ceedings of the XIII EURALEX International Congress. Vol. 1. Barcelona, 425–432.

Landau, S. (2001): Dictionaries: The art and craft of lexicography. Cambridge.

Lee, D. (2001): Genres, registers, text types, domains, and styles: clarifying the concepts and navigating a path through the BNC jungle. Language Learning & Technology 5,3, 37–72.

Moon, R. (1998): Fixed expressions and idioms in English. Oxford.

Rayson, P./Garside, R. (2000): Comparing corpora using frequency profiling. In: Proceedings of the ACL workshop on comparing corpora. Hong Kong, 1–6.

Schmid, H. (1994): Probabilistic part-of-speech tagging using decision trees. In: Proceedings of the International Conference on New Methods in Language Processing. Manchester, XX–XX.

Sinclair, J. McH. (1996): The search for units of meaning. Textus 9, 1, 75–106.

Sinclair, J. McH. (1998): The lexical item. In: Wiegand, E. (ed.), Contrastive lexical semantics. Amsterdam, 1–24.

Sinclair, J. McH. (2003): Reading concordances. Harlow.

Wible, D./Kuo, C./Chien, F./Wang, C. (2002): Toward automating a personalized concordancer for data-driven learning: A lexical difficulty filter for language learners. In: Kettemann, B./Marko, G. (eds.), Teaching and learning by doing corpus analysis. Amsterdam, 147–154.

Wiechmann, D./Fuhs, S. (2006): Concordancing software. In: Corpus Linguistics and Lingustic Theory 2,1, 107–127.

Marco Baroni, Rovereto (Italy)
Silvia Bernardini, Forlì (Italy)

103. Tools and procedures for the acquisition of morphological and syntactic information from corpora

1. Introduction

Over the past decades, the importance of the lexicon has increased in both natural language processing (NLP) and linguistic theory. Within NLP, much of the early research focused on isolated 'toy' tasks, treating the lexicon as a peripheral component. These days, the focus is on constructing systems suitable for the treatment of large, naturally occurring texts, and therefore rich lexical resources have become crucial for NLP systems dealing with real-world applications. At the same time, the importance of the lexicon has increased for theoretical reasons as within many linguistic theories, it has taken on an increasingly central role in the description of both idiosyncratic and regular properties of language.

Obtaining large, explicit lexicons rich enough for computational and linguistic use has, however, proved challenging. Manual construction of a large-scale lexicon is a major task involving many years of lexicographic work. Manual work is costly, the resulting resources require extensive labour-intensive porting to new tasks, and they often lack information important for current NLP techniques, in particular statistical information about the frequencies of lexical items in data.

Automatic acquisition or updating of lexical information from relevant repositories of text (such as the web and corpora of published text) is a more promising avenue to pursue long term. Although the approach is challenging, it is now viable, cost-effective and gathers statistical information as a side-effect of the acquisition process. Statistical information can easily be adapted to new domains and usage patterns provided relevant corpus data is available. The resources and techniques required for the automatic approach are now available. Several large corpora have been constructed for many languages, along with the web, which is particularly helpful in overcoming the problem of data sparseness, severe even in the largest

corpora (Keller et al. 2002). Methods for automatic text analysis and machine learning have also now developed to the point that they can be usefully deployed.

Methods for automatic lexical acquisition have been developed for many areas since the past decade or two, starting from collocations (Dunning 1993), word senses (Pereira et al. 1993), prepositional phrase attachment ambiguity (Hindle/Rooth 1993), word semantic classes (Zernik 1989), selectional preferences (Resnik 1993), diathesis alternations (McCarthy 2001), subcategorization (e.g. Brent 1991, 1993), and multiword expressions (Sag et al. 2003). Many of the methods are still under development and need further refinement before they can successfully be applied to large-scale lexical acquisition. However, some are now sufficiently developed that they are starting to produce large-scale lexical resources, which include frequency and usage information tuned to genres and sub-languages. These include in particular those, which acquire syntactic information from corpora.

