that more than 75% of the proper names take part in coreference relations, as it is in their nature to refer to a specific named entity, but still 6% of them take part in biotype or homolog relations, due to the fact that a gene, its homologs, and the protein it synthesizes usually share the same name. 44% of quantified NPs take part in set-member relations, as they usually refer to more than one entity. 56% of indefinite NPs are discourse new.

Table 5.4 shows the distribution of anaphoric relations according to the distance between anaphor and antecedent in our corpus. The majority of coreferent relations occur between NPs in different sections of the paper, while the majority of associative relations occur between NPs in adjacent sentences. We can see that very few biotype relations cross section boundaries, and that the majority of set-member relations occur within the same sentence (most likely due to the List cases described in Section).

Class/Distance	Same sentence	Previous sentence	Same paragraph	Previous paragraph	Same section	Other sections
coreferent	195	323	295	220	181	410
biotype	49	100	59	36	35	6
homolog	6	10	2	3	1	2
set-member	162	130	103	64	48	36

Table 5.4: Distance between anaphor and antecedent according to anaphoric relation

We can form coreference chains by following the coreferent links between noun phrases in the corpus, so that all noun phrases in the text that refer to the same entity are part of the same coreference chain. The more noun phrases in a chain, the longer it is; Figure 5.1 shows the number of chains of different size in our corpus. We have in total 357 chains with at least two elements (and 715 single noun phrases that are not part of any chain). Our longest chain is composed by 68 noun phrases, and the average chain size is 5.5.

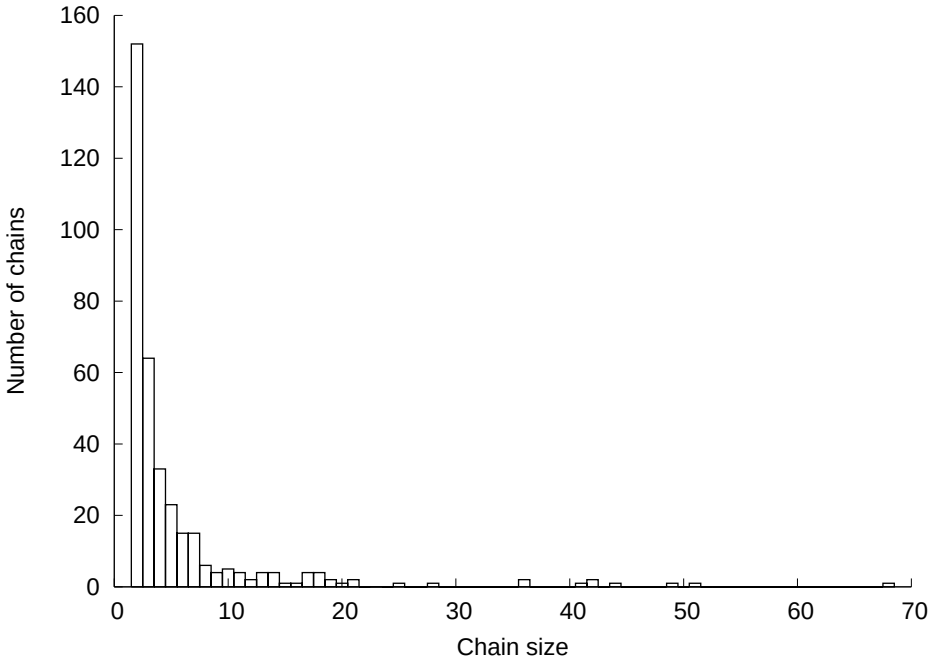


Figure 5.1: Number of coreference chains by chain size

The corpus and the annotation guidelines are available to the scientific community via the FlySlip project website http://www.wiki.cl.cam.ac.uk/rowiki/NaturalLanguage/FlySlip .

5.4 Summary

This chapter presents a scheme for annotating coreferent and associative anaphoric relations in biomedical papers. Our scheme takes into account the domain of the text, classifying the anaphoric relations according to the domain relation that supported the linguistic relation. Upon our annotation scheme, we have built a corpus of five scientific full-text articles that, according to our best knowledge, is the first corpus of biomedical articles with anaphora information not to be built from paper abstracts.

We use this corpus as evaluation data for the baseline anaphora resolution system presented in Chapter 6 and as training and evaluation data for the probabilistic anaphora resolution system presented in Chapter 7.

Chapter 6

Rule-based baseline system[1]

We have developed a knowledge-based baseline anaphora resolution system for the biomedical domain. The system identifies coreferential relations between biotyped entities as well as associative links. We have created a small set of rules to identify the antecedents of NPs of interest in the text. The rules aim to encode the well-defined characteristics of the coreferent and associative relations. We have created rules only for biotype and set-member types of associative relations, since there is no clear pattern for homolog relations.

The system does not require training. It makes use of lexical, syntactic, semantic and positional information to link anaphoric expressions. The lexical information consists of the words themselves, as well as the number, singular or plural, of each noun phrase. The syntactic information consists of noun phrase boundaries and the distinction between head and pre-modifiers extracted using RASP (as described in Chapter 4). The semantic information comes from the gene-name recognition and biotype tagging processes (also described in Chapter 4). The distance between the anaphoric expression and its possible antecedent is taken into account as positional information. The system assumes as discourse new NPs for which it could not find any antecedent.

The next section describes how we use the available information to resolve anaphora.

6.1 Resolving anaphora cases

We take all biotyped NPs as potential anaphors to be resolved. As potential antecedents for an anaphor we take all biotyped NPs that occur before it in the text. For each anaphor we look for its closest antecedent. For linking anaphors to their antecedents we consider the features presented in Table 6.1.

The algorithm to find the antecedent for each anaphor is given in Figure 6.1. Our matching among heads and modifiers is case-insensitive, allowing, for example, "msl gene" to be related to "MSL protein" given their common modifiers. Head nouns and modifiers are lemmatized, so the words "protein" and "proteins" would match (however they disagree in number).

Coref$_i$, if found, is considered coreferent to A$_i$, and Assoc$_i$, associative. For example, in the passage:

(25) ``Dosage compensation, which ensures that the expression of **X-linked genes**:C_j is equal in males and females ... the hypertranscription of **the X-chromosomal genes**:A_i in males ...''

C_j is taken to be coreferential with the anaphor indexed as A_j. Additionally, in:

(26) ``... the role of **the roX genes**:C_n in this process ... which MSL proteins interact with **the roX RNAs**:A_m ...''

[1]Part of the work presented in this chapter has been published in [Gasperin, 2006].

Feature	Description
$head_{an}$	anaphor head noun
$head_{a}$	antecedent head noun
mod_{an}	set of anaphor pre-modifiers[2]
mod_{a}	set of antecedent pre-modifiers
num_{an}	anaphor number
num_{a}	antecedent number
$biotype_{an}$	anaphor biotype
$biotype_{a}$	antecedent biotype
d	distance from the anaphor

Table 6.1: Features used by the baseline system

- Input: a set A with all anaphors; a set C with all antecedent candidates.
- Consider Coref_i as coreferent antecedent of A_i; Assoc_i as associative antecedent of A_i; Assoc-Biotype_i as biotype antecedent of A_i; $\text{Assoc-Set-Member}_i$ as set-member antecedent of A_i;
- For each anaphor A_i:
  - Let Coref_i be the closest preceding NP C_j such that
    $\text{head}(\text{C}_j)=\text{head}(\text{A}_i)$ and
    $\text{num}(\text{C}_j)=\text{num}(\text{A}_i)$ and
    $\text{biotype}(\text{C}_j)=\text{biotype}(\text{A}_i)$
  - Let Assoc-Biotype_i be the closest preceding NP C_j such that
    $\text{head}(\text{C}_j)=\text{head}(\text{A}_i)$ or
    $\text{head}(\text{C}_j)=\text{mod}(\text{A}_i)$ or
    $\text{mod}(\text{C}_j)=\text{head}(\text{A}_i)$ or
    $\text{mod}(\text{C}_j)=\text{mod}(\text{A}_i)$ but
    $\text{biotype}(\text{C}_j)\neq \text{biotype}(\text{A}_i)$
  - Let $\text{Assoc-Set-Member}_i$ be the closest preceding NP C_j such that
    $\text{head}(\text{C}_j)=\text{head}(\text{A}_i)$ and
    $\text{biotype}(\text{C}_j)=\text{biotype}(\text{A}_i)$ but
    $\text{num}(\text{C}_j)\neq\text{num}(\text{A}_i)$
  - Let Assoc_i be the closest between Assoc-Biotype_i and $\text{Assoc-Set-Member}_i$
  - If Coref_i is closer to A_i than Assoc_i, Assoc_i is ignored.
  - If Coref_i nor Assoc_i are found, A_i is assumed to be discourse new.
- Output: a set of (Coref_i,Assoc_i-A_i) relations.

Figure 6.1: Rule-based algorithm for anaphora resolution

Class	perfect			relaxed		
	P	R	F	P	R	F
coreferent	47.5	56.8	51.8	61.7	61.0	61.3
assoc-biotype	22.4	18.0	20.0	23.3	18.7	20.8
assoc-set-member	36.8	2.5	4.7	36.8	2.5	4.7
discourse new	37.4	30.2	33.4	37.4	30.2	33.4

Table 6.2: Performance of the baseline system

C_n meets the conditions to form an associative link to A_m. The same is true in the following example in which there is an associative relation between C_y and A_x:

(27) ``**The genes ced-4 and ced-9:**C_y have been shown to ... **the ced-9 gene:**A_x is ...''

However, the system is not able to find the correct antecedent when there is no string (head or modifier) matching, such as in the coreferent relation between ``Dark/HAC-1/Dapaf-1'' and ``The Drosophila homolog''.

6.2 Results

We evaluated our system against the five hand-annotated full-text articles described in Chapter 5. We have achieved the precision and recall scores presented in the first column ('perfect') of Table 6.2. The 'perfect' scores consider exact match between the anaphor-antecedent pairs returned by the system and those manually annotated in the corpus. These performance scores are reached when considering hand-corrected input, that is, perfect gene name recognition, NP extraction and biotype tagging.

The performance for coreferent cases is clearly higher than for associative cases. This indicates that our rules are more accurate in identifying the former than the latter. Associative relations are known to be less straightforward than coreferent, and so more difficult to encode as rules. The recall for set-member cases is extremely low, since the system relies on head-noun matching for resolving those but the majority of set-member cases in our corpus (66%) does not have matching heads (41% do not have any string matching).

The performance scores of the system increase if we consider as correct the cases for which it is able to find an antecedent other than the closest, but which is from the same coreference chain as the closest antecedent. These are cases like the following:

(28) ``The function of Drosophila mib (**D-mib**) is not known ... we have studied the function of **the Drosophila D-mib gene.** We report here that **D-mib** appears to ...''

where the system returns the first "D-mib" as the coreferent antecedent for the last "D-mib", instead of returning "the Drosophila D-mib gene" as the closest antecedent. In order to take such cases into account, we have used the MUC scoring strategy, as presented in Section 3.4 in Chapter 3, to evaluate the resolution of the coreferent cases. Using this evaluation strategy, the baseline reaches the scores presented in the second column ('relaxed') of Table 6.2 for coreferent cases. This evaluation is possible since coreference chains can be derived from our corpus annotation. When the restriction to find the closest antecedent is relaxed, the system manages to achieve almost 10% gain in F-measure for coreferent cases.

The MUC scoring, however, does not deal with associative cases. To evaluate these in a

Class	coreferent			associative			discourse new		
	P	R	F	P	R	F	P	R	F
pn	82.4	75.6	78.8	25.0	8.4	12.5	22.1	70.3	33.7
defnp	35.3	46.8	40.2	22.3	7.1	10.8	51.0	22.2	30.9
demnp	62.7	47.0	53.7	25.0	3.4	6.0	-	-	-
indefnp	16.9	31.5	22.0	13.3	16.0	14.5	78.7	35.6	49.0
quantnp	18.0	37.0	24.2	41.1	9.5	15.5	40.0	5.2	9.3
other np	28.4	42.4	34.0	30.1	11.4	16.5	52.0	23.4	32.3

Table 6.3: Performance of the baseline system per NP form

less strict way than perfect matching to the annotation, we also considered as correct the cases for which the antecedent selected by the system is coreferent with the associative antecedent assigned to the anaphor in the manual annotation. That is, since the two NPs involved in an associative relation may each be part of a different coreference chain, the relaxed scoring for the associative relation assumes that if the anaphor has been linked by the system to another member of the correct antecedent's coreference chain, the link is correct.

Treating these cases as positive we reach the scores presented in the 'relaxed' column of Table 6.2. We can observe a slight increase in the performance scores for biotypes cases, but none for set-member.

Table 6.3 reports the 'perfect' performance of the baseline system according to each type of NP. The best performance for coreferent cases is achieved for proper names. This is because in our corpus 78% of coreferent relations where the anaphor is a proper name involve head-noun matching[3], so the system was able to resolve 96% of these. As proper names do not usually have head modifiers, head-noun matching and biotype matching cover the majority of cases. However, in our corpus 74% of definite NPs also involve head-noun matching, but in their cases this is not an indicator of coreference as precise as for proper names, since definite NPs with mismatching modifiers (e.g. "the faf gene" and "the roX gene") can refer to different entities; this is the main source of error in the resolution of coreferent same-head definite NPs, since we select the closest NP with same head-noun. The same problem arrises with demonstrative NPs, which in our corpus account for 80% of coreferent cases with head-noun matching, but only 68% of these were correctly resolved.

The performance for associative cases is very low for all types of NPs. The low recall is due to both rules for associative cases (which aim to cover "ideal" cases of biotype and set-member types of associative relations) being very restrictive, covering only 34.8% of the associative cases in our corpus, thus 34.8% would be the maximum recall that the baseline system is able to resolve. The low precision is caused by the lack of a distance measure to be used instead of selecting the closest candidate that conforms to the rules.

6.3 Limitations

The system relies heavily on string matching and will not link cases where there is no string overlapping. In our corpus in 21% of the coreference relations and in 37% of the associative relations there is no string matching (neither head noun nor modifier) between the anaphor and

[3]Those that do not are usually coreferent relations that involve apposition, such as "Only **one mammalian CED-4 homolog**, **Apaf-1**, has been ...", but also regular cases can occur, e.g "Reports of **a potential functional mammalian analog** of Reaper, Hid, and Grim have been published. Although **Diablo/Smac** shares no sequence homology with Reaper, Hid, or Grim, it too can bind IAPs.".

the antecedent, so these cases have no chance of being resolved by the baseline system. Eliminating the string-matching requirement would lead to very low precision, and using additional less-intuitive features to compensate that becomes complicated in a rule-based system, since the way to integrate the features is less clear than when combining basic intuitive features such as string matching, and number and semantic class (dis)agreement. For example, it is known that different NP types exhibit different anaphoric behaviour, but encoding NP types as part of rules is not straightforward.

Relaxing string matching in the rules would require adding other types of constraints to the rules in order to avoid the expected loss of precision. It is relatively clear how the current additional factors, biotype and number matching, contribute to the anaphoric relations being treated (as specified in our rules), but it is not as straightforward to model the expected behaviour of other factors available to us, such as NP type, distance, and syntactic clues.

Our baseline system selects as antecedent the closest candidate that fits the string, number and biotype matching criteria. However, instead of choosing the closest candidate, the system should be able to use a distance measure to rank the candidates according to distance and the other features. The closest candidate is not always the right one, and different NP types have different ranges of distance from their antecedents.

6.4 Integration with curation tool

The baseline anaphora resolution system presented here has been integrated to the tool created as part of the FlySlip project for facilitating the curation of biomedical literature. The curation process requires the identification of biomedical entities of interest present in the text and the extraction of particular information written about them in order to filling a template. FlyBase curators focus on extracting information related to genes and alleles mentioned in the text by reading the text using a PDF viewer or a print-out of the paper, and fill a template for each of them.

The templates should contain all the information written about a specific gene or allele in the given paper. That includes any information given also about the gene products, parts of the gene, its mutated versions, gene family, homologs, etc.

The curation tool developed, called PaperBrowser [Karamanis *et al.*, 2007] aims to make the curation process more efficient; it provides two distinct ways to browse a biomedical paper being curated: a Paper view and an Entities view. The Paper view lists the gene names (which have been recognised by the Vlachos *et al.* NER system) in the order in which they appear in each section of the paper. The Entities view (Figure 6.2) is built upon the output of the anaphora resolution system: it lists groups of noun phrases recognised as referring to the same gene (coreferent relations, marked 'C'; the coreferent anaphor-antecedent pairs are merged into coreference chains) or to a biologically related entity (associative relations, marked 'a').

Clicking on a node in Entities view highlights in the same colour in the text all noun phrases listed together with the clicked node. In this way the selected node and all anaphorically related noun phrases become more visible in the text, making the curation process easier and faster; it helps the curator to focus on the information available in the paper related to a single gene at a time.

In order to assess the effect of PaperBrowser on the curation process, Karamanis *et al.* [2008] have observed and recorded how the curators navigate the article in order to find curatable information. In their experiment, for half of the articles the curators used PaperBrowser and for the other half they used a generic file viewer, which provided only a "Find" function to look for strings in the text. The curators' task was to highlight portions of text that contained information that was required for filling the templates. To estimate the efficiency of each navigation mechanism (PaperBrowser or Find), they counted the number of navigation actions (clicks on Paper view or Entities view, or searches using Find) that preceded each highlighting event. The

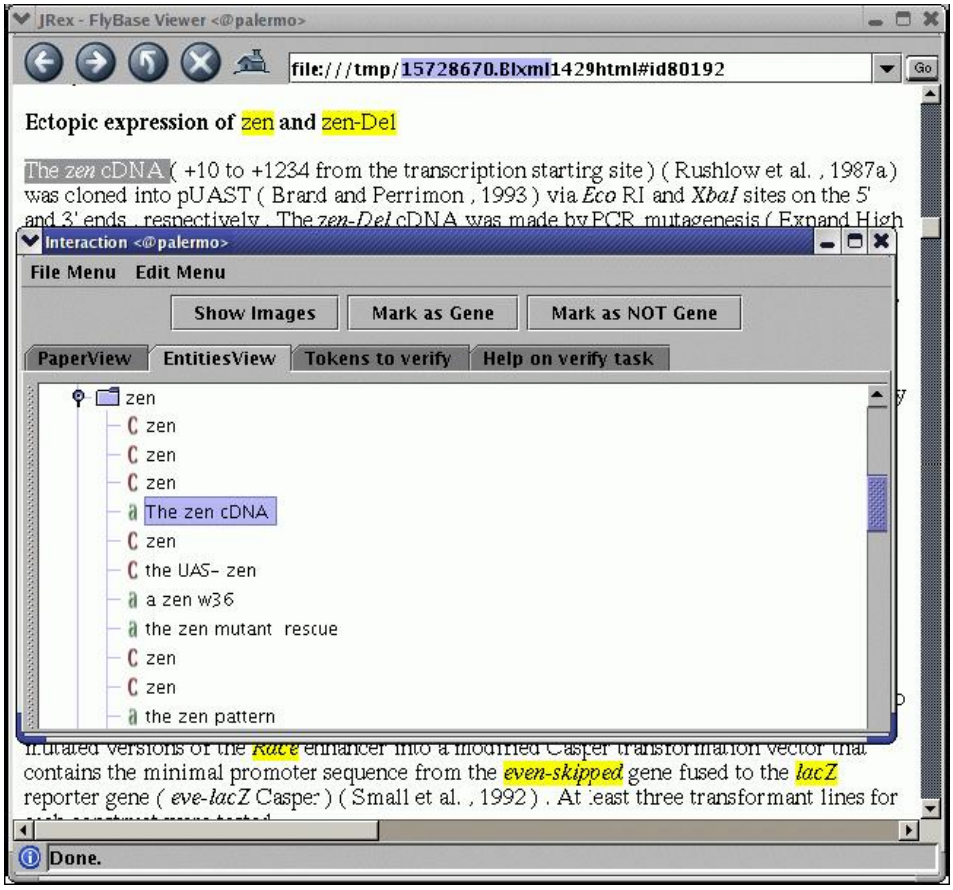


Figure 6.2: Entities view from PaperBrowser

fewer the actions, the more efficiently the curator accesses information. The authors report that PaperBrowser in its entirety makes curation 58% more efficient than using the simple "Find" function, although they have not measured the effect of the use of Entities view (the module based on anaphora resolution) alone in the curation process.

Since PaperBrower is used by humans (curators), and given that Entities view offers guidance to the curators instead of automatically extracting information, it is in principle easier for them to get around precision and recall errors made by our baseline anaphora resolution system than it would be for an automated system to do so.

6.5 Summary

In this chapter we have described our baseline system for anaphora resolution, which is rule-based and relies on string, number and biotype matching between the anaphor and antecedent candidates. It does not need training data, which is a considerable advantage but, on the other hand, it is not flexible enough to allow less obvious relations that do not conform to the restrictions encoded by the rules. In order to be able to relax the current rules, mainly the requirement for string matching, we would need to include other factors (features) into the rules; however, it is not straightforward how these could be combined to encode the characteristics of anaphoric relations.

The resulting links between the anaphoric entities are integrated into an interactive tool which aims to facilitate the curation process by highlighting and connecting related bio-entities: curators are able to navigate among different mentions of the same and related entities in order

to find the information they need to curate easily. The tool has made the curation process 58% more efficient.

In the next chapter, we present a probabilistic anaphora resolution system that aims to overcome the limitations of the baseline system. By computing the probability of the relation between the anaphor and a candidate, it aims 1) to balance the values of the features in order to be able to solve even cases where there is no string matching (relying on the other features), 2) to take into account the different types of NPs so that their specific behaviour is considered, and 3) to consider a distance measure to quantify the distance between anaphor and candidate.

Chapter 7

Probabilistic model[1]

Probabilistic classifiers and, in particular, the naive Bayes classifier, are popular in the machine learning community and in many applications [Garg and Roth, 2001]. These classifiers are derived from generative probability models which provide a way to study statistical classification in complex domains such as natural language processing. The study of probabilistic classification is that of approximating a joint distribution with a product distribution. Bayes' theorem is used to estimate the conditional probability of a class label, and then assumptions are made in the model, to decompose this probability into a product of conditional probabilities.

While the use of Bayes rule is harmless, the final decomposition step introduces independence assumptions which may not hold in the data. The most common model used in classification, however, is the naive Bayes model in which independence assumptions are extreme, that is, independence is assumed among all features. Although the naive Bayes algorithm makes unrealistic probabilistic assumptions, it has been found to work surprisingly well in practice for several applications such as text classification and spam filtering.

The good performance of naive Bayes motivates experiments to verify whether we can improve the performance of Bayesian classifiers by avoiding unrealistic assumptions about independence, that is, by making more conscious choices of independence assumptions. Friedman and Goldszmidt [1996], for example, have proposed an augmented version of the naive Bayes classifier, where a restricted Bayesian networks inducer learns at most one dependency to be maintained for each feature, while independence is assumed for all other features.

We have opted for empirically determining which dependencies to keep and consequently, which independence assumptions to make. That is the approach taken by Ge *et al.* [1998] to develop a probabilistic model for anaphora resolution of pronouns. Inspired by their approach, we have developed our model for resolution of non-pronominal anaphora in biomedical texts.

One of the advantages of probabilistic models is that they return a confidence measure (probability) for each decision they make, while decision trees, for example, do not. Another advantage of this type of model is that they consider the prior probability of each class, while other machine-learning techniques such as SVMs and neural networks do not. The use of the prior probability is especially important when training data is scarce or expensive.

Our probabilistic model results from a simple decomposition process applied to a joint probability equation that involves several features. The decomposition aims to decrease the effect of data sparseness on the model, so that even small training corpora can be viable. The decomposed model can be understood as a more sophisticated Bayesian classifier, since we consider the dependence among some of the features instead of full independence as in naive Bayes.

Our model seeks to classify the relation between an anaphoric expression and an antecedent candidate as coreferent, associative or none at all. It computes the probability of each pair anaphor-candidate for each class. To compute the probability of one pair the model does not take into account information about other pairs. The candidate with the highest overall probability

[1]Part of the work presented in this chapter has been published in [Gasperin and Briscoe, 2008].

Feature	Possible values
f_A	Form of noun phrase of the anaphor A: 'pn', 'defnp', 'demnp', 'indefnp', 'quantnp', or 'np'.
f_a	Form of noun phrase of the antecedent candidate a: same values as for f_A.
$hm_{a,A}$	Head-noun matching: 'yes' if the anaphor's and the candidate's head nouns match, 'no' otherwise.
$hmm_{a,A}$	Head-modifier matching: 'yes' if the anaphor's head noun matches any of the candidate's pre-modifiers, or vice-versa, 'no' otherwise.
$mm_{a,A}$	Modifier matching: 'yes-yes' if anaphor and candidate have at least one head pre-modifier in common but also have other mismatching modifiers, 'yes-no' if anaphor and candidate have at least one modifier in common and no mismatching modifiers, 'no-yes' and 'no-no.
$num_{a,A}$	Number agreement: 'yes' if anaphor and candidate agree in number, 'no' otherwise.
$sr_{a,A}$	Syntactic relation between anaphor and candidate: 'none', 'apposition', 'subj-obj', 'pp', and few others.
$bm_{a,A}$	Biotype matching: 'yes' if anaphor's and candidate's biotype (semantic class) match, 'no' otherwise.
$gp_{a,A}$	is biotype *gene* or *product*? 'yes' if the anaphor biotype or candidate biotype is *gene* or *product*, 'no' otherwise. This feature is mainly to distinguish which pairs can hold biotype relations.
$d_{a,A}$	Distance in sentences between the anaphor and the candidate.
$dm_{a,A}$	Distance in number of entities (markables) between the anaphor and the candidate.

Table 7.1: Features used by the probabilistic model

for each class is selected as the antecedent for that class, or no antecedent is selected if the probability of no relation is higher than the positive probabilities; in this case, the expression is considered to be new in the discourse.

Among the associative cases, we focused on biotype and set-member relations. We were not successful in dealing with the associative homolog cases, and we discuss our tentative ideas concerning the resolution of these cases in Section 7.5.

7.1 Features

We have chosen 11 features to describe the anaphoric relations between an antecedent candidate a and an anaphor A. The features are presented in Table 7.1. Most features are relational, that is, they combine information from both the anaphor and the antecedent candidate (such as $hm_{a,A}$ and $bm_{a,A}$), while the others refer to one or the other. All features but $gp_{a,A}$ are domain-independent; $gp_{a,A}$ is specific for the biomedical domain. Our feature set covers the basic aspects that influence anaphoric relations: the form of the anaphor's NP, string matching, semantic class matching, number agreement, and distance. We have designed our feature set based on previous work. Soon *et al.* [2001] and Ng and Cardie [2002c] have used several binary features to identify the type of the NP to be resolved, while we have the multi-valued feature f_A (and f_a for the candidate antecedent) with the same purpose.

All anaphora resolution systems have some kind of string matching features: for example, Soon *et al.* discard determiners and then compare strings; Strube *et al.* [2002] use a minimum-edit-distance measure to compare strings; Castano *et al.* [2002] favour the candidate with the longest substring match; Vieira and Poesio distinguish head-noun matching and modifier matching. We also distinguish head and modifier matching; we believe each type of matching plays a

role in distinct anaphoric relations (e.g. in biotype relations, modifier-modifier or head-modifier matching are more common than head-head matching).

Number agreement is commonly used to identify coreferent NPs, but we also rely on number disagreement to identify set-member relations. We have derived the number (singular or plural) of the NP head nouns from part-of-speech tags; however, we also classified as plural the NPs whose head noun was 'family' or which had coordinated head nouns.

Syntactic relations between the anaphor and antecedent candidate, in particular apposition, have been used by Soon *et al.* to identify coreference. Besides apposition, which we also use to indicate coreference, our $sr_{a,A}$ feature also encodes other syntactic relations between NPs, such as subject-object of a verb and prepositional attachment, in order to indicate the non-existence of associative relations between the syntactically related NPs (on the other hand, prepositional attachment indicates the presence of possessive relations, which we discuss separately in Section 7.6).

Semantic matching, in our case biotype matching, is also a common feature; Soon *et al.* and Ng and Cardie use WordNet semantic relations to indicate coreference, while Vieira and Poesio use WordNet in order to find semantic relations that support associative relations. Castaño *et al.* use UMLS types also to support coreference. We consider positive matching to be indicative of coreference and set-member relations, and negative matching as an indication of biotype relations.

Distance is considered in different ways in previous work. Vieira and Poesio, for example, look for an antecedent from right to left and select the closest candidate which conforms to their rules. This could be seen as distance in terms of number of relevant NPs between anaphor and candidate. Soon *et al.* measure distance in number of sentences, while Ng and Cardie have included in their system distance in number of paragraphs. We have experimented with distance in terms of paragraph and sections of the paper, but these did not contribute to performance improvement, and we kept only the distance in number of NPs (in our case only those referring to biomedical entities), $dm_{a,A}$, and in number of sentences, $d_{a,A}$.

As we do not try to resolve pronouns, we have not adopted features that are usually related to them, such as binding constraints.

Our only domain-dependent feature is $gp_{a,A}$. It indicates which anaphor-candidate pairs can take part in biotype relations, since some combinations of biotypes do not take part in biotype relations, e.g. an NP biotyped as 'part-of' a gene should not be linked to an NP biotyped as 'part-of-product'.

7.2 The resolution model

Given the above features, for each antecedent candidate a of an anaphor A, we compute the probability P of a specific class of anaphoric relation C between a and A. P is formalised as follows:

$$P(C = \text{'class'} | f_A, f_a, hm_{a,A}, hmm_{a,A}, mm_{a,A}, num_{a,A}, sr_{a,A}, bm_{a,A}, gp_{a,A}, d_{a,A}, dm_{a,A})$$

For each pair of a given anaphor and an antecedent candidate we compute P for *C='coreferent'*, *C='biotype'*, and *C='set-member'*. We also compute *P(C='none')*, which corresponds to the probability of no relation between the NPs.

We decompose the probability P and assume independence of some of the features in order to handle the sparseness of the training data. In the following equations, we omit the subscripted indices of the relational features for clarity.

$$P(C|f_A, f_a, hm, hmm, mm, num, sr, bm, gp, d, dm)$$

$$= \frac{P(C|f_A)P(f_a, hm, hmm, mm, num, sr, bm, gp, d, dm|C, f_A)}{P(f_a, hm, hmm, mm, num, sr, bm, gp, d, dm|f_A)} \tag{7.1}$$

Equation 7.1 is obtained by applying Bayes' theorem to the initial conditional probability. $P(C|f_A)$ is the prior probability of each class for each type of NP, it will encode the distribution of the classes in the training data. We kept f_A on the right side of P in order to emphasize the role of the anaphor's NP type in the model, although it could be decomposed in the same way as the other features (as described below) with no effect on the resulting model.

Since the denominator contains relational features, whose value change according to the candidate under consideration, we cannot eliminate it in the usual fashion. We keep the denominator in order to normalise P across all candidates, so that the resulting probability for the various candidates is comparable. We also want to keep the result as a probability value – if we cut the denominator it becomes an unrestricted score.

From 7.1, we then apply the chain rule to both numerator and denominator, as follows:

$$= P(C|f_A)\frac{P(f_a|C, f_A)}{P(f_a|f_A)}\frac{P(hm, hmm, mm, num, sr, bm, gp, d, dm|C, f_A, f_a)}{P(hm, hmm, mm, num, sr, bm, gp, d, dm|f_A, f_a)} \tag{7.2}$$

$$= P(C|f_A)\frac{P(f_a|C, f_A)}{P(f_a|f_A)}\frac{P(d, dm|C, f_A, f_a)}{P(d, dm|f_A, f_a)}\frac{P(hm, hmm, mm, num, sr, bm, gp|C, f_A, f_a, d, dm)}{P(hm, hmm, mm, num, sr, bm, gp|f_A, f_a, d, dm)} \tag{7.3}$$

$$= P(C|f_A)\frac{P(f_a|C, f_A)}{P(f_a|f_A)}\frac{P(d, dm|C, f_A, f_a)}{P(d, dm|f_A, f_a)}\frac{P(sr|C, f_A, f_a, d, dm)}{P(sr|f_A, f_a, d, dm)} \frac{P(hm, hmm, mm, num, bm, gp|C, f_A, f_a, d, dm, sr)}{P(hm, hmm, mm, num, bm, gp|f_A, f_a, d, dm, sr)} \tag{7.4}$$

$$= P(C|f_A)\frac{P(f_a|C, f_A)}{P(f_a|f_A)}\frac{P(d, dm|C, f_A, f_a)}{P(d, dm|f_A, f_a)}\frac{P(sr|C, f_A, f_a, d, dm)}{P(sr|f_A, f_a, d, dm)} \frac{P(bm, gp|C, f_A, f_a, d, dm, sr)}{P(bm, gp|f_A, f_a, d, dm, sr)}\frac{P(hm, hmm, mm, num|C, f_A, f_a, d, dm, sr, bm, gp)}{P(hm, hmm, mm, num|f_A, f_a, d, dm, sr, bm, gp)} \tag{7.5}$$

$$= P(C|f_A)\frac{P(f_a|C, f_A)}{P(f_a|f_A)}\frac{P(d, dm|C, f_A, f_a)}{P(d, dm|f_A, f_a)}\frac{P(sr|C, f_A, f_a, d, dm)}{P(sr|f_A, f_a, d, dm)} \frac{P(bm, gp|C, f_A, f_a, d, dm, sr)}{P(bm, gp|f_A, f_a, d, dm, sr)}\frac{P(num|C, f_A, f_a, d, dm, sr, bm, gp)}{P(num|f_A, f_a, d, dm, sr, bm, gp)} \frac{P(hm, hmm, mm|C, f_A, f_a, d, dm, sr, bm, gp, num)}{P(hm, hmm, mm|f_A, f_a, d, dm, sr, bm, gp, num)} \tag{7.6}$$