Basic lexico-syntactic information, subcategorization, is of particular importance for both NLP and linguistic work. Access to an accurate and comprehensive subcategorization lexicon is vital for the development of successful parsing technology (Carroll et al. 1998), important for many NLP tasks such as automatic verb classification (Schulte im Walde 2006), useful for any applications which can benefit from information about predicate-argument structure (e.g. Information Extraction (IE) (Surdeanu et al. 2003)) and highly interesting for empirically based theoretical research on language, e.g. psycholinguistic research on sentence processing (Lapata et al. 2001) and linguistic research in child language acquisition (Buttery/Korhonen 2007).

In this article we survey tools and procedures for subcategorization acquisition in particular. We start by discussing the definition of subcategorization in section 2. The general process of automatic subcategorization acquisition is described in section 3, together with the different methods proposed so far, and their evaluation and performance. Section 4 summarises the state-of-the-art and discusses directions for future work. Section 5 concludes by describing the first large-scale public-domain subcategorization lexicon that has been built automatically for English.

2. What is subcategorization?

Subcategorization aims to capture the diverse behaviour of words – the fact that different subcategories of words make different demands on their arguments. For example, the English verb put requires 3 syntactic arguments (1a) and doesn't permit any fewer (1b–1d):

(1) a. Sam put the book on the table
b. *Sam put the book
c. *Sam put on the table
d. *Sam put

Subcategorization concerns arguments of a predicate. These may be either obligatory or optional, in which case they should be separated from adjuncts. While arguments are closely associated with the predicate and understood to complete its meaning (2a), adjuncts are understood to complete the meaning of the central predication as a whole (2b).

(2) a. He ate chocolate
b. He sat eating chocolate

From the perspective of the linguistic theory, argument-adjunct distinction is only a partially solved problem, challenged by many individual cases, which tend to fall in the grey area between arguments and adjuncts. The work on automatic subcategorization acquisition has had to adopt a simplified, empirically based approach to argument-adjunct distinction which resembles that adopted by the COMLEX Syntax lexicographers (Grishman et al. 1994). The COMLEX lexicographers have demonstrated that the problematic grey area cases aside, arguments can be distinguished fairly accurately from adjuncts using five criteria and five heuristics for argument-hood and six criteria and two heuristics for adjunct-hood. These criteria and heuristics are culled mostly from the linguistics literature and supplemented with rough generalizations. For example, they state that NPs, PPs headed by to, and finite clauses without gaps tend to be arguments, while purpose clauses, PPs and ADVPs expressing place, time and manner are usually adjuncts. They also advise that an argument usually occurs with the verb at significantly higher frequency than with most other verbs, while an adjunct occurs with a large variety of verbs with roughly the same frequency and meaning. Conflicts between the criteria are resolved in various ways. For example, the complement-hood criteria override the adjunct-hood criteria in all but a few well-defined cases, a

single complement-hood criterion warrants argument analysis, and so forth.

Given argument-adjunct distinction, subcategorization concerns the specification, for a predicate, the number and type of arguments which it requires for well-formedness. Subcategorization structures are frequently characterized in terms of syntactic frames called subcategorization frames (SCFs). These provide generalization over various syntactic contexts required by verbs associated with the same syntactic behaviour. For example, we can use the frame NP+PP to characterize the subcategorization structure in (1a), as well as those in Sam put the book on the table yesterday and John flew the plane to Rome. More or less specific SCF classifications can be made, depending e.g. on whether the frames are parameterized for lexically-governed particles and prepositions, whether any semantic knowledge is incorporated, and so forth. For example, the fairly detailed SCF classification proposed by Briscoe (2000) incorporates as many as 168 SCF distinctions for English. It abstracts over specific lexically-governed particles and prepositions and specific predicate selectional preferences, but includes some semi-productive bounded dependency constructions, such as particle and dative movement.

Fully to define the association between a particular subcategorization structure and a given predicate, however, one must go beyond listing of syntactic frames. Full account of subcategorization requires specifying the number and type of arguments that a particular predicate requires, predicate sense in question, semantic representation of the particular predicate-argument structure, mapping between the syntactic and semantic levels of representation, semantic selectional restrictions or preferences on arguments, control of understood arguments in predicative complements, diathesis alternations, and possibly also further details of predicate-argument structure.