Following the decomposition, we begin to eliminate the dependencies among the features that we consider unnecessary. We based our independence assumptions on linguistic intuitions. First, we consider that the lexical (string matching) features *hm*, *hmm*, and *mm* are only dependent on the NP types (f_A, f_a) and on biotype matching *bm*. That is, depending on the NP type of the anaphor and the antecedent, it may be more likely for their words to match. For example, if the

anaphor is a proper name, it is very likely that *mm* (modifier matching) will have value "no", since proper names rarely carry modifiers. Also, once anaphor and antecedent have the same biotype, the chance that their words will match is higher than otherwise. We consider then that *hm*, *hmm*, and *mm* are not dependent on distance *d* or *dm* (the distance between two noun phrases does not influence their chance of matching), nor on syntactic relations *sr*, number agreement *num* (since we compare the lemmatised versions of the words, number does not influence string matching) or biotype *gp*. With that in mind, we can simplify the corresponding factor of our probability model as follows:

$$P(hm, hmm, mm|C, f_A, f_a, d, dm, sr, bm, gp, num) \approx P(hm, hmm, mm|C, f_A, f_a, bm)$$
and
$$P(hm, hmm, mm|f_A, f_a, d, dm, sr, bm, gp, num) \approx P(hm, hmm, mm|f_A, f_a, bm)$$

We consider that number agreement *num* depends only on the type of anaphor and antecedent NPs. For example, proper names are usually singular, so if both anaphor and antecedent NPs are proper names, the probability that *num* will have value "yes" is very high. We model *num* as independent of the distance between anaphor and antecedent (*d*, *dm*), syntactic relations *sr* between them, and semantic information (*bm*, *gp*): the chance of both NPs agreeing in number is virtually the same no matter how far from each other they are, or whether they have equal biotypes, or if they share a syntactic relation. So we assume:

$$P(num|C, f_A, f_a, d, dm, sr, bm, gp) \approx P(num|C, f_A, f_a)$$
and
$$P(num|f_A, f_a, d, dm, sr, bm, gp) \approx P(num|f_A, f_a)$$

We also consider that the semantic features biotype-matching *bm*, and gene-or-product *gp* are independent from all features but class *C*. We understand that for anaphor and antecedent candidate to have the same biotype it is not necessary that the remaining features have any specific value: there is no requirement for an entity of an specific biotype to be refered to using an specific type of NP, nor distance neither syntactic relations restrict the biotypes of the entities mentioned. We then model:

$$P(bm, gp|C, f_A, f_a, d, dm, sr) \approx P(bm, gp|C)$$
and
$$P(bm, gp|f_A, f_a, d, dm, sr) \approx P(bm, gp)$$

We also model syntactic relation *sr* to be independent of the anaphor's and antecedent's NP types f_A and f_a, since all types of NPs can take part in the syntactic relations under consideration. For example, in a subject-object relation (one NP is the subject and the other the object of the same verb), any NP type can assume subject and object positions. In some types of syntactic relations, however, one type of NP may be more frequent than others. For example, in appositive constructions it is common for the apposition to be a proper name, given the frequent occurrence of "`<extented name> ( <abbreviation> )`" constructions. We have decided though to ignore this aspect because our syntactic relation feature is very sparse: in our corpus only 15% of the intrasentential anaphoric relations (mostly coreferent) involve any syntactic relation, and we wanted to avoid further fragmentation of the statistics derived from these cases. *sr* is dependent on distance, because all syntactic relations are intrasentencial, so in the presence of a positive value for *sr* (all but "none"), the value for *d* will always be 0 (zero).

$$P(sr|C, f_A, f_a, d, dm) \approx P(sr|C, d, dm)$$
and
$$P(sr|f_A, f_a, d, dm) \approx P(sr|d, dm)$$

And finally we take d and dm to be independent of f_a, since it is the anaphor's NP type f_A that determines the distance it can be from the antecedent. As we have seen earlier, each NP type allows for a smaller or larger scope in which to find the antecedent.

$$P(d, dm|C, f_A, f_a) \approx P(d, dm|C, f_A)$$
and
$$P(d, dm|f_A, f_a) \approx P(d, dm|f_A)$$

The final equation then becomes:

$$P(C|f_A, f_a, hm, hmm, mm, num, sr, bm, gp, d, dm) \approx$$

$$P(C|f_A)\frac{P(f_a|C, f_A)}{P(f_a|f_A)}\frac{P(d, dm|C, f_A)}{P(d, dm|f_A)}\frac{P(sr|C, d, dm)}{P(sr|d, dm)}\frac{P(bm, gp|C)}{P(bm, gp)}\frac{P(num|C, f_A, f_a)}{P(num|f_A, f_a)}\frac{P(hm, hmm, mm|C, f_A, f_a, bm)}{P(hm, hmm, mm|f_A, f_a, bm)} \quad (7.7)$$

7.2.1 Comparison to Ge *et al.* model

There are a few basic differences between our model and Ge *et al.*'s model for pronoun resolution. Besides the different feature set, we have adapted other aspects of their framework to the needs of non-pronominal anaphora resolution.

The authors do not mention any special treatment of negative instances, which correspond to the anaphor-candidate pairs in which the candidate is not the annotated antecedent. Ge [Ge, 2000] describes how for each anaphor they pre-select (using the Hobbs algorithm) a maximum of 25 antecedent candidates. This limits considerably the number of negative instances generated from the training data. In our case, antecedents of nominal expressions can be found much further away than antecedents of pronouns, which impedes us from limiting the number of antecedents to be considered (although we limit distance in sentences). This results in an overwhelming number of negative instances, which has to be reduced. We discuss our strategy for reducing the number of negative instances in the next section.

In the decomposition process the authors cut out terms of the equation that they do not consider to influence the resolution process, that is, that do not change from one candidate to another. However, ignoring these terms means the outcome value of the equation is not a probability value anymore (between 0 and 1) but a score within a broader range. Since we wanted to keep the result values as probabilities (in order for them to be used in the active learning experiments described in Chapter 8), we opted to keep all terms of the equation.

In order to treat pleonastic pronouns (the pronoun "it" when not used anaphorically, e.g. "It is important to ...", which has no antecedent), Ge adds another parameter to the probability equation. She considers an additional feature that represents the occurrence of particular lexical patterns with the pronoun, which indicate their pleonastic role. This feature is considered to be independent of the other features, and it is very precise in pointing pleonastic pronouns. Unfortunately, non-anaphoric nominal expressions do no exhibit as clear patterns as pleonastic pronouns. For that reason, we decided to group features that indicate non-anaphoric NPs in a separate model, described in Section 7.7.1.

7.2.2 Training

As described in Chapter 5, for each mention of a biomedical entity, we annotated its closest coreferent antecedent (if found) and its closest associative antecedent (if found), from one of the associative classes. From our annotation we can infer coreference chains by merging the coreferent links between mentions of a same entity.

The annotated relations and the features derived from them are used as training data for the probabilistic model above. We have also considered negative training instances, which result from the absence of an anaphoric relation between a NP that precedes an anaphoric expression and was not marked as its antecedent (neither marked as part of the same coreference chain of its antecedent). The negative instances outnumber considerably the number of positive instances (annotated cases): while we have 2,452 positive instances (distributed across all NP types and anaphoric classes), we have 873,731 negative instances.

To balance the ratio between positive and negative training instances, we have clustered the negative instances and kept only a portion of each cluster, proportional to its size. All negative instances that have the same values for all features are grouped together (consequently, a cluster is formed by a set of identical instances) and only a fraction of each cluster members is kept as negative training data. We have experimented keeping $\frac{1}{2}$, $\frac{1}{3}$, $\frac{1}{5}$ and $\frac{1}{10}$ of each cluster. The higher the number of negative instances, the higher the precision of the resolution, but the lower the recall. Our best results were achieved using $\frac{1}{10}$ of each cluster, which add up to 85,314 negative instances. In this way, small clusters (with fewer than 10 members), which are likely to represent noisy instances (similar to positive ones), are eliminated, and bigger clusters are shrunk; however the shape of the distribution of the negative instances is preserved. For instance, our biggest cluster (feature values are: f_A='pn', f_a='pn', hm='no', hmm='no', mm='no', bm='yes', gp='yes', num='yes', sr='none', d='16<', dm='50<') with 33,998 instances is reduced to 3,399 – still considerably more numerous than any positive sample.

Our strategy for reducing the number of negative instances approximates the frequency values of positive and negative instances, but still maintains a considerable gap between them, without altering the shape of the distribution of negative instances. However, it alters the overall distribution of anaphoric relations, affecting the prior probabilities of each class.

Other works have used a different strategy to reduce the imbalance between positive and negative instances [Soon *et al.*, 2001, Ng and Cardie, 2002c, Strube *et al.*, 2002], where only instances with a negative antecedent that is closer than that annotated are considered. We discuss and compare the performance of both strategies in Section 7.8.

7.3 Results

Given the small size of our corpus, we did not hold out a test set. Instead, we have measured the average performance achieved by the model on a 10-fold cross-validation setting, using the whole of the annotated corpus.

We consider as antecedent candidates all noun phrases that precede the anaphor. For a given anaphor, we first select as antecedent according to each anaphora class the candidate with the highest value for P for that class. We also compute $P(C = \text{'none'})$ for all candidates. If $P(C = \text{'coreferent'}) > P(C = \text{'none'})$ for the selected coreferent antecedent, it is kept as the resulting antecedent. The same procedure is used for the selected associative antecedent with the highest probability, independent of the type of associative class. For set-member cases, where an anaphor can have multiple antecedents, if more than one candidate has the same highest probability, all these candidates are kept: this happens when the anaphor refers to the set (instead of to a member) and in such cases it is common for the antecedents to carry similar features. For instance, in the following example, "`...Reaper, Hid and Grim ...these genes...`", the three correct antecedent candidates will be at the same distance from the anaphor (since they are in the same sentence), have the same biotypes, and none will have string matching.

When no coreferent or associative antecedent is found (or when $P(C = \text{'none'})$ is higher on both cases) the anaphor is classified as discourse new.

Table 7.2 presents the performance scores we achieved for each anaphora class. The first column, 'perfect', shows the result of a strict evaluation, where we consider correct all pairs that

Class	Perfect			Relaxed		
	P	R	F	P	R	F
coreferent	55.7	57.2	56.5	73.8	63.8	68.4
assoc-biotype	29.7	35.0	32.2	31.9	37.1	34.3
assoc-set-member	35.5	39.0	37.1	38.7	42.5	40.5
discourse new	44.2	52.3	47.9	44.2	52.3	47.9

Table 7.2: Performance of the probabilistic model

exactly match an antecedent-anaphor pair in the annotated data. On the other hand, column 'relaxed' considers correct also the pairs where the assigned antecedent is not the exact match in the annotated data but is coreferent to it. For coreferent cases, the 'relaxed' scores corresponds to the MUC-style scoring (described in Section 3.4, Chapter 3). For associative cases, we also considered as correct the cases for which the antecedent selected by the system is coreferent with the associative antecedent assigned to the anaphor in the manual annotation.

It is clear that the results for coreferent cases are much better than for associative cases, but the latter are known to be more challenging. Moreover, the 'relaxed' column shows considerable improvements in comparison to 'perfect'. This means that several anaphors are being linked to the correct coreference chain, despite not being linked to the closest antecedent. This happens mainly in cases where there is no string matching between the closest antecedent and the anaphor, causing an earlier mention of the same entity with matching head and/or modifiers to get higher probability. We believe we can approximate 'perfect' to 'relaxed' results if we extend the string matching features to represent the whole coreference chain, that is, consider a positive matching when the anaphor match any of the elements in a chain, similarly to the idea presented in [Yang *et al.*, 2004].

We believe that the lower overall performance for associative cases is due to the difficulty of selecting features that capture all aspects involved in associative relations. Our set of features is clearly failing to cover some of these aspects, and a deeper feature study should be the best way to boost the scores. However, despite being lower, these performance scores are higher than those from previous approaches for newspaper texts, which used for instance the WordNet [Poesio *et al.*, 1997] or the Internet [Bunescu, 2003] as the source of semantic knowledge.

Table 7.3 shows the 'perfect' performance scores according to each type of NP. The resolution of proper names achieves the highest scores among all types of NPs for most classes. This is due to their limited structure, since proper names usually do not have elaborated premodification or modification at all, so our string matching features carried simpler patterns for these NPs. Also, 5% of the correctly-resolved coreferent cases and 88% of the correctly-resolved set-member cases where the anaphor is a proper name did not contain positive values for the string matching features; they were resolved mainly due to *sr* and *d* features, respectively. Based on these same two features, 1.5% of correctly-resolved coreferent definite NPs, 18% of correctly-resolved coreferent demonstrative NPs, 61% of correctly-resolved coreferent indefinite NPs, 5.5% of correctly-resolved coreferent other NPs, and 40% of correctly-resolved set-member demonstrative NPs where no string matching occurred could be resolved. No biotype relation was correctly resolved when there was no positive value for at least one of the string matching features. Indefinite and quantified NPs achieved the lowest scores for coreferent cases, since the highest percentage of training instances for these NPs are not coreferent (as seen in Table 5.3). Indefinite NPs, as expected, have the best scores for discourse new cases.

Class	coreferent			biotype			set-member			discourse new		
	P	R	F	P	R	F	P	R	F	P	R	F
pn	77.5	71.9	74.6	26.8	25.5	26.1	53.7	65.7	59.1	35.1	59.3	44.1
defnp	48.0	47.3	47.6	26.3	28.1	27.2	29.2	26.1	27.6	38.8	51.8	44.4
demnp	57.8	48.5	52.8	-	-	-	71.4	57.6	63.8	-	-	-
indefnp	27.0	34.2	30.2	14.2	12.5	13.3	21.0	28.5	24.2	63.4	54.7	58.8
quantnp	11.2	12.9	12.0	-	-	-	28.5	37.6	32.5	37.1	34.2	35.6
other np	41.3	41.4	41.4	30.9	48.2	37.7	19.3	19.4	19.4	49.7	56.0	52.6

Table 7.3: Performance of the probabilistic model ('perfect') per NP form

Class	hm, hmm, mm, num, bm, dm			col.1 + f_A, f_a			col.1 + gp			col.1 + sr		
	P	R	F	P	R	F	P	R	F	P	R	F
coreferent	55.9	57.4	56.7	59.5	55.9	57.7	56.5	57.6	57.1	57.0	59.7	58.3
assoc-biotype	24.0	28.8	26.2	28.7	33.9	31.1	24.1	29.9	26.7	24.4	28.5	26.3
assoc-set-member	31.1	19.9	24.3	28.4	30.2	29.3	30.9	19.0	23.5	30.4	20.5	24.5
discourse new	36.8	56.5	44.6	41.1	56.5	47.6	36.5	56.5	44.4	39.2	56.3	46.2

Table 7.4: Incremental performance of the probabilistic model

7.4 Feature analysis

Focusing on the features we have used, Table 7.4 shows how different features contribute to the final scores. The first column uses a basic set of features that rely only on string matching, number agreement, biotype matching and distance (the same information that has been used by our baseline system presented in Chapter 6). The remaining columns present the scores for adding the corresponding features to the set in the first column.

We can observe that the string matching features *hm*, *hmm*, and *mm*, the number agreement feature *num*, biotype matching *bm*, and distance in markables *dm* are the core features and achieve reasonable performance (the performance of the probabilistic model using only these features is already higher that the performance of our baseline system; we discuss this in Section 7.9). However, f_A and f_a play an important role; they increase the precision of coreferent cases and boost considerably the performance of the associative ones. This is due to the different distribution of NP types across the relations as shown is Table 5.3. The remaining features focused on specific cases: *gp* improved biotype recall, by boosting the probability of a biotype relation when anaphor or candidate have specific biotypes; and *sr* improved precision and recall of coreferent cases.

7.5 Homolog relations

In Chapter 5, we described associative homolog relations between NPs, besides associative biotype and set-member relations. Homolog relations correspond to only 2.8% of associative relations and 0.95% of anaphoric relations in general. Given this, the prior probability of homolog relations is very low and not even prominent features specifically designed to identify these relations were able to make a difference. We have experimented adding to our probabilistic

resolution model two extra features aimed at distinguishing homolog relations:

- *hom*: 'yes' if the anaphor's or antecedent candidate's head noun or pre-modifiers matches the word 'homolog' or 'homologous'; 'no' otherwise.
- *spe_np*: 'yes' if the anaphor or candidate NP contains a word that is tagged as a species name (e.g. Drosophila, mammalian); 'no' otherwise.
- *spe_sent*: 'yes' if the sentence of the anaphor or candidate contains a word that is tagged as a species name; 'no' otherwise.

We have manually annotated all words in the text that refer to organisms and different species of these, so that features *spe_np* and *spe_sent* could be collected.

The above features were not able to distinguish homolog from coreferent relations, since NPs that have positive values for these features also take part in coreferent relations. We could not design any feature that would make this distinction, which in the case of biotype relations is mainly due to the biotype (mis)matching feature and in the case of set-member relations, the number (dis)agreement feature.

7.6 Possessive relations

As described in Chapter 5, we also tried to resolve possessive relations, which we do not consider anaphoric but which previous work has often regarded as associative anaphora. We consider possessive relations to be simply syntactic relations between NPs, and tested our resolution model on finding these relations. $sr_{a,A}$ feature is responsible for encoding the syntactic relation between the NPs.

Table 7.5 presents the results on the resolution of possessive relations. Most of the recall errors are due to cases where the parser failed to recover the syntactic relation between the NPs.