While the ultimate goal of automatic subcategorization acquisition is the acquisition of all this information form corpus data, the work conducted so far has focussed on what can be extracted from corpora relatively reliably now using the state-of-the-art NLP technology: the syntactic SCFs associated with different predicates and their frequency in corpus data.

3. Automatic subcategorization acquisition

3.1. Resources

Systems exist for acquiring SCFs from both unannotated and manually annotated corpora (e.g. treebanks). Employing fully accurate input data the latter type of systems (e.g. Sarkar/Zeman 2000, O'Donovan et al. 2005) yield very promising results. However, we focus here on the former type of systems because not being dependent on the availability of manually annotated corpora, they can be applied to any domain and task, provided adequate corpus data is available. Assuming this approach, subcategorization acquisition requires the following type of language resources:

- *Corpora*: Large collections of naturally occurring language (e.g. text corpora or the Web) are required for training and testing purposes. Subcategorization is a typical natural language phenomenon in the sense that it shows a Zipfian distribution (Korhonen, 2002): only the most frequent SCFs are very frequent, while most SCFs are extremely infrequent in language. Adequate experimentation and construction of a comprehensive SCF lexicon requires therefore access to a large corpus. The number of corpus citations required for a single predicate depends on the nature of a predicate. Korhonen (2002) has estimated that for a predicate taking multiple SCFs c. 250 corpus citations at minimum is required for sufficient performance with subcategorization acquisition. For a predicate taking fewer SCFs less data may suffice. Given the distribution of predicates themselves is Zipfian too, this requires access to large data.
- *Lexical Resources or Grammars*: Building a classifier for SCFs and/or evaluating its output requires basic knowledge of the grammar of the language in question, in particular concerning the type of arguments predicates tend to take and the way they realize syntactically in the language. Where manually built syntax dictionaries are available (e.g. the ANLT (Boguraev et al. 1987) and COMLEX (Grishman et al. 1994) dictionaries for English) subcategorization acquisition work typically exploits these as a starting point.
- *NLP tools*: Although it is possible to acquire some SCF patterns from raw corpus data, building a comprehensive classifier and lexicon requires access to basic text processing tools which can analyse the grammatical category of words and the syntactic context where they occur. In fact, the more accurate tagging, lemmatizing and parsing tools are available for a language, the better the results typically with subcategorization acquisition (see the below section 3.2 for examples and discussion on this).

3.2. Methods

The first systems capable of automatically learning a small number of verbal subcategorization frames (SCFs) from unannotated English corpora emerged over a decade ago (Brent 1991 and 1993, Manning 1993). Subsequent research has yielded systems for English (Carroll/Rooth 1998, Briscoe/Carroll 1997) capable of detecting comprehensive sets of SCFs with promising accuracy and demonstrated success in application tasks (Carroll et al. 1998, Korhonen et al. 2003, Buttery/Korhonen 2007), the most ambitious of which deal not only with verbs but also with nouns and adjectives (Preiss et al. 2007). During the past years, systems have been proposed for languages other than English, including German (Schulte im Walde 2006), Spanish (Esteve Ferrer 2004), Portuguese (Gamallo et al. 2003), Greek (Maragoudakis et al. 2001), Japanese (Kawahara/Kurohashi 2002), Chinese (Han/Zhao 2006), Bengali (Banerjee et al. 2009), Polish (Debowski 2009), Italian (Ienco et al. 2008) and French (Chesley/Salmon-Alt 2006, Messient 2008), among others.

Various methods of subcategorization acquisition share a common objective: given corpus data, to identify verbal predicates in this data and record the type and/or number of SCFs taken by these predicates. The systems vary according to how SCFs are defined (depending on the assumed definition of the argument-adjunct distinction, and whether the SCFs are parameterized for prepositions and particles) and whether the SCFs are prespecified or learned from data. However, regardless of these differences, the systems typically proceed in two steps, by

- generating hypotheses for SCFs, and
- selecting reliable hypotheses for the final lexicon.

For hypothesis generation (i), most systems first tag and/or parse ambiguous corpus data (i.e. no word sense disambiguation is typically performed) using a chunker, a partial parser, or a full parser which returns intermediate analyses. They then identify basic SCF patterns in the resulting data (verbs and their local context), and classify them to SCFs on the basis of hand written or automatically generated rules. Because many criteria for argument-adjunct distinction rest on judgements which cannot (yet) be made automatically and since no lexical information is typically used during parsing (because this is the information we aim to acquire) the output from hypothesis generator is inevitably noisy, containing many incorrect SCFs.