Class	P	R	F
Possessive	53.5	70.5	60.9

Table 7.5: Performance of the resolution of possessive relations

7.7 Anaphoricity determination

As described above, our model aims to find the correct antecedent(s) for each mention of a biomedical entity, and when it is not able to find an antecedent, the mention is classified as discourse new. However, some researchers [Ng and Cardie, 2002b, Bean and Riloff, 1999, Uryupina, 2003] have investigated the advantage of identifying which mentions are anaphoric or not beforehand, so that the resolution system would only look for the antecedents of mentions that were found to be anaphoric.

Ng and Cardie [Ng and Cardie, 2002b] trained a decision-tree and a rule-learning classifier to distinguish between anaphoric and non-anaphoric NPs using a set of 37 features (lexical, grammatical, semantic and positional). They have reached around 65% F-measure for the MUC-6 and MUC-7 corpora. Vieira and Poesio [2000] developed a set of heuristics to identify discourse-new definite NPs, mostly based on particular syntactic constructions, reaching 75% recall and 86% precision for these cases. Bean & Riloff [Bean and Riloff, 1999] used basically the same heuristics as Vieira & Poesio, but additionally they verified whether the definite NPs were in the first sentence and also whether the NP is a 'definite only', i.e. its head always

occurs with the definite article. They have reached 78% recall and 87% precision in recognizing discourse-new definite NPs.

Uryupina trains a rule-learning classifier with discourse-new vs. discourse-old instances, using the same syntactic features used by Vieira and Poesio, plus a measure of "definite probability" derived from internet counts (how many times the NP appears with the definite article, with the indefinite article, and independent of determiner). Her system reached 88.5% precision and 84.3% recall in distinguishing anaphoric and non-anaphoric NPs.

However, these systems ultimately showed little influence on the performance of the resolution system [Poesio *et al.*, 2004]; Poesio *et al.* give an overview of the strategies that have been developed in previous work and reinforces the need for anaphoricity determination: they present results for Vieira and Poesio's anaphora resolution system and for a system similar to Ng and Cardie's without a discourse-new detector and with perfect (hand-coded) discourse-new detection, where the last brings about a 25-30% gain in precision.

In [Ng, 2004], the author argues that the usual way in which an anaphoricity determination system and a coreference resolution system are integrated, i.e. where the first system is developed independently of the second, and the second uses the output of the first as a constraint to bypass cases that were not considered anaphoric, might not be the most effective. He discusses and tests different ways of combining the systems: 1) the anaphoricity system could so be developed to optimise the results of the resolution system, instead of being developed independently; and 2) the anaphoricity information could be used as a additional feature to the resolution system, instead of being used as a by-pass hard constraint. The author tests all combinations of strategies 1 and 2 and reports best results by tuning the anaphoricity system according to the effect on the resolution system, and by considering the anaphoricity information a hard constraint. He regards as baseline the results obtained by the coreference resolution system alone (with no anaphoricity determination); he uses as coreference resolution system that described in [Soon *et al.*, 2001], and as anaphoricity determination system, that described in [Ng and Cardie, 2002b].

In the corpora used by the systems above, the majority of NPs were not anaphoric; for example, Ng and Cardie got 63.8% and 73.2% accuracy by classifying as discourse-new all NPs in the MUC-6 and MUC-7 corpora respectively. On the other hand, in our corpus, the majority of cases are anaphoric, so classifying all NPs as anaphoric can be considered the baseline performance for an anaphoricity determination system. This baseline reaches 83.3% accuracy. So, from the start, our incentive for integrating an anaphoricity determination system to our resolution system is smaller.

Aiming to beat the baseline performance, which is already quite high, we have developed a probabilistic model for anaphoricity determination in the same way that we did for our anaphora resolution model but based in a different set of features. We wanted to investigate the best way to combine it with our resolution model in the same way as proposed by Ng [2004]. However, our system was only slightly better than the baseline, having low recall for discourse-new NPs. For this reason, we have decided not to integrate it with our resolution system.

Below we describe the features we used and the model for distinguishing discourse-new and anaphoric NPs.

7.7.1 Discourse new vs. anaphoric model

Like the resolution model, this model is implemented as a decomposed probability function.

We have selected 7 features to represent the anaphoric and non-anaphoric expressions in our data, which are presented in Table 7.6. The features are related to the anaphoric expression itself, not considering any specific potential antecedent at this stage.

Ng and Cardie and Vieira and Poesio have used more elaborate syntactic patterns to identify discourse-new definite NPs, such as the presence of a proper name as head-noun modifier, or

Feature	Possible values
f_A	Form of noun phrase of the anaphor A: 'defnp' (definite NP), 'demnp' (demonstrative NP), 'pn' (proper name), or 'np' (all other NPs).
ahm	Head-noun matching: 'yes' if there is any NP preceding the anaphor that has the same head noun, 'no' otherwise.
$ahmm$	Head & modifier matching: 'yes' if there is any NP preceding the anaphor that has the same head noun AND at least one head pre-modifier in common, 'no' otherwise.
amm	Pre-modifier matching: 'yes' if there is any NP preceding the anaphor that has at least one head pre-modifier in common, 'no' otherwise.
syn_A	Syntactic pattern of the NP: 'cmod' indicating clausal modifiers, or 'none'.
num_A	Number of the NP: 'singular' or 'plural'.
pos_A	Position of the NP: 'title' (title of the paper), 'sent' (first sentence of the paper), or 'other'.

Table 7.6: Features used by the discourse-new vs. anaphoric model

the occurrence of specific constructions such as apposition, post-modification by relative clause, among others. As proper names are very frequent in biomedical texts, they very often occur as pre-modifiers and are not an indication of non-anaphoricity. We have only considered relative clause post-modification as a relevant syntactic pattern, after testing with apposition, verbal-phrase modifiers and non-clausal modifiers, which did not contribute to the performance of the model.

From these features we define

$$P(C = `class' | f_A, ahm, ahmm, amm, syn_A, num_A, pos_A)$$

We want to compute P for $C = `discourse\_new'$ and $C = `anaphoric'$ for each NP of interest in the text, and choose the class according to the higher value for P.

To reduce the influence of data sparseness on training this model, we decomposed the above probability and assumed independence among some of the features, as follows.

$P(C|f_A, ahm, ahmm, amm, syn_A, num_A, pos_A)$

$$= \frac{P(C|f_A)\ P(ahm, amm, ahmm, syn_A, num_A, pos_A|C, f_A)}{P(ahm, amm, ahmm, num_A, syn_A, pos_A|f_A)} \tag{7.8}$$

$$\propto P(C|f_A)\ P(ahm, amm, ahmm, num_A, syn_A, pos_A|C, f_A) \tag{7.9}$$

Equation 7.8 is obtained by applying Bayes' theorem to the initial equation, and Equation 7.9 is obtained by eliminating the denominator from the previous equation, which is invariant in this case. We continue the decomposition process by applying the chain rule, as in our resolution model, and get to the following equation.

$$\begin{aligned} = P(C|f_A)\ P(pos_A|C, f_A)\ P(syn_A|C, f_A, pos_A) \\ P(num_A|C, f_A, syn_A, pos_A)\ P(ahm, amm, ahmm|C, f_A, syn_A, num_A, pos_A) \end{aligned} \tag{7.10}$$

Class	P	R	F
anaphoric	86.4	96.4	91.1
discourse new	58.2	24.4	34.4

Table 7.7: Performance of the anaphoricity determination model

Class	anaphoric			discourse new		
	P	R	F	P	R	F
pn	93.7	98.8	96.2	52.6	15.6	24.0
defnp	87.4	95.0	91.0	50.0	26.8	34.9
demnp	100	100	100	-	-	-
indefnp	64.1	70.4	67.1	73.1	67.1	70.0
quantnp	75.5	97.2	85.0	50.0	7.8	13.6
other np	79.9	96.6	87.5	48.7	11.4	18.5

Table 7.8: Performance of the anaphoricity determination model per NP form

After assuming independence between some features, we get to the final equation:

$$P(C|f_A, ahm, ahmm, amm, syn_A, pos_A) = \\ P(C|f_A)\ P(pos_A|C)\ P(syn_A|C, f_A)\ P(num_A|C, f_A)\ P(ahm, amm, ahmm|C, f_A) \quad (7.11)$$

$P(C|f_A)$ represents the prior probability of each class (discourse new or anaphoric) according to each type of NP; it encodes the distribution of the classes.

7.7.1.1 Performance

The above model is able to reach an overall accuracy of 84.4%, which is very close to the performance of the baseline for anaphoricity determination (83.3% accuracy when considering all NPs to be anaphoric). The performance scores per class are presented in Table 7.7, and according to NP type in Table 7.8. Recall of discourse-new cases is low, indicating that the positive effect of the anaphoricity determination step over the resolution system would be minor, in addition to the negative impact of decreased precision of anaphoric cases.

7.8 Variations of the selection of instances

In [Ng and Cardie, 2002a, Uryupina, 2004] the authors discuss positive and negative sample selection. They reinforce that the number of negative instances is many times higher than the number of positive instances and that it is necessary to treat that, and also claim that positive instances that are too "hard" to learn should be filtered out.

For positive sample selection, Uryupina relies on empirical criteria for each type of NP to define which instances could be excluded, while Ng and Cardie iteratively train a rule inducer only with positive instances in order to identify which rules can be induced from the data and keeping only the instances which fit such rules. Both Ng and Cardie and Uryupina consider a positive instance the link between the anaphor and any member of its coreferent chain (while we only regard as positive instances the anaphor-antecedent links that were explicitly annotated in

Class	Perfect			Relaxed			Perfect for pn		
	P	R	F	P	R	F	P	R	F
coreferent	56.1	60.5	58.2	74.2	67.2	70.5	81.0	81.4	81.2
assoc-biotype	28.0	33.6	30.6	30.1	35.7	32.7	25.5	23.9	24.7
assoc-set-member	34.4	39.5	36.8	37.5	43.0	40.1	51.2	65.7	57.6
discourse new	49.8	51.2	50.5	49.8	51.2	50.5	44.3	52.2	47.9

Table 7.9: Performance of the probabilistic model with filtering of positive instances

our corpus). This is an important aspect of Uryupina's positive-sample selection strategy, since for definite NPs, for instance, she keeps only pairs where head nouns match (if there is none as part of the chain, then a different strategy is adopted).

In our case we cannot adopt the same positive sampling strategy because we do not deal only with coreference relations (as in Ng and Cardie and Uryupina), but also with associative relations. Associative relations are local and it is not realistic to replicate the associative link with a member of a chain to all other members of it. For this reason, we keep coreference relations local and only sample the anaphor-closest antecedent pairs, so that the probability distribution of the relations being considered is maintained.

However, we decided to run one experiment where we filter the set of positive instances. In order to try to improve the resolution of cases where there was no string matching, we decided to eliminate from the original training data the positive instances in which anaphor and antecedent were identical strings and had the same biotype. These cases could be resolved as coreferent independently of our model, since such NPs are very likely to corefer. On the other hand, the absence of these cases from the training data would increase the probability of partial/no-string matching as a feature of anaphoric cases (as compared with the probability of having no relation). We trained the model on the filtered instances and, for each anaphor, we (1) select the closest candidate with full string matching and same biotype, if found, (2) select the candidate with the highest probability according to the model trained on the filtered instances, and finally (3) chose the closest one between the two as the antecedent. Table 7.9 presents the results for this experiment. These results were obtained through 10-fold cross-validation over the same data used on the initial experiment presented in Section 7.3 (expect for the filtering of positive instances as described here).

Comparing with the original results from Table 7.2, as expected we observe that the recall of the model on coreference cases has improved. This strategy mainly improves the resolution of proper names (last column of Table 7.9, which can be compared with the first row of Table 7.3). We observed that most of the performance gain on coreferent cases is due to the high-precision matching of identical proper names, and a few cases of proper names with no proper-name antecedents, where there is no string matching, could be resolved given the new probability distribution of these cases. However, the impact of the new distribution of no/partial-string-matching cases was not as positive as expected. Overall, the performance scores of the resolution of NPs other than proper names decreased. The precision and recall of biotype cases for all NPs has also clearly decreased, while the scores for set-member cases have changed slightly. This is mainly because the probability of an anaphor being coreferent to a candidate with no/partial string matching became more competitive (but not necessarily more precise) than the probability of the associative cases in the same situation.We concluded that this new distribution of coreferent cases is not adequate because it disturbs the resolution of the associative cases and

Class	Prob+Closer 'perfect'			Prob+Closer 'relaxed'		
	P	R	F	P	R	F
coreferent	66.2	50.0	56.9	80.9	50.7	62.4
assoc-biotype	31.1	10.1	15.2	34.4	11.1	16.8
assoc-set-member	46.3	17.5	25.4	51.4	19.4	28.1
discourse new	31.3	88.1	46.2	31.3	88.1	46.2

Table 7.10: Performance of the probabilistic model with 'closer' negative sampling

also the resolution of coreferent cases where anaphors are not proper names. However, the high-precision matching of identical NPs can be adopted on top of our original probabilistic model.

For negative sample selection Soon *et al.* [2001], Ng and Cardie [2002c] and Strube *et al.* [2002] select as negative instances all links between the anaphor and NPs that are not its antecedent and which are closer than its antecedent. In [Ng and Cardie, 2002a] the strategy is slightly different: since the positive sampling considers the whole coreference chain, the negative sampling also does and all links between the anaphor and NPs that are not part of its coreference chain and which are closer to it than the farthest member of its chain are taken into account.

In [Ng and Cardie, 2002a] the authors argue that the candidates further away than the (farthest) coreferent antecedent are not needed for the classification, since it occurs from right to left from the anaphor until the classifier finds the antecedent. This does not hold in our case, because we rank all candidates, and instances derived from one particular anaphoric relation influence the probabilities for resolution of other relations.

We have tried training our probabilistic model using Soon *et al.* 's strategy for selecting negative instances instead of ours, described in Section 7.2.2; Table 7.10 presents the 'perfect' and 'relaxed' performance scores achieved by Soon's strategy. In our dataset, this strategy was able to reduce the number of negative instances to about $\frac{1}{3}$, while our strategy reduces it to $\frac{1}{10}$. The larger number of negative instances increases the precision scores and reduces the recall scores for all positive classes, while the opposite happens for the negative class, which defines the discourse-new scores. We reckon that the considerable drop on recall for the associative cases would make the system less viable, while the low precision for discourse-new cases shows that many anaphoric cases are left unresolved.

We regard our strategy, based on the clustering of negative instances and consecutive cluster size reduction, as more effective at proportionally eliminating negative instances that are less frequent and that are more likely to be noisy. Our approach does not alter the shape of the distribution of the negative instances; it simply approximates the frequency values of positive and negative instances to decrease the data skewness, and by doing so, increases the recall of the model.

7.9 Comparison with other approaches

We have comapred our model with three others: our rule-based baseline system (presented in the previous chapter), a naive-Bayes model, and a decision-tree-based model.

7.9.1 Rule-based baseline

When we compare our probabilistic model with our rule-based baseline system (performance scores in Table 6.2), we can observe a gain in performance for coreferent cases and an even larger improvement for associative cases. Our probabilistic model overcomes the baseline even if we use the same features for both systems, as presented in the first column of Table 7.4. The

Class	Naive Bayes			Naive Bayes relaxed		
	P	R	F	P	R	F
coreferent	34.7	48.6	40.5	57.3	65.6	61.2
assoc-biotype	13.0	33.3	18.7	13.7	34.7	19.6
assoc-set-member	18.1	23.3	20.4	21.8	27.9	24.5
discourse new	43.6	6.6	11.5	43.6	6.6	11.5

Table 7.11: Performance of naive bayes model

baseline system relies on some sort of string matching between anaphor and antecedent, and is not able to infer a relation between expressions when the matching does not occur. That is one of the main aspects that the probabilistic system aims to overcome by balancing all features together. The baseline also considers distance in a different way: it selects the closest antecedent that matches the rules, while the probabilistic system balances distance and the other features and is able to select antecedents that are not the closest.

7.9.2 Naive-Bayes baseline

As described in the introduction of this chapter, our probabilistic model can be seen as a more elaborated version of the naive Bayes classifier. While naive Bayes assumes independence among all features, in our model we carefully selected which dependences should be preserved for each feature and assumed independence in relation to those remaining. We have implemented a naive Bayes classifier using the same features we have used for our probabilistic model. The results for our anaphora resolution task using a naive Bayes model are presented in Table 7.11.

We can observe that the performance scores for all classes are considerably lower than for our probabilistic model. The discourse-new recall scores are the lowest, showing that the naive Bayes model almost always chooses an antecedent for the anaphor. The independence of the features has made it difficult for the model to identify the "none" cases, and this has contributed to its low precision for positive cases.

The considerably lower performance of naive Bayes in relation to our probabilistic model shows that preserving genuine dependencies among the features can have a big impact on the precision and recall of the resolution process.

7.9.3 Decision trees

We also compared our model with a system based on decision trees, since this approach has been taken by several other corpus-based anaphora resolution systems [Ng and Cardie, 2002c, Soon *et al.*, 2001, Strube *et al.*, 2002]. We have induced a decision tree using the C4.5 algorithm [Quinlan, 1993] implemented in the Weka tool [Witten and Frank, 2005] (in fact we induced 10 trees considering different folds for the cross-validation evaluation); we have used the same features used by our probabilistic model. We selected as the antecedent the candidate which is the closest to the anaphor for which a class other than 'none' is assigned to it in the decision tree. The 'perfect' and 'relaxed' scores for C4.5 are presented in Table 7.12. The difference between 'perfect' and 'relaxed' scores is not as large as on our probabilistic model; this shows that decision trees are more often getting even the coreference chain wrong, not only the closest antecedent. We assume that this is due to the lack of ranking among the candidates, since we opt for the obvious strategy of selecting the closest candidate that gets a positive class according to the tree (in the same way that Soon *et al.* and Strube *et al.* do).

The main disadvantage of both the baseline and decision tree systems compared with the probabilistic model is, besides the lower performance, that they do not provide a probability

Class	C4.5			C4.5 relaxed		
	P	R	F	P	R	F
coreferent	49.6	58.1	53.5	57.9	55.5	56.6
assoc-biotype	21.7	28.5	24.6	22.9	29.9	26.0
assoc-set-member	28.5	31.3	29.8	30.4	33.3	31.8
discourse new	48.5	32.5	38.9	48.5	32.5	38.9

Table 7.12: Performance of decision-tree system

assigned to each decision they make, which makes it impossible to learn how confident the model is in different cases and to take advantage of that information to improve the system. This aspect also makes it difficult to develop a consistent strategy for returning multiple antecedents for set-member cases, since there is no obvious way to do it.

7.10 Summary

In this chapter we have presented our probabilistic model for resolving coreferent and associative anaphora in biomedical texts. The model is simple, based on the decomposition of a joint probability into the product of several conditional probabilities. The model performs better than our baseline system and also better that a decision-tree model trained using the same feature set.

We have compared our strategy for negative sample selection with the more popular strategy proposed by Soon *et al.* and shown that ours is more appropriate for a probabilistic system, which ranks all candidates instead of searching them from right to left from the anaphor.

Our model, despite being simple and trained on a very small corpus, has coped well with its task of finding antecedents for coreferent and associative cases of anaphora. We have outperformed a naive Bayes classifier and a decision-tree-based classifier trained on the same data using the same features.

Our model returns a probability for each classification it makes, and this can be used as a confidence measure that can be exploited to improve the system itself or in external applications. In the next chapter, we exploit this to simulate the selection of additional instances by adopting an active learning technique.

Chapter 8

Active learning

To improve the performance of our probabilistic model for anaphora resolution, we decided to increase the amount of training data incrementally and selectively. Instead of simply selecting more articles to be fully annotated, we have adopted an active learning strategy to select particular instances that are considered more "significant" according to a given criterion, aiming to reduce considerably the amount of data to be annotated.

There are two main approaches to active learning: the uncertainty-based approach and the committee-based approach [Thompson *et al.*, 1999]. The uncertainty-based approach consists in: 1) training a classifier on a set of labelled instances; 2) applying the classifier on a set of unlabelled instances; 3) computing a measure of how confident the classifier is about the class assigned to each of the instances; and 3) selecting those about which the classifier is least confident.

The committee-based approach is different in the sense that it uses more than one classifier: 1) trains a set of classifiers on the same set of labelled instances; 2) applies the classifiers to a set of unlabelled instances; 2) computes a measure of disagreement between the classifiers for each sample; and 3) selects the instances about which the classifiers most disagree.

The efficiency of active learning can be measured in two ways: the reduction of the amount of training data necessary to achieve a given performance (usually the performance of a supervised system), or the increase in performance for a fixed amount of training data.

Here we focus on the uncertainty-based approach, using our probabilistic model as the classifier. The uncertainty-based approach has been applied, for instance, to named-entity recognition by Shen *et al.* [2004] who report at least 80% reduction in annotation costs, parsing by Tang *et al.* [2002] who reports a $\frac{2}{3}$ saving, and parse selection by Baldridge and Osborne [2003] who report a 60% saving. We are not aware of any work that has applied active learning to anaphora resolution.

Most of the works which have experimented with active learning usually simulate the acquisition and annotation of unlabelled data, that is, the data are in fact labelled and the selected instances are extracted from it automatically, simulating the manual annotation of those instances.

We have done the same: we have divided our training data in two parts, one for the initial training and the other for active learning (simulating unlabelled data), and have compared the classifier performance when trained on instances selected by active learning with its performance when trained on the same number of randomly selected instances.

In the next section we describe the strategy we have adopted to select the instances to take part in the active learning, and in Section 8.2 we describe our experiments.

8.1 Uncertainty measure

In order to measure how confident our model is about the class it assigns to each candidate, and consequently the one it chooses as the antecedent of an anaphor, we experiment with the following entropy-based measures.

We first compute what we call the "local entropy" among the probabilities for each class—*P(C="coreferent")*, *P(C="biotype")*, *P(C="set-member")* and *P(C="none")*—for a given pair anaphor(A)-candidate(a), which is defined as

$$LE(A, a) = -\sum_{C} P(C) log_2 P(C) \tag{8.1}$$

where $P(C)$ represents Equation 7.7 in the previous chapter, that is, the probability assigned to the anaphor-candidate relation by our probabilistic model for a particular class. The more similar the probabilities, the more uncertain the model is about the relation, so the higher the local entropy. This measure is similar to other entropy-based measures used in previous work for different problems.

We also compute the "global entropy" of the distribution of candidates across classes for each anaphor. The global entropy aims to combine the uncertainty information from all antecedent candidates for a given anaphor (instead of considering only a single candidate-anaphor pair as for LE). The higher the global entropy, the greater the uncertainty of the model about the antecedent for an anaphor. The global entropy combines the local entropies for all antecedent candidates of a given anaphor. We propose two versions of the global entropy measure. The first is simply a sum of the local entropies of all candidates available for a given anaphor and is defined as

$$GE1(A) = \sum_{a} LE(A, a) \tag{8.2}$$

The second version averages the local entropies across all candidates and is defined as

$$GE2(A) = \frac{\sum_a LE(A, a)}{|a|} \tag{8.3}$$

where $|a|$ corresponds to the number of candidates available for a given anaphor.

We consider that in general the further away a candidate is from the anaphor, the lower the local entropy of the pair is (given that when distance increases, the probability of the candidate not being the antecedent, *P(C="none")*, also increases), and consequently the less it contributes to the global entropy. This is the intuition behind $GE1(A)$.

However, in some cases, mainly when the anaphor is a proper name, there may be several candidates at a long distance from the anaphor that still get a reasonable probability assigned to them due to positive string matching. Therefore we decided to experiment with averaging the sum of the local probabilities by the number of candidates, so $GE2(A)$.

8.2 Experiments

Initially our training data were divided in 10-folds for cross-validation evaluation of our probabilistic model for anaphora resolution. For the active learning experiments we kept the same folds, using one for the initial training, eight for the active learning phase, and the remaining one for testing. We have experimented with all initial-training/active-learning/testing splits derived from different combinations of the 10 folds, and the results in this section correspond to the average of the results from the different data splits. A fold contains the positive and negative samples derived from about 270 anaphors; it contains about 7000 candidate-anaphor pairs (an average of about 26 antecedent candidates per anaphor). The anaphors that are part of each fold were randomly selected.

The purpose of our experiments is to check if the instances selected by using the entropy-based measures described above, when added to our training data, can improve the performance of the model more than in the case of adding the same amount of randomly selected instances.

Class	1 fold			2 folds			1 fold+$LE(A,a)$			1 fold+$GE1(A)$			1 fold+$GE2(A)$		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
coreferent	48.5	51.3	49.9	50.5	54.8	52.6	51.4	24.8	33.5	54.5	50.3	52.3	49.3	55.7	52.3
biotype	22.9	23.3	23.1	26.6	26.1	26.3	22.2	7.7	11.5	30.9	24.4	27.2	23.7	27.7	25.6
set-member	24.4	28.8	26.4	27.8	34.5	30.8	25.4	12.1	16.4	29.6	30.7	30.2	28.0	33.9	30.6
discourse new	39.8	48.1	43.6	43.6	47.1	45.3	23.4	82.5	36.5	37.7	55.6	44.9	43.4	45.2	44.3

Table 8.1: Performance of active learning

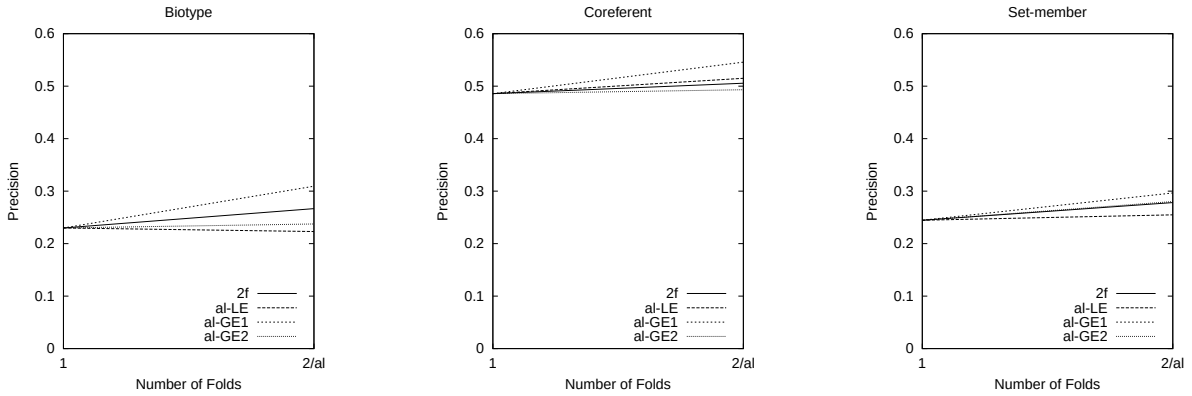
For this, we checked the performance of our model using:

1. One fold of training data.
2. Two folds of training data: we view the second fold as containing randomly selected instances, since the instances were initially randomly distributed along the 10 folds.
3. One fold of data plus 7000 instances selected using $LE(A,a)$ from the eight folds reserved for active learning: we select the same number of instances that form a fold, so that we can compare the performance with that from 2. For LE, each instance corresponds to a candidate-antecedent pair, and a fold contains on average 7000 of these. These instances are selected from the 8 folds reserved for active learning.
4. One fold of data plus all instances derived from 270 anaphors selected using $GE1(A)$ from the eight folds reserved for active learning: for GE1 we select anaphors instead of candidate-anaphor pairs, and a fold contains about 270 of them. For each anaphor selected we generate all positive and negative instances associated with it.
5. The same as 4 but using $GE2(A)$.

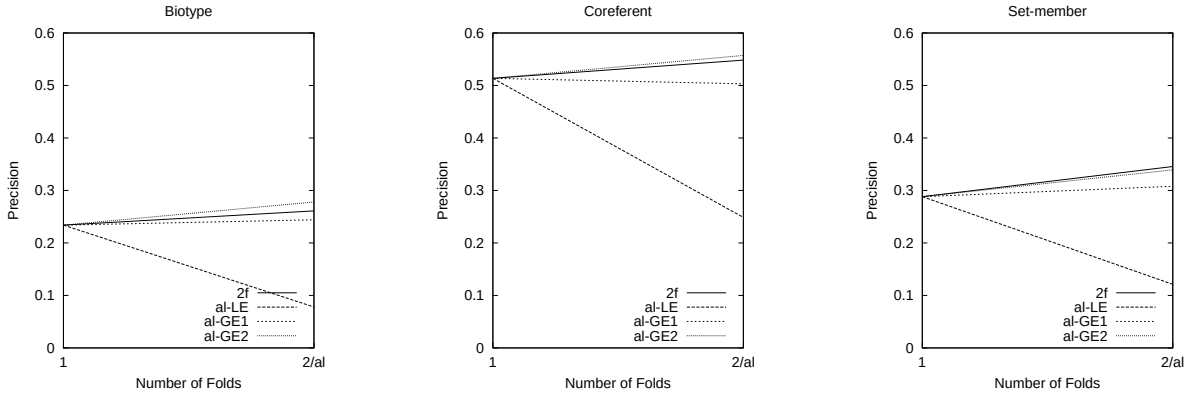
We expected (3), (4) and/or (5) to achieve better performance than (2), however this has not happened. Table 8.1 shows the 'perfect' performance scores for (1), (2), (3), (4) and (5) (these scores can be compared with those achieved using all available folds, 9, for training presented in the previous chapter).

We observe that none of the uncertainty measures that we tested has performed consistently better than random sampling, which goes against our expectations. We expected the entropy-based measures to be able to select training instances that could improve the performance of the model more than randomly selected instances (like those contained in a arbitrary data fold). $LE(A,a)$ presents the most dramatic results: it worsens the general performance of the model for all classes. However, we can observe some differences between the impact of using $GE1(A)$ and $GE2(A)$ to select instances. The precision and recall scores for '1 fold+$GE1(A)$' go in the opposite direction of the scores of the other settings: while '2 fold' and '1 fold+$GE2(A)$' training achieves better recall than precision, '1 fold+$GE1(A)$' training reaches better precision than recall. Figure 8.1 presents the graphs for precision, recall, and F-measure values for each anaphora class for '1 fold+$LE(A,a)$', '1 fold+$GE1(A)$' and '1 fold+$GE2(A)$'.

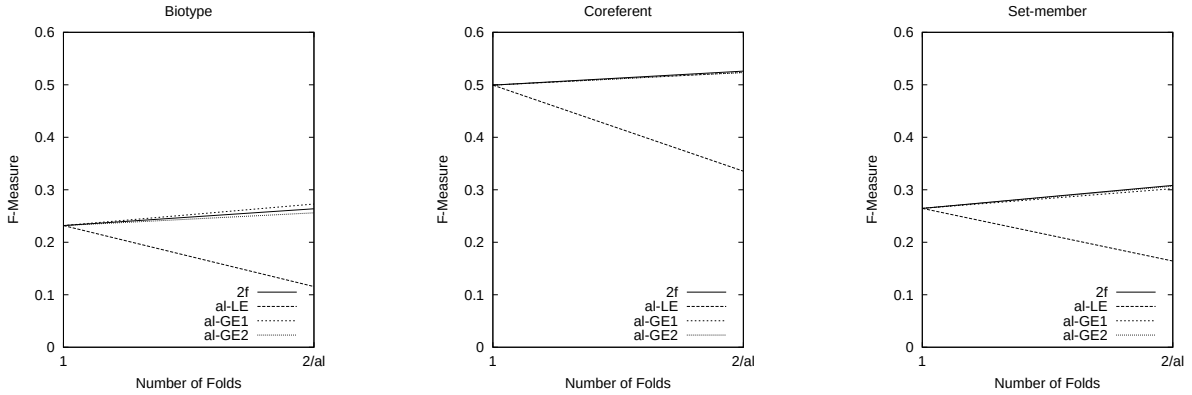
Looking at the instances selected by each active learning strategy, we observe the following. $LE(A,a)$, which considers anaphor-candidate pairs, selects mostly negative instances, given that these are highly frequent. This can explain the increase in precision and drop in recall for the positive cases (observed for coreferent and set-member, the most frequent positive classes), and



(a) Precision



(b) Recall



(c) F-measure

Figure 8.1: Graphs of the performance of active learning using $LE(A, a)$, $GE1(A)$ and $GE2(A)$

the considerable increase in recall for the discourse-new cases, since that is expected with the increase of negative instances.

$GE1(A)$ and $GE2(A)$ select a proportional number of positive and negative instances, since these measures consider an anaphor and all possible antecedent candidates, generating all instances that derive from each selected anaphor (usually one or two positive instances and several negative). However, we can observe some differences between the effect of using $GE1(A)$ and $GE2(A)$ to select instances. We observe that 70% of the instances selected by $GE1(A)$ were proper names, while the distribution of NP types among the instances selected by $GE2(A)$ is similar to the original distribution in the data. This confirms the problem we expected to have with $GE1(A)$, since exact matches of proper names that occur at a considerable distance from the anaphor still get a higher probability assigned to them, which does not happen as often with other types of NPs. On the other hand, the correct antecedent of 30% of $GE2(A)$-selected instances were in the same sentence as the anaphor, while the same happens to only 8% of $GE1(A)$-selected instances. $GE2(A)$ behaviour in this case is counter-intuitive, since antecedents in the same sentence should be found by the model with lower uncertainty than antecedents further away from the anaphor. Another counter-intuitive behaviour of $GE2(A)$ is that only 3% of the selected anaphors have no string matching with their antecedents (33% have no head-noun matching), while these cases correspond to 30% of instances selected by $GE1(A)$ (62% of instances have no head-noun matching). We expected instances involving no string matching to be selected because they are usually the ones about which the model is most uncertain.

8.3 Discussion

Despite the different behaviour presented by each of the measures, none was successful in improving the performance of the model in relation to that of random sampling. While $GE1(A)$ seems to follow the expected approach when selecting instances for active learning, it is biased in favour of selecting more proper-name anaphors.

While entropy-based measures for sample selection seem the obvious option given that we use a probabilistic model, they did not give positive results in our case. A future study of different ways of combining the local entropies is necessary, as well as the study of other non-entropy-based measures for sample selection.