Although some systems treat hypothesised SCFs as absolute SCF indicators, most treat them as probabilistic indicators. The latter systems typically employ a separate filtering component which performs hypothesis selection (ii). During hypothesis selection reliable SCFs are selected for the final lexicon. It is typically performed using either statistical hypothesis tests (e.g. the binomial hypothesis test, long likelihood ratio, t-test) or simple thresholding on the frequencies or relatively frequencies of SCFs. For better accuracy, smoothing may be performed prior to filtering which corrects the automatically acquired SCF distribution with probability estimates based e.g. on the likely semantic class of a predicate. For maximum accuracy (in well-defined tasks), one may choose to use already known SCFs e.g. from a dictionary and just use the system to associate corpus-based frequencies with the known SCFs.

The need for hypothesis selection and the optimal method for it depends on the intended use of the lexicon. For example, if the aim is to use SCF frequencies to aid parsing, a user may want to maximise the accuracy (rather than the coverage) of the lexicon, while some tasks may benefit from a lexicon which provides good coverage at the expense of accuracy. For example, Sun and Korhonen (2009) obtained the best results with automatic verb classification when using an unfiltered SCF lexicon as input data. In this case noisy SCFs contained information useful for the task (e.g. adjuncts).

To provide a more concrete picture of the process of subcategorization acquisition, we will survey the different approaches so far proposed for this work. Our survey will mostly focus on English since the approaches proposed for other languages have essentially been language-specific adaptations of well-known English techniques, i.e. their core methodology is very similar to that employed for English. We divide various SCF acquisition methods into three groups which we discuss in the subsequent sections. This grouping reflects chronological development from preliminary systems capable of acquiring only a small number of SCFs towards more ambitious systems suitable for large-scale subcategorization acquisition. It also shows how methods have developed with respect to the different factors mentioned above.

3.2.1. Preliminary work

Work on automatic subcategorization extraction was initiated by Brent (1991, 1993) who proposed a preliminary method for acquiring just six SCFs from corpus data. The set of SCFs targeted was manually composed and restricted to those involving basic NP, sentential and infinitival phrases. Brent's purpose was only to exploit unambiguous and determinate information in raw (un-tagged) corpora. A number of lexical cues was defined in his approach, mostly involving closed class items, which reliably cue verbs and SCFs. Although only highly reliable cues are used, the correspondence between cues and syntactic structure is still not perfect, and the output of the hypothesis generator contains some noise. A filter based on the binomial hypothesis test (BHT) is employed which calculates (i) the overall error probability that a particular SCF will be hypothesised and (ii) the amount of evidence for an association of that SCF with the verb in question to decide which hypotheses are reliable enough to warrant a conclusion.

The main problem with Brent's approach is that it generates high accuracy hypotheses at the expense of coverage. Reliant on raw corpus data, the method is dependent on lexical cues. However, for many verbs and SCFs, no such cues exist. For example, some verbs subcategorize for the preposition in (e.g. *They assist the police in the investigation*), but the majority of occurrences of 'in' after a verb are NP modifiers or non-subcategorized locative phrases (e.g. *He built a house in the woods*). Thus the approach is not extendable to all SCFs and at any rate leads to ignoring a great deal of information potentially available. Use of only unambiguous data means that corpus analysis will be incomplete and no accurate frequency information can be gathered.

3.2.2. Further developments

Given the problems of Brent's method, subsequent approaches to SCF acquisition have opted to seek evidence from all examples in corpus data. This has necessitated the use of annotated input data. The approach has been to extract POS tags from corpora and chunk the POS tagged data into non-recursive cores of major phrases, e.g. verb groups, bare unpostmodified NPs, PPs and so forth. Essentially, chunking allows factoring data into those pieces of structure which can be recovered without knowledge of the phenomena that we are trying to acquire (i.e. SCFs).