The main difference between our application of active learning to anaphora resolution and previous successful applications of active learning to other tasks is the number of probabilities involved in the calculation of the uncertainty of the model. We believe this is the reason why our active learning experiments were not successful. While, for example, named-entity recognition involves a binary decision, and parse selection involves a few parsing options, in our case there are several antecedent candidates to be considered. For anaphora resolution, when using a pairwise resolution model, it is necessary to combine the predictions for one candidate-anaphor pair with the others in order to predict the global uncertainty of the model.

Given the unexpected and negative results of our simulation of active learning, we have abandoned the intention of running a genuine active learning experiment, where the manual annotation of additional unlabelled data was going to be necessary.

8.4 Summary

In this chapter we have presented our experiments with active learning to improve the performance of our probabilistic anaphora resolution system. We have adopted entropy-based measures to select new instances to be added to our training data. However, the actively selected instances were not more successful in improving the performance of the system than the same number of randomly selected instances. The three variants of the entropy-based measure we used behaved differently when selecting new instances, but none of them achieved remarkable

performance. Further studies on active sample selection for anaphora resolution are necessary.

Chapter 9

Conclusions and future directions

In this thesis we have presented our study on anaphora and anaphora resolution in biomedical scientific articles. In order for this work to take place we first investigated the nature of biomedical texts and the resources available in this domain, as well as the state of the art in the field of anaphora resolution.

In Chapter 2, we discussed different properties of biomedical texts, such as the particular NP-form distribution in comparison with other genres of text, the presence of gene and protein names and ambiguity between them, and the background knowledge necessary for understanding the relations between the entities in the text. We have studied the resources (ontologies, terminologies, databases) available in the biomedical domain, which we used to explore as source of semantic knowledge for the anaphora resolution process. We have also investigated which tasks have usually been performed on biomedical texts, such as named-entity recognition, so we could identify the procedures available for identifying the biomedical entities in the text.

We have also studied the works that have been developed for resolution of anaphoric non-pronominal NPs, both within and outside the biomedical domain. We have discussed the difference between anaphora and coreference and have opted for treating the combination of both. We have studied systems that have been developed aiming to resolve only coreference and coreferent anaphora (such as [Ng and Cardie, 2002b]) and also systems aiming to solve both coreferent and associative anaphora (such as [Vieira and Poesio, 2000]). We have reviewed knowledge-based and corpus-based systems; the former require hand-crafting or hand-tuning of rules and are consequently more conservative; while the latter rely on training data to infer the appropriate behaviour. We have also researched works on anaphora resolution in the biomedical domain and have seen that the very few existing systems are developed and tested on abstracts of scientific papers, instead of full-text articles.

Given the need of biology researchers for information extraction, we have decided to focus on anaphora resolution in the biomedical domain, which would be able to resolve both coreferent and associative anaphora in full-text scientific articles. Anaphora resolution is an important step towards increasing the recall of information extraction systems.

We have opted for a corpus-based approach since it has been increasingly popular in anaphora resolution [Soon *et al.*, 2001, Ng and Cardie, 2002b, Strube *et al.*, 2002] and showed promising results. Besides, corpus-based approaches are more flexible than knowledge-based approaches and this allowed us to explore the problem more openly.

We have built a framework in which to fit our anaphora resolution system. We have set a pipeline for identifying and classifying NPs referring to biomedical entities. We have employed a named-entity recogniser to identify gene and protein names [Vlachos *et al.*, 2006], the RASP parser [Briscoe *et al.*, 2006] to identify NP boundaries and NP subconstituents (e.g. head noun, pre-modifiers), and have developed a strategy for identifying and classifying the NPs referring to biomedical entities based on the Sequence Ontology [Eilbeck and Lewis, 2004]. Using the terminology present in the Sequence Ontology we were able to recognise mentions of entities

of interest in the text; to increase the coverage of the Sequence Ontology we have enriched it with specific sets of terms derived from the UMLS Metathesaurus. We have used the ontological relations from the Sequence Ontology (mainly *is-a*, *part-of* and *derived-from* relations) to determine a set of seven biotypes that were used to classify the mentions: "gene", "product", "subtype", "part-of", "part-of-product", "supertype", and "variant". The main limitation of this classification approach was the lack of a mechanism for resolving ambiguous terms, that is, terms that could have multiple biotypes assigned to them (e.g. "sequence").

We identified the four most prominent anaphoric relations present in biomedical articles: coreference and coreferent anaphora, associative biotype relations, associative homolog relations and associative set-member relations. Coreferent relations are practically the same as in other domains, they are relations between two distinct mentions of the same entity. Associative relations are anaphoric relations between mentions of related entities. The associative biotype relation is the relation between two mentions of different biotype, such as a gene and its parts or products. The associative homolog relation occurs between mentions of entities of the same biotype but which are homologs of each other (corresponding entities in different organisms). The associative set-member relation is common in other domains as well: it is the relation between a mention of a single (or subset) entity and that of a group of entities that contains the single one; in this relation anaphors may have multiple antecedents or multiple anaphors that point to the same antecedent. Coreferent relations are considerably more frequent than associative relations. Set-member relations are the most frequent among associative relations, followed by biotype relations. Homolog relations are rare.

We have developed guidelines for annotating these relations in a set of biomedical articles and have built the first corpus of full-text biomedical articles with anaphora information. For each anaphor, a coreferent and an associative antecedent were annotated if found. The corpus is composed of 5 papers, which contain 2696 NPs referring to biomedical entities (most of them gene products) and 3037 anaphoric relations between them. With this corpus we were able to conduct experiments on a corpus-based anaphora resolution system.

We wanted to adopt a corpus-based approach which is reliable despite being trained on such a small corpus and in which all aspects of the data could be explored. We wanted this approach to be able to resolve the types of anaphoric relations that we had defined: coreferent, associative biotype, homolog and set-member. We decided to implement a non-parametric statistical approach based on the decomposition of a probability conditional on several features. The decomposition of the model reduces the impact of the sparseness of the data on the performance of the model. This model collects statistics from the training corpus and it considers the prior distribution of each class of anaphoric relation. We have selected a set of 11 features to represent a relation between an anaphor-antecedent pair; these features involve string matching features which distinguish between head-noun matching and pre-modifier matching, a number agreement feature, a semantic matching feature which relies on the biotypes assigned to each class, distance features that represent the distance between the anaphor and the antecedent candidate, the types of the anaphor NP and the candidate NP, and any syntactic relation between anaphor and antecedent. From the training corpus we can also infer negative samples, which represent the absence of an anaphoric relation between pairs of NPs that were not linked in the annotation process. Such samples correspond to the overwhelming majority of samples extracted from the corpus and skew the distribution of the relations. We have developed a strategy for proportionally reducing the number of negative samples in order to leave more room for the positive samples to influence the resolution model; our strategy proved to be more effective than the most popular one for our probabilistic system. We trained our system on 90% of our corpus and tested it on 10% of it in a cross-validation setting. Our system reached a good level of performance, 56-68% F-measure, on coreferent relations and reasonable performance of associative

biotype and set-member relations. We were not able to resolve homolog relations since we could not find features which would distinguish these from coreferent relations. Our model is the first corpus-based technique that aims to solve associative anaphora.

We compared the performance of our model with that of a simple rule-based baseline system, a naive-Bayes-based system, and a decision-tree-based system. Our model outperformed all three. The baseline system is composed of a set of rules whose aim was to to identify the most intuitive cases of coreferent and associative cases of anaphora; the system relies on string matching, biotype matching, number agreement and distance. The baseline system proved to be too conservative and achieved very low recall for associative cases, it could not solve cases that did not contain some sort of string matching. The naive-Bayes and the decision-tree systems were trained on the same corpus and used the same features that we have used to train our probabilistic model. The performance of the naive-Bayes system was considerably lower than for our probabilistic model for all classes of anaphoric relations. We concluded that the complete independence among the features as modeled by naive Bayes has made it difficult for the model to identify the cases where there is no relation between the anaphor and a candidate, and this has contributed to its low precision for positive cases. The decision-tree system achieved performance similar to that of our probabilistic system on coreference cases but lower recall and precision on associative cases. We could, moreover, observe that a considerable number of cases which were incorrectly resolved by the decision-tree system were not only incorrect in terms of not choosing the correct closest antecedent but it selected antecedents that were unrelated (not part of the same coreferent chain) to the correct antecedent. Our probabilistic system, on the other hand, could identify several antecedents that, despite not being the closest to the anaphor, were coreferent to it. This happens mainly because our probabilistic system is able to rank all candidates and choose the one with the highest probability, while when using a decision-tree approach we select the first antecedent accepted by the tree.

Since our model offers a confidence measure of its decisions, that is, the resulting probability assigned to each relation, we decided to experiment with active learning in order to investigate whether it could help us improve the performance of the system by selecting a reduced number of significant samples to be added to the training data. We have adopted three variations of an entropy-based measure to identify anaphors about which the model was uncertain, in order for these to be selected for manual annotation and added to the training data. Unfortunately our measures were not able to capture the uncertainty of the model and did not select samples that caused an improvement in the performance of the model higher than that caused by the addition of randomly selected samples. We believe this is related to the number of cases involved in the calculation of the entropy measure, since we take into account the probabilities assigned to each antecedent candidate available to the anaphor.

We believe that our work has made a significant contribution to the field of anaphora resolution in biomedical texts. The anaphora relations that we identified (and the corpus we created with them) can be used by the community as a basis for investigation of relations of interest. Our strategy for semantic typing of the NPs of interest can be reused to type other corpora in this domain. Our probabilistic model has proved able to cope with the task of coreferent and associative anaphora resolution despite the small amount of training data. And finally, we have opened the question of active learning for anaphora resolution and the problems one may encounter in its attempt.

The next section suggests some directions of future work that could follow from the work done in this thesis.

9.1 Future work

Here we outline a few extensions of the work we have done in this thesis, which could be pursued in the future.

Our biotype tagging strategy can be refined to include biomedical events besides entities. The other-bio biotype could be refined by using, for example, the Gene Ontology to distinguish NPs that refer to molecular function or biological processes. This would make possible anaphora resolution between events and also between entities and events.

Since the inter-annotator agreement for associative relations in our corpus was not ideal, it would be necessary to refine our guidelines so that better agreement can be reached. We could clarify a few issues and enforce a particular attitude towards them (e.g. not to annotate associative relations when anaphor and antecedent are explicitly related by a syntactic relation). For the sake of consistency in the annotation, it would be important to define a single way to proceed in face of unusual cases, for example, pointing to a specific action to be taken when "mixed" relations are found (when the anaphor could have different associative relations with more than one antecedent).

We believe the types of anaphoric relations we identified could also be refined. For example, set-member relations could be split among different types of relations, since they cover variations of the set-membership behaviour. For example, the relation between a set and a single member could be distinguished from the opposite, member to set, relation. It would be interesting to investigate if such distinctions can reveal specific problems of the resolution of set-member cases.

We think the homolog relations are both biologically and linguistically interesting and deserve further attention. We believe the information contained in this relation is relevant to information extraction efforts, since the properties of an entity are usually shared by homolog entities. Linguistically, as shown by Example 39 in Chapter 5, homolog relations can be very subtle and it is possible that deeper discourse-related features are necessary to distinguish them from coreferent relations.

Our feature set should be expanded in order to try to better represent the characteristics of associative relations. A feature study would be required to identify which features to include. We believe features related to the coreference chain of the anaphor can contribute to the resolution of the associative cases. Yang *et al.* [2004], for example, consider the words present in the entire coreference chain of a candidate when matching it against an anaphor; the authors use this to resolve coreferent cases but we believe it can have an influence in the resolution of associative cases as well once coreference is known. However, the inclusion of additional features would require more training data to be annotated to compensate for the increase in the sparseness of the data.

We would like to validate the anaphoric relations we have found in our corpus and our whole resolution framework on articles in other subdomains of biology, besides the fruit fly genomics.

Uses of active learning in anaphora resolution definitely deserve further attention. Since annotating anaphora is a costly task, reducing the amount of annotated data needed to train a model is particularly important. It is necessary to investigate how to represent the uncertainty of a model when it faces several decisions in order to make a final one, as in the case of anaphora resolution, since the process has to choose from several candidates which will be the antecedent.

It would be interesting to make a comprehensive study of the dimension of the contribution of anaphora resolution to information extraction applications in the biomedical domain. In Chapter 6 we have shown how our baseline system contributes to a tool for facilitating the curation of biomedical articles. It is important to investigate how much anaphora resolution can affect other applications, for example, the extraction of relations between entities (and events) in the text.

Appendix A

Coreference and anaphora annotation guidelines

The annotation process consist of establishing links between mentions of biomedical entities (which have been previously marked). We divided this task in two phases. In the first phase, the annotator will link mentions that refer to a same entity (coreferent), while in the second phase, the annotator will link mentions who are related to each other, but which do not refer to the same entity (associative). The next sections explain both phases.

A.1 First phase: Linking coreferent mentions

Different mentions that refer to a same entity are called coreferent mentions. For each mention in the text, the annotator will have to check if there is another mention previously in the text (called antecedent) that refer to the same entity as the current mention. Always look for the closest previous mention. Coreferent mentions should have the same biotype.

When a coreferent antecedent for the current mention is found, both mention and antecedent should be added together to a set in order to keep track of the whole coreference chain. In the following examples of coreferent relations each mention is represented with an 'id' attribute that is used to identify the mention. The 'set' attribute refers to a set of coreferent mentions.

(29) ''...the expression of <m id="1" biotype="gene" set="set_1">**X-linked genes**</m> is equal in males and females...the hypertranscription of <m id="2" biotype="gene" set="set_1">**the X-chromosomal genes**</m> in males...''

(30) ''...is composed of <m id="10" biotype="product" set="set_2">**five proteins**</m> encoded by the male-specific lethal genes... <m id="15" biotype="product" set="set_2">**The MSL proteins**</m> colocalize to hundreds of sites...''

A.1.1 Special cases

A.1.1.1 Apposition

When you find appositive mentions like:
''the remaining protein, MSL3,''
the annotator should link the appositive mention to the main one as coreferent, as in example 31:

(31) ''<m id="30" biotype="product" set="set_10">**the remaining protein**</m>, <m id="31" biotype="product" set="set_10">**MSL3**</m>''

A.1.1.2 Predicates

Predicative mentions, as in example 32, should also be annotated as coreferent.

(32) ``<m id="40" biotype="gene" set="set_11">**ced-4**</m> is <m id="41" biotype="gene" set="set_11">**an pro-apoptotic gene**</m>''

A.2 Second phase: Linking associative anaphoric mentions

Associative anaphoric relations between mentions rely on factual relations between the biomedical entities referred to by them. These factual relations are assumed by the writer of the paper to be known by the reader, they represent the actual relations between the biomedical entities, independent of the text, for example, the fact that a gene encodes a protein, or that a gene is composed by DNA sequences. The annotator should consider as associative cases the instances where these factual relations imply a dependency between the mention and its antecedent, that is, the meaning of the mention would not be fully understood (or the mention would seem to be out of place) if it was not for its relation with the antecedent.

Given that the associative relations are more subtle than the coreference relation (annotated in the previous phase), the span of associative anaphoric links is usually shorter, that is, associative antecedents are usually close to the anaphor, while coreferent antecedents can be further away. The annotator should look for associative antecedents mainly within the same section of the paper as the anaphoric mention, it is unlikely that they will be far from the current mention.

Also, once the entity referred to by the current mention has already been mentioned recently in the text (in which case a coreferent relation should have been marked in the previous phase), it is very unlikely that this mention would have an associative relation with a previous mention. This means that an entity that is salient in the readers mind (because of its recent mention) does not need indirect (associative) relations to be referred to.

In summary, the annotator should make his decision about the existence of an associative link to a previous mention by weighting his/her perception of the dependency between the mentions, their distance from each other and the salience of the entity the mention refers to in the text.

We are interested in three types of associative relations:

- biotype associative: when the related mentions have different biotypes (e.g. a gene and its protein)
- homolog associative: when the related mentions are homolog and have the same biotype (e.g. a gene and its homolog from another organism)
- set-member associative: when one of the related mentions refers to a set that contains the referent of the other mention (e.g. a gene and its family)

For each mention in the text, the annotator will have to check if there is another mention previously in the text which holds an associative relations with the current mention and, if so, the relation should be classified according to the above classes. Always look for the closest previous mention.

Below we show examples of all types of associative relations. In the examples, a mention that holds an associative relation with its antecedent gets a pointer link to the related mention; the 'ante' attribute represents this pointer, it refers to the identifier of the closest associative-related mention. The 'rel' attribute identifies the type of associative relation.

If the annotator feels that there is an associative relation between the current mention and an antecedent but such relation does not fit any of the above classes, the relation should not be marked.

A.2.1 Biotype relation

We call biotype relation an associative anaphoric relation between two mentions that refer to different entities with different biotypes, as in examples 33 and 34. In such cases, the associative relation should be marked as 'biotype'.

(33) ''There was considerable excitement in the field when potential mammalian and Drosophila homologs for <m id="20" biotype="gene">**ced-3**</m> were discovered.<m id="25" biotype="product" ante="20" rel="biotype">**The CED-3 protein**</m> is one of...''

(34) ''...the role of <m id="30" biotype="gene">**the roX genes**</m> in this process... interact with <m id="35" biotype="partof" ante="30" rel="biotype">**the roX RNAs**</m>''

In example 33, the gene "ced-3" is introduced in the text, and in the next sentence the writer starts talking about the "CED-3" protein, assuming that the reader knows the relation between the gene and the protein. In example 34, a similar situation occurs; a mention to the "roX RNAS" would seem out of place in this context if the "roX genes" had not been mentioned before.

Antecedents for a biotype relation are usually close to the current mention, in the same paragraph or same section of the paper. Biotype relations can occur between mentions of the following biotypes:

	Gene	Subtype	Variant	Supertype	Part-of	Part-of-product	Product	Otherbio
Gene		X	X	X	X	X	X	X
Subtype	X		X	X	X	X	X	X
Variant	X	X		X	X	X	X	X
Supertype	X	X	X					
Part-of	X	X	X					
Part-of-product	X	X	X				X[1]	
Product	X	X	X			X[1]		X[1]
Otherbio	X	X	X				X[1]	

This table basically shows that one element in a biotype relation is always a gene mention (gene, subtype and variant biotypes), with the exception of the cases where the relation is between a product mention and a part-or-product or otherbio mention.

However, there are some misleading cases which we do not consider as a biotype relation; these are cases where mentions to related entities are close to each other in the text but there is no implicit dependency between them and the latest mention can be understood independently. For instance, in example 35, the relation between "pro-domains" and "caspases" is explicit in the text, so we do not consider the relation between these mentions as associative anaphora and it should not be marked. Example 37 also presents a case where mentions should not be linked, the preposition 'of' makes the relation explicit; associative anaphora in general does not happen between mentions in the same clause.

(35) ''<m id="40" biotype="product">**Initiator caspases**</m> usually have <m id="41" biotype="partof_product">**long pro-domains**</m>''

[1]These are the only biotype relations that do not involve gene mentions; they involve gene products as a central role instead.

(36) ``<m id="50" biotype="partof"> **The 38-bp consensus TIR**</m> of <m id="51" biotype="subtype">**Transib transposons**</m> consists of ...''

The passage below (example 37) also shows a case which should not be considered as a biotype relation.

(37) ``The expression of <m id="60" biotype="gene" set="set_5">**reaper**</m> has been shown to be regulated by distinct stimuli, including X-irradiation, steroid hormone signaling and a block in cell differentiation. Recently, a Drosophila p53 ortholog was identified by searching the genome database, and it was shown to bind to <m id="61" biotype="partof" ante="60" rel="biotype">**the reaper promoter**</m> ... The observation that transcription of <m id="62" biotype="gene" set="set_5">**reaper**</m> can be induced by ...''

In this example, the relation between mentions 61 and 60 is a genuine biotype relation; but on the other hand there is no biotype relation between mentions 62 and 61. That is because "reaper" (mention 62) has recently been mentioned in the text (there is a coreference relation between mentions 60 and 62) so there is no need for the reader to imply the existence of "reaper" (62) (or to recall its role in the text) from its relation with "reaper promoter" (61), since it is already being talked about directly in adjacent text.

A.2.2 Homolog relation

Another type of associative relation that we are interested in is the homolog relation (e.g. between homolog genes or homolog proteins). In this case, the related entities have the same biotype but are homologs. Two genes or gene products are homologs when they share a common ancestor, occurring within one species or in different organisms. However, entities of all biotypes (except 'otherbio') can take part in homolog relations, since authors can refer to, for instance, homolog sequences (that can be part of a gene, or part of a protein sequence). See example 38, where the associative relation is marked as 'homolog'.

(38) ``... <m id="70" biotype="gene">**mammalian Bok**</m> ... <m id="75" biotype="gene" ante="40" rel="homolog">**the Drosophila Bok homolog**</m> ...''

In this example, the Drosophila homolog is only introduced as such (i.e. as homolog instead of simply as a gene) because the text had been talking about the mammalian gene. That is, referring to "the Drosophila Bok homolog" would seem out of place if there was no previous mention to the gene it is homologous to. Homologous mentions should only be linked as 'homolog' associative anaphora when it is the homology relation that make both mentions related.

Usually at least one of the mentions which take part in homolog relation contain the word "homolog" or some species name (e.g. Drosophila, mammalian, mouse, insect).

As for the biotype relations, homolog relations usually occur between mentions that are close to each other, within the same paragraph or same section of the paper.

The annotator should not consider as 'homolog' associative relations the cases where two entities, even if known to be homologous, are referred to independently of their homology. For instance, see the example below:

(39) ``<m id="80" biotype="product" set="set_4">**CED-4**</m> translocates to the nuclear membrane where it activates CED-3, resulting in programmed cell death. Only <m id="81" biotype="product" set="set_5" ante="80" rel="homolog">**one mammalian CED-4 homolog**</m>, <m id="82" biotype="product" set="set_5">**Apaf-1**</m>, has been extensively characterized to date. Like <m id="83" biotype="product" set="set_4">**CED-4**</m>, <m id="84" biotype="product" set="set_5">**Apaf-1**</m> requires dATP for caspase activation.''

The relation between mentions 81 and 80 is a genuine homolog relation, and mention 82 is coreferent to 80. But, despite we knowing that "CED-4" and "Apaf-1" are homologs, there is no homolog relation between mentions 84 and 83.

A.2.3 Set-member relation

The third type of associative relation is the set-member relation, which happens when an entity is related to a set of entities that it is part of or vice-versa. The single entity and the entities in the set have the same biotype. It happens mostly in the presence of coordinated NPs (Subsection A.2.3), plural NPs (Subsection A.2.3), and NPs referring to families of bio-entities (Subsection A.2.3). Set-member relations also have a short range, with the majority of antecedents happening within the same section of the paper, but in some cases the links can cross sections boundaries.

Coordination

It is common to find in the text some mentions like:
"the genes reaper, hid, and grim"

Mentions like the one above, which contains coordination, can have multiple antecedents. When this is the case, the relation between the mentions should be marked as 'set-member'. The same is valid for the case when a simple mention refers to one like the one above. For example, for the following portion of text:

(40) ``... <m id="30" biotype="gene">**reaper**</m>, <m id="31" biotype="gene">**hid**</m>, and <m id="32" biotype="gene">**grim**</m> are regulators of apoptosis ... <m id="35" biotype="gene" ante="30,31,32" rel="set-member">**the genes reaper, hid, and grim**</m>...''

List

When a set of entities is mentioned and its mention is followed by a list of its members, as in an apposition construction, the members should be linked to the set by a 'set-member' relation. Members can be listed between commas, as in Example 41 or in brackets, as in Example 42.

(41) ``... <m id="40" biotype="product">**two proteins**</m> encoded by the recombination-activating genes, <m id="41" biotype="product" ante="40" rel="set-member">**approximately 1040-aa RAG1**</m> and <m id="42" biotype="product" ante="40" rel="set-member">**approximately 530-aa RAG2**</m>,...''

(42) ``... <m id="50" biotype="product">**surface receptors**</m> of vertebrate B and T immune cells (<m id="51" biotype="product" ante="50" rel="set-member">**BCRs**</m> and <m id="52" biotype="product" ante="50" rel="set-member">**TCRs**</m>) .''

Plural

Plural mentions will be treated the same way as coordinated mentions, as they can also have multiple antecedents or be the antecedent of multiple mentions. The relation is associative and should be marked as 'set-member'. See example 43.

(43) ``... <m id="60" biotype="gene">**ced-4**</m> and <m id="61" biotype="gene">**ced-9**</m> ... <m id="65" biotype="gene" ante="60,61" rel="set-member">**the genes**</m> ...''

Family

An entity mention can be related to a mention of its family, and this is also a case of set-member associative relation. See example 44 and 45.

(44) ``... <m id="70" biotype="product">**the mammalian anti-apoptotic protein Bcl-2**</m> ... <m id="75" biotype="product" ante="70" rel="set-member">**Bcl-2 family**</m> ...''

(45) ``... <m id="80" biotype="product">**the MSLs**</m> ...
<m id="85" biotype="product" ante="80" rel="set-member">**MSL-1**</m> ...''

Subset

We also consider as 'set-member' relation the relation between a set and a subset of it, as in the example below.

(46) ``<m id="90" biotype="otherbio">**D-mib mutant discs**</m> have no wing pouch ... The complete loss of D-mib activity in <m id="92" biotype="otherbio" ante="90" rel="set-member">**D-mib1 mutant discs**</m> ...''

Other

This is a special case of set-member relations, which include mentions that contain the word 'other' (or similar words, like 'remaining'), as in Example 47.

(47) ``...distribution in females ectopically expressing <m id="5" biotype="product">**MSL2**</m> but lacking <m id="6" biotype="product" ante="5" rel="set-member">**other MSL proteins**</m>.''

In these cases, the 'other' mentions should be linked to its complement, that it, its antecedent is the mention referring to the item excluded from the set.

However, as said before, if there is no implicit dependency between the mentions for the understanding of the later one, there is no anaphoric relation and the mentions should not be linked. See example 48:

(48) ``<m id="90" biotype="product">**CED-9**</m> and <m id="91" biotype="product">**EGL-1**</m> belong to <m id="92" biotype="product">**a large family**</m> of <m id="93" biotype="product">**proteins**</m> related to ...''

In this example, "a large family" should not me linked to mentions 90 and 91, because the set-membership is being explicit in the text (by the verb 'belong'). "proteins" should also not be linked to "a large family" for the same reason (the preposition 'of' makes the relation explicit).

As for biotype relations, set-member relations should not be marked when the given mention is already coreferent to an antecedent in the same section, as below, where mention 69 should not be linked to mention 65, because mention 69 is coreferent to mention 60, which is close by.

(49) ``... <m id="60" biotype="gene" set="set_10">**ced-4**</m> and <m id="61" biotype="gene">**ced-9**</m> ... <m id="65" biotype="gene" ante="60,61" rel="set-member">**the genes**</m> ... <m id="69" biotype="gene" set="set_10">**ced-4**</m> ...''

A.2.4 Mixed relations

There are cases where the type of relation between two mentions is mixed, that is, it could be seen as more that one of the above types of associative relation. In Example 50, the relation between mentions 12 and 10 can be seen as biotype (gene-otherbio relation) and set-member.

(50) ``While <m id="10" biotype="gene">**the neur and mib genes**</m> are evolutionarily conserved, ... events requiring <m id="12" biotype="otherbio">**neur activity**</m>.''

In such cases, the annotator should select the type of relation that he/she feels to be more prominent.

A.2.5 General remarks

The text to be annotated may contain table or figure captions. Mentions that are part of captions can have associative links to mentions in the body of the text, but NO mention in the body of the text should be assigned a link to a mention in a caption.

Bibliography

[ACE, 2004] ACE Project, http://www.ldc.upenn.edu/Projects/ACE/docs/EnglishLNKV3-0.PDF. *ACE Annotation Guidelines for Entity Link Tracking (LNK), Version 3.0*, 2004.

[Aone and Bennett, 1995] C. Aone and S. Bennett. Evaluating automated and manual acquisition of anaphora resolution strategies. In *Proceedings of ACL'95*, pages 122–129, 1995.

[Azzam *et al.*, 1998] Saliha Azzam, Kevin Humphreys, and Robert Gaizauskas. Evaluating a focus-based approach to anaphora resolution. In *Proceedings of COLING-ACL'98*, pages 74–78, 1998.

[Bagga and Baldwin, 1998] Amit Bagga and Breck Baldwin. Algorithms for scoring coreference chains. In *Proceedings of LREC-1998 Workshop on Linguistics Coreference*, 1998.

[Baldridge and Osborne, 2003] Jason Baldridge and Miles Osborne. Active learning for hpsg parse selection. In Walter Daelemans and Miles Osborne, editors, *Proceedings of CoNLL-2003*, pages 17–24. Edmonton, Canada, 2003.

[Bean and Riloff, 1999] David Bean and Ellen Riloff. Corpus-based identification of non-anaphoric noun phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*, pages 373–380, 1999.

[Blaschke *et al.*, 1999] Christian Blaschke, Miguel A. Andrade, Christos Ouzounis, and Alfonso Valencia. Automatic extraction of biological information from scientific text: protein-protein interactions. In *Proceedings of ISMB'99*, pages 60–67, 1999.

[Blaschke *et al.*, 2004] Christian Blaschke, Lynette Hirschman, and Alexander Yeh (eds.). *Proceedings of the BioCreative Workshop, Granada*. 2004.

[Bodenreider, 2006] Olivier Bodenreider. Lexical, terminologival, and ontological resourses for biomedical text mining. In Sophia Ananiadou and John McNaught, editors, *Text Mining for biology and biomedicine*, pages 43–66. Artech House, 2006.

[Boguraev and Kennedy, 1999] Branimir Boguraev and Christopher Kennedy. Salience-based content characterisation of text documents. In Inderjeet Mani and Mark T. Maybury, editors, *Advances in Automatic Text Summarization*. MIT Press, 1999.

[Briscoe and Carroll, 2002] Edward J. Briscoe and John Carroll. Robust accurate statistical annotation of general text. In *Proceedings of LREC 2002*, pages 1499–1504, Las Palmas de Gran Canaria, 2002.

[Briscoe *et al.*, 2006] Edward J. Briscoe, John Carroll, and Rebecca Watson. The second release of the RASP system. In *Proceedings of ACL-COLING 06*, Sydney, Australia, 2006.

[Bundschus *et al.*, 2008] Markus Bundschus, Mathaeus Dejori, Martin Stetter, Volker Tresp, and Hans-Peter Kriegel. Extraction of semantic biomedical relations from text using conditional random fields. *BMC Bioinformatics*, 9(207), 2008.

[Bunescu and Mooney, 2005] Razvan Bunescu and Raymond Mooney. Subsequence kernels for relation extraction. In *Proceedings of the 19th Conference on Neural Information Processing Systems*, 2005.

[Bunescu, 2003] Razvan Bunescu. Associative anaphora resolution: A web-based approach. In *Proceedings of EACL 2003 - Workshop on The Computational Treatment of Anaphora*, Budapest, 2003.

[Carbonell and Brown, 1988] Jaime Carbonell and Ralf Brown. Anaphora resolution: A multi-strategy approach. In *Proceedings of the 12th International Conference on Computational Linguistics*, pages 96–101, 1988.

[Carletta, 1996] Jean Carletta. Assessing agreement on classification tasks: The Kappa statistic. *Computational Linguistics*, 22(2):249–254, 1996.

[Castaño *et al.*, 2002] José Castaño, Jason Zhang, and James Pustejovsky. Anaphora resolution in biomedical literature. In *Proceedings of International Symposium on Reference Resolution for NLP 2002*, Alicante, Spain, 2002.

[Cohen and Hersh, 2005] Aaron M. Cohen and William R. Hersh. A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, 6(1):57–71, 2005.

[Cohen *et al.*, 2005] K. Bretonnel Cohen, Lynne Fox, Philip Ogren, and Lawrence Hunter. Corpus design for biomedical natural language processsing. In *Proceedings of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases*, Detroit, 2005.

[Cohen, 1995] W. W. Cohen. Fast effective rule induction. In *Proceedings of the 12th International Conference on Machine Learning*, pages 115–123, 1995.

[Collier *et al.*, 1999] Nigel Collier, Hyun Seok Park, Norihiro Ogata, Yuka Tateisi, Chikashi Nobata, Takeshi Sekimizu, Hisao Imai, and Jun'ichi Tsujii. The genia project. In *Proceedings of EACL'99*, pages 271–272, 1999.

[Craven and Kumlein, 1999] Mark Craven and Johan Kumlein. Constructing biological knowledge bases by extracting information from text sources. In *Proceedings of ISMB'99*, pages 77–86, 1999.

[Ding *et al.*, 2002] J. Ding, D. Berleant, D. Nettleton, and E. Wurtele. Mining medline: Abstracts, sentences, or phrases. In *Proceedings of the Pacific Symposium on Biocomputing - PSB 2002*, pages 326–337, Hawaii, 2002.

[Eilbeck and Lewis, 2004] Karen Eilbeck and Suzanna E. Lewis. Sequence ontology annotation guide. *Comparative and Functional Genomics*, 5:642–647, 2004.

[Franzén *et al.*, 2002] Kristofer Franzén, Gunnar Eriksson, Fredrik Olsson, Lars Asker Per Lidén, and Joakim Cöster. Protein names and how to find them. *International Journal of Medical Informatics*, 67(1-3):49–61, 2002.

[Friedman and Goldszmidt, 1996] Nir Friedman and Moises Goldszmidt. Building classifiers using bayesian networks. In *Porcessings of AAAI/IAAI 2006*, pages 1277–1284, 1996.

[Fundel *et al.*, 2007] Katrin Fundel, Robert Küffner, and Ralf Zimmer. RelEx – relation extraction using dependency parse trees. *Bioinformatics*, 23(3):365–371, 2007.

[Gaizauskas and Humphreys, 2000] Robert Gaizauskas and Kevin Humphreys. Quantitative evaluation of coreference algorithms in an information extraction system. In Simon Botley and Tony McEnery, editors, *Corpus-based and Computational Approaches to Discourse Anaphora*, pages 145–169. John Benjamins, Amsterdam, 2000.

[Gaizauskas *et al.*, 2000] Robert Gaizauskas, George Demetriou, and Kevin Humphreys. Term recognition and classification in biological science journal articles. In *Proceedings of the Workshop on Computational Terminology for Medical and Biological Applications of the 2nd International Conference on Natural Language Processing (NLP2000)*, pages 37–44, 2000.