Ushioda et al. (1993), Manning (1993) and Gahl (1998) represent chunking-based SCF acquisition. They all opt for partial parsing via finite state regular expression pattern matching. Parsing is deterministic, and ambiguities in analysis are typically solved using the longest match heuristic: if there are two possible parses that can be produced for the same substring, the parser chooses the longer match. SCF recognition is usually aided by the use of a small number of lexical cues. Ushioda et al. (1993) applies nine SCF extraction rules to the chunked sentences (written as regular expressions and obtained through examination of occurrences of verbs in a training text) in order to recognise six SCF types. Manning (1993) proposes a similar but more ambitious system capable of recognizing 19 distinct SCFs. Hypotheses are evaluated and filtered by a modified version of the BHT employed by Brent. Gahl's (1998) work differs from Ushioda's and Manning's in that she performs SCF acquisition in the context of a corpus query system. Gahl presents an extraction tool for use with the British National Corpus (BNC) (Leech 1992) which she uses to create subcorpora containing different SCFs for verbs, nouns and adjectives, given the frames expected for each predicate A user has the choice of 27 searchable SCFs, based on a selection of those occurring in the COMLEX syntax dictionary. No filtering for hypothesis selection is used in her work.

Extracting SCF information from chunked data increases the number of cues available and allows also for low reliability cues. Running in linear time, partial parsing is a quick way to seed the SCF acquisition process with some a priori grammatical knowledge. The disadvantage, however, is the high level of noise in output, caused by the limitations of partial parsing and the inadequacy of the longest match heuristic.

3.2.3. Towards large-scale subcategorization acquisition

Subsequent work on SCF acquisition has opted for more knowledge-based hypothesis generation. Instead of acquiring SCFs from partially parsed data, recent systems have acquired this information from data parsed using an intermediate parser. Rather than simply chunking the input, an intermediate parser produces singly rooted trees which re-

quire global coherence from syntax and therefore impose greater grammatical constraint on analysis. The parsers used have been probabilistic.

Carroll and Rooth (1998) introduce a technique based on a robust statistical parser and automatic tuning of the probability parameters of the grammar. They use an iterative approach to estimate the distribution of SCFs given head words, starting from a hand-written headed context-free grammar (CFG) whose core is a grammar of chunks and phrases. A probabilistic version of this grammar is first trained from a POS-tagged corpus using an unsupervised machine learning technique. Lexicalised event counts (frequency of a head word accompanied by a SCF) are collected, the probabilistic CFG is lexicalised on rule heads, after which the machine learning algorithm is run iteratively, and finally the conditional probability estimates for verb and SCF combinations from all the runs are combined. Carroll and Rooth target 15 SCFs. Carroll and Rooth do not employ filtering for hypothesis selection, but include all hypotheses generated in the final lexicon.

Briscoe and Carroll (1997) describe a system capable of categorizing 161 different SCFs. This comprehensive set of SCFs was obtained by merging the SCFs classifications of the ANLT and COMLEX dictionaries and manually adding into this set new SCFs discovered from the corpus data. While the previous approaches to SCFs acquisition employ only syntactic SCFs, Briscoe and Carroll's frames also incorporate semantic information (e.g. about control of predicative arguments). Their system parses data with a robust statistical parser which uses a shallow tag sequence grammar written in a unification based grammar formalism. The parser yields intermediate phrase structure analyses. Local syntactic frames are then extracted from the parsed data (including the syntactic categories and head lemmas of constituents) from sentence subanalyses which begin/end at the boundaries of specified predicates. The resulting extracted subcategorization patterns are then classified as SCFs on the basis of the feature values of syntactic categories and the head lemmas in each pattern. Briscoe and Carroll employ BHT for hypothesis selection, refining it with a priori estimates of the probability of membership in different SCFs.

Building on the basic approach of Briscoe and Carroll (1997), Preiss et al. (2007) have proposed an improved and extended system which exploits the Robust Accurate Statistical Parsing (RASP) system (Briscoe/Carroll 2002). This more recent system is the first system for English which can be used to acquire comprehensive lexicons for verbs, nouns and adjectives. It incorporates a rule-based classifier which identifies 168 verbal, 37 adjectival and 31 nominal frames from grammatical relations output by a RASP. The rules employed are similar to those employed by Briscoe and Carroll (1997), with the difference that they operate on grammatical relations rather than parse trees.