[Gaizauskas *et al.*, 2003] Robert Gaizauskas, George Demetriou, Peter J. Artymiuk, and Peter Willett. Protein structures and information extraction from biological texts: the PASTA system. *Bioinformatics*, 19(1):135–143, 2003.

[Garg and Roth, 2001] Ashutosh Garg and Dan Roth. Understanding probabilistic classifiers. In Luc De Raedt and Peter A. Flach, editors, *Lecture Notes In Computer Science: Proceedings of the 12th European Conference on Machine Learning*, pages 179–191. Springer-Verlag, 2001.

[Gasperin and Briscoe, 2008] Caroline Gasperin and Ted Briscoe. Statistical anaphora resolution in biomedical texts. In *Proceedings of COLING 2008*, Manchester, UK, 2008.

[Gasperin *et al.*, 2007] Caroline Gasperin, Nikiforos Karamanis, and Ruth Seal. Annotation of anaphoric relations in biomedical full-text articles using a domain-relevant scheme. In *Proceedings of DAARC 2007*, pages 19–24, Lagos, Portugal, 2007.

[Gasperin, 2006] Caroline Gasperin. Semi-supervised anaphora resolution in biomedical texts. In *Proceedings of BioNLP'06*, New York, 2006.

[Ge *et al.*, 1998] Niyu Ge, John Hale, and Eugene Charniak. A statistical approach to anaphora resolution. In *Proceedings of the Sixth Workshop on Very Large Corpora - COLING-ACL'98*, Montreal, Canada, 1998.

[Ge, 2000] Niyu Ge. *An approach to anaphoric pronouns.* PhD thesis, Brown University, 2000.

[Grosz *et al.*, 1995] Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. Centering: a framework for modelling the local coherence of discourse. *Computational Linguistics*, 21(2):203–226, 1995.

[Grosz, 1978] Barbara J. Grosz. Focusing in dialog. In *Proceedings of the Workshop on Theoretical Issues in Natural Language Processing*, pages 96–103, Urbana-Champaign, Illinois, 1978.

[Hanisch *et al.*, 2003] D. Hanisch, J. Fluck, H. T. Mevissen, and R. Zimmer. Playing biology's name game: Identifying protein names in scientific text. In *Proceedings of the Pacific Symposium on Biocomputing - PSB 2003*, pages 403–414, Hawaii, 2003.

[Hawkins, 1978] John A. Hawkins. *Definiteness and Indefiniteness.* Humanities Press, Atlantic Highland, NJ, 1978.

[Hearst, 1992] Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics*, Nantes, 1992.

[Hirschman *et al.*, 2002] Lynette Hirschman, John Park, Junichi Tsujii, Limsoon Wong, and Cathy Wu. Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, 18(12):1553–1561, 2002.

[Hirshman and Chinchor, 1997] Lynette Hirshman and Nancy Chinchor. MUC-7 coreference task definition. In *MUC-7 Proceedings*, Online at http://www.itl.nist.gov/iaui/894.02/related_projects/muc/ proceedings/co_task.html, 1997.

[Hirst, 1981] Graeme Hirst. *Anaphora in Natural Language Understanding: A survey*, volume 119 of *Lecture Notes in Computer Science*. Springer-Verlag, 1981.

[Hobbs, 1978] Jerry Hobbs. Resolving pronoun references. *Lingua*, 44:311–338, 1978.

[Hoste, 2005] Véronique Hoste. *Optimization Issues in Machine Learning of Coreference Resolution*. PhD thesis, University of Antwerp, 2005.

[Huang *et al.*, 2004] Minlie Huang, Xiaoyan Zhu, Yu Hao, Donald G. Payan, Kunbin Qu, and Ming Li. Discovering patterns to extract protein.protein interactions from full texts. *Bioinformatics*, 20(18):3604–3612, 2004.

[Huang, 2000] Yan Huang. Discourse anaphora: Four theoretical models. *Journal of Pragmatics*, 32(2):151–176, 2000.

[Johnson *et al.*, 2007] Helen L. Johnson, Jr. William A. Baumgartner, Martin Krallinger, K. Bretonnel Cohen, and Lawrence Hunter. Corpus refactoring: a feasibility study. *Journal of Biomedical Discovery and Collaboration*, 2007.

[Karamanis *et al.*, 2007] Nikiforos Karamanis, Ian Lewin, Ruth Seal, Rachel Drysdale, and Ted Briscoe. Integrating natural language processing with flybase curation. In *Proceedings of the Pacific Symposium in Biocomputing 2007*, pages 245–256, 2007.

[Karamanis *et al.*, 2008] Nikiforos Karamanis, Ruth Seal, Ian Lewin, Peter McQuilton, Andreas Vlachos, Caroline Gasperin, Rachel Drysdale, and Ted Briscoe. Natural language processing in aid of flybase curators. *BMC Bioinformatics*, 9(193), 2008.

[Kennedy and Boguraev, 1996] Christopher Kennedy and Branimir Boguraev. Anaphora for everyone: pronominal anaphora resoluation without a parser. In *Proceedings of COLING 1996*, pages 113–118, 1996.

[Kim and Park, 2004] Jung-Jae Kim and Jong C. Park. BioAR: Anaphora resolution for relating protein names to proteome database entries. In *Proceedings of the Workshop on reference resolution and its applications - ACL 2004*, Barcelona, 2004.

[Krallinger and Hirschman, 2007] Martin Krallinger and Lynette Hirschman. *Proceedings of the Second BioCreative Challenge Evaluation Workshop, Madrid*. 2007.

[Krauthammer *et al.*, 2000] M. Krauthammer, A. Rzhetsky, P. Morozov, and C. Friedman. Using blast for identifying gene and protein names in journal articles. *Gene*, 259(1–2):245–252, 2000.

[Kulick *et al.*, 2004] Seth Kulick, Ann Bies, Mark Liberman, Mark Mandel, Ryan McDonald, Martha Palmer, Andrew Schein, and Lyle Ungar. Integrated annotation for biomedical information extraction. In *Proceedings of HLT/NAACL'2004*, 2004.

[Lappin and Leass, 1994] Shalom Lappin and Herbert Leass. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4), 1994.

[Lewin, 2007] Ian Lewin. BaseNPs that contain gene names: domain specificity and genericity. In *Proceedings of BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 163–170, 2007.

[McCarthy and Lehnert, 1995] Joseph McCarthy and Wendy Lehnert. Using decision trees for coreference resolution. In C. Mellish, editor, *Proceedings of the Fourteenth International Conference on Articial Intelligence*, pages 1050–1055, 1995.

[Meyer and Dale, 2002] Josef Meyer and Robert Dale. Using the wordnet hierarchy for associative anaphora resolution. In *Proceedings of SemaNet'02: Building and Using Semantic Networks*, Taipei, Taiwan, 2002.

[Mitkov, 1998] Ruslan Mitkov. Robust pronoun resolution with limited knowledge. In *Proceedings of the COLING 1998/ACL 1998*, pages 869–875, Montreal, 1998.

[Mitkov, 1999] Ruslan Mitkov. Anaphora resolution: The state of the art. Technical report, http://clg.wlv.ac.uk/papers/mitkov-99a.pdf, 1999.

[Morgan *et al.*, 2003] Alex Morgan, Lynette Hirschman, Alexander Yeh, and Marc Colosimo. Gene name extraction using flybase resources. In *Proceedings of ACL 2003 Workshop on Natural Language Processing in Biomedicine*, Sapporo, Japan, 2003.

[Morgan *et al.*, 2004] A. A. Morgan, L. Hirschman, M. Colosimo, A. S. Yeh, and J. B. Colombe. Gene name identification and normalization using a model organism database. *J. of Biomedical Informatics*, 37(6):396–410, 2004.

[MUC, 1995] *Proceedings of the 6th conference on Message understanding*. Morgan Kaufman Publishers, 1995.

[MUC, 1998] *Proceedings of the 7th conference on Message understanding*, http://www-nlpir.nist.gov/related_projects/muc/proceedings/muc_7_toc.html, 1998.

[Müller and Strube, 2001] Christoph Müller and Michael Strube. Annotating anaphoric and bridging expressions with MMAX. In *Proceedings of the 2nd SIGDIAL Workshop on Discourse and Dialogue*, pages 90–95, Aalborg, Denmark, 2001.

[Nelson *et al.*, 2001] Stuart J. Nelson, Douglas Johnston, and Betsy L. Humphreys. Relationships in medical subject headings. In Carol A. Bean and Rebecca Green, editors, *Relationships in the organization of knowledge*, pages 171–184. Kluwer Academic Publishers, 2001.

[Ng and Cardie, 2002a] Vincent Ng and Claire Cardie. Combining sample selection and error-driven pruning for machine learning of coreference rules. In *Proceedings of EMNLP 2002*, pages 55–62, Philadelphia, 2002.

[Ng and Cardie, 2002b] Vincent Ng and Claire Cardie. Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In *Proceedings of the 19th Coling*, Taipei, Taiwan, 2002.

[Ng and Cardie, 2002c] Vincent Ng and Claire Cardie. Improving machine learning approaches to coreference resolution. In *Proceedings of ACL 2002*, Philadelphia, 2002.

[Ng, 2003] Vincent Ng. Machine learning for coreference resolution: Recent successes and future challenges. Technical report, http://ecommons.library.cornell.edu/handle/1813/5630, Dec. 2003.

[Ng, 2004] Vincent Ng. Learning noun phrase anaphoricity to improve coreference resolution: Issues in representation and optimization. In *Proceedings of ACL 2004*, pages 152–159, Barcelona, 2004.

[Park and Kim, 2006] Jong C. Park and Jung-Jae Kim. Named entity recognition. In Sophia Ananiadou and John McNaught, editors, *Text Mining for biology and biomedicine*, pages 43–66. Artech House, 2006.

[Park *et al.*, 2001] Jong C. Park, Hyun-Sook Kim, and Jung-Jae Kim. Bidirectional incremental parsing for automatic pathway identication with combinatory categorial grammar. In *Proceedings of the Pacific Symposium on Biocomputing - PSB 2001*, 2001.

[Poesio and Vieira, 1998] Massimo Poesio and Renata Vieira. A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216, 1998.

[Poesio *et al.*, 1997] Massimo Poesio, Renata Vieira, and Simone Teufel. Resolving bridging descriptions in unrestricted texts. In *Proceedings of the Workshop on Operational Factors In Practical, Robust Anaphora Resolution for Unrestricted Texts*, Madrid, 1997.

[Poesio *et al.*, 2002] Massimo Poesio, Tomonori Ishikawa, Sabine Schulte Im Walde, and Renata Vieira. Acquiring lexical knowledge for anaphora resolution. In *Proceedings of LREC 2002*, Las Palmas De Gran Canaria, 2002.

[Poesio *et al.*, 2004] Massimo Poesio, Olga Uryupina, Renata Vieira, Mijail Alexandrov-Kabadjov, and Rodrigo Goulart. Discourse-new detectors for definite description resolution: A survey and a preliminary proposal. In *Proceedings of ACL 2004 - Workshop on Reference resolution*, Barcelona, 2004.

[Poesio, 2000] Massimo Poesio. *The GNOME Annotation Scheme Manual.* GNOME Project, http://cswww.essex.ac.uk/Research/nle/corpora/GNOME/anno_manual_4.htm, 2000.

[Poesio, 2004] Massimo Poesio. The mate/gnome proposals for anaphoric annotation, revisited. In *Proceedings of 5th SIGdial Workshop on Discourse and Dialogue*, Cambridge, MA, 2004.

[Pustejovsky *et al.*, 2002] James Pustejovsky, José Castaño, Roser Saurí, A. Rumshisky, J. Zhang, and W. Luo. Medstract: creating large-scale information servers for biomedical libraries. In *Proceedings of the ACL'02 Workshop on NLP in the biomedical domain*, pages 85–92, Philadelphia, 2002.

[Quinlan, 1993] J. Quinlan. *C4.5: Programs for machine learning.* Morgan Kaufmann, San Mateo, CA, 1993.

[Rich and LuperFoy, 1988] E. Rich and S. LuperFoy. An architecture for anaphora resolution. In *Proceedings of the Second Conference on Applied Natural Language Processing*, pages 18–24, 1988.

[Sanchez-Graillet *et al.*, 2006] O. Sanchez-Graillet, M. Poesio, M. Kabadjov, and R. Tesar. What kind of problems do protein interactions raise for anaphora resolution? - a preliminary analysis. In *Proceedings of the SMBM 2006*, Jena, 2006.

[Shen *et al.*, 2004] Dan Shen, Jie Zhang, Jian Su, Guodong Zhou, and Chew-Lim Tan. Multi-criteria-based active learning for named entity recognition. In *Proceedings of ACL 2004*, Barcelona, 2004.

[Sidner, 1979] Candace Sidner. *Towards a computational theory of definite anaphora comprehension in English discourse*. PhD thesis, MIT, 1979.

[Soon *et al.*, 2001] Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544, 2001.

[Stapley and Benoit, 2000] Benjamin J. Stapley and Gerald Benoit. Biobibliometrics: Information retrieval and visualization from co-occurrences of gene names in medline asbtracts. In *Proceedings of the Pacific Symposium on Biocomputing - PSB 2000*, pages 529–540, 2000.

[Strube and Hahn, 1999] Michael Strube and Udo Hahn. Functional centering-grounding referential coherence in information structure. *Computational Linguistics*, 25(3):309–344, 1999.

[Strube *et al.*, 2002] Michael Strube, Stefan Rapp, and Christoph Müller. The influence of minimum edit distance on reference resolution. In *Proceedings of the EMNLP 2002*, pages 312–319, Philadelphia, 2002.

[Tanabe and Wilbur, 2002] L. Tanabe and W. J. Wilbur. Tagging gene and protein names in biomedical text. *Bioinformatics*, 18(8):1124–1132, 2002.

[Tanabe *et al.*, 2005] Lorraine Tanabe, Natalie Xie, Lynne H. Thom, Wayne Matten, and W. John Wilbur. Genetag: a tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics*, 6(Suppl. 1)(S3), 2005.

[Tang *et al.*, 2002] Min Tang, Xiaoqiang Luo, and Salim Roukos. Active learning for statistical natural language parsing. In *Proceedings of ACL 2002*, pages 120–127, Philadelphia, Pennsylvania, 2002. Association for Computational Linguistics.

[Tetreault, 2001] J. Tetreault. A corpus-based evaluation of centering and pronoun resolution. *Computational Linguistics*, 27(4):507–520, 2001.

[Thompson *et al.*, 1999] Cynthia A. Thompson, Mary Elaine Califf, and Raymond J. Mooney. Active learning for natural language parsing and information extraction. In *Proceedings of the 16th International Conference on Machine Learning*, pages 406–414, 1999.

[Uryupina, 2003] Olga Uryupina. High-precision identification of discourse-new and unique noun phrases. In *Proceedings of the ACL 2003 Student Workshop*, pages 80–86, 2003.

[Uryupina, 2004] Olga Uryupina. Linguistically motivated sample selection for coreference resolution. In *Proceedings of DAARC 2004*, Furnas, 2004.

[van Deemter and Kibble, 2000] Kees van Deemter and Rodger Kibble. On coreferring: Coreference in MUC and related annotation schemes. *Computational Linguistics*, 26(4), 2000.

[Vieira and Poesio, 2000] Renata Vieira and Massimo Poesio. An empirically-based system for processing definite descriptions. *Computational Linguistics*, 26(4):525–579, 2000.

[Vieira *et al.*, 2002] Renata Vieira, Susanne Salmon-Alt, Caroline Gasperin, Emmanuel Schang, and Gabriel Othero. Coreference and anaphoric relations of demonstrative noun phrases in multilingual corpus. In *Proceedings of the DAARC 2002*, Estoril, 2002.

[Vieira, 1998] Renata Vieira. *Definite description processing in unrestricted text.* PhD thesis, University of Edinburgh, Edinburgh, 1998.

[Vilain *et al.*, 1995] Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. A model-theoretic coreference scoring scheme. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 45–52, 1995.

[Vlachos *et al.*, 2006] Andreas Vlachos, Caroline Gasperin, Ian Lewin, and Ted Briscoe. Bootstrapping the recognition and anaphoric linking of named entities in Drosophila articles. In *Proceedings of the Pacific Symposium on Biocomputing - PSB 2006*, Hawai, 2006.

[Watson *et al.*, 2003] Rebecca Watson, Judita Preiss, and Edward J. Briscoe. Contribution of domain-independent robust pronominal anaphora resolution to open-domain question-answering. In *Proceedings of the International Symposium on Reference Resolution*, Venice, Italy, 2003.

[Witten and Frank, 2005] Ian H. Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques, 2nd Edition.* Morgan Kaufmann, San Francisco, 2005.

[Yang *et al.*, 2004] X. Yang, J. Su, G. Zhou, and C. L. Tan. An NP-cluster based approach to coreference resolution. In *Proceedings of COLING 2004*, Geneva, Switzerland, 2004.

[Zweigenbaum *et al.*, 2007] Pierre Zweigenbaum, Dina Demner-Fushman, Hong Yu, and K. Bretonnel Cohen. New frontiers in biomedical text mining: Session introduction. In *Proceedings of Pacific Symposium on Biocomputing - PSB 2007*, pages 205–208, Hawaii, 2007.