For reasons given earlier, employing probabilistic parsing in SCF acquisition is an improvement over the use of partial parsing and the longest match heuristic. In sum we may say that, while the early work minimised noise at the expense of coverage (both in terms of SCFs and data), the follow-up work maximised coverage at the expense of accuracy, and recent work has aimed to maximise both coverage and accuracy. However, at the present state of development, most intermediate parsers still yield fairly noisy output, mainly due to the lack of lexical and semantic information during parsing. As the output from the hypothesis generator is noisy, filtering is needed when aiming for a high accuracy lexicon. Hypothesis selection techniques adopted by recent approaches are similar to those selected in early work – most approaches employ BHT, as originally introduced by Brent (1993). Although different modifications to this test have been proposed, many early and recent approaches report unreliable performance, especially with low frequency SCFs. Although simple thresholding based on the relative frequencies of SCFs may perform better (e.g. Preiss et al. 2007) it is not an ideal solution because it is based on ignoring low frequency SCFs.

Korhonen (2002) has addressed this problem by developing a hypothesis selection component to be used in conjunction with the systems of Briscoe and Carroll (1997) and Preiss et al. (2007) which doesn't employ any hypothesis testing, but which involves smoothing noisy SCF distributions from the system's hypothesis generator with back-off estimates based on lexical-semantic classes of verbs. For example, in this approach the verb 'fly' is classified in the class of MOTION verbs (Levin 1993), and the noisy SCF distri-

bution for the verb 'fly' is smoothed using a prototypical SCF distribution for MOTION verbs (constructed by merging the SCF distributions of a few other MOTION verbs). This method, when combined with simple filtering (thresholding on the probability estimates from smoothing), improves subcategorization acquisition performance significantly since it helps to correct the automatically acquired distributions with semantic information and also helps to deal with sparse data. It can currently only be applied to verbs, but could be easily extended to nouns and adjectives as well. Wide and/or domain-specific application of this method would require, however, an automatic method for classifying verbs to lexical-semantic classes. Although such methods are under development (e.g. Sun/Korhonen 2009) they require further research before they can be successfully incorporated as part of automatic subcategorization acquisition.

3.3. Evaluation

3.3.1. Measures

SCF acquisition systems are typically evaluated in terms of types or tokens. Types are the set of SCFs acquired. Type-based evaluation involves assessment of the lexical entries in a lexicon. It is usually performed on unseen test data, with a number of randomly selected test verbs. The SCF types acquired are compared with those found in some gold standard. The gold standard is usually obtained either through manual analysis of corpus data, or from lexical entries in a large dictionary. Both approaches have their advantages and disadvantages. Manual construction of a gold standard is time-consuming, but yields an accurate measure when obtained from the data that the system used to acquire the entries. Meanwhile, obtaining a gold standard from a dictionary is fast, but the resulting resource may not be relevant for the test data. This is because dictionaries may contain SCFs absent from the corpus data or lack SCFs present in the corpus data.

Tokens are the individual occurrences of SCFs in corpus data. They are evaluated against manually analysed corpus tokens. Evaluation may be performed on the corpus data from which the acquired SCFs were obtained, to estimate the coverage of the training data, i.e. the coverage of the lexicon the system has learned. This indicates e.g. an estimate of the parsing performance that would result from providing a parser with the SCFs acquired. Alternatively, token-based evaluation may be performed on a different corpus to examine how well the acquired information generalizes.

Evaluation is frequently performed using precision and recall. Obtaining these measures requires recording the number of

- true positives (TPs) – correct SCF types or tokens proposed by the system
- false positives (FPs) – incorrect SCF types or tokens proposed by the system
- false negatives (FNs) – correct SCF types or tokens not proposed by the system.

When evaluating SCF information, precision and recall are usually reported over types. Type precision is the percentage of SCFs that the system proposes which are correct (in the gold standard), while type recall is the percentage of SCFs in the gold standard that the system proposes. F-measure combines precision and recall into a single measure of overall performance.

Some methods have been evaluated using measures of distributional similarity (e.g. KL distance, entropy, skew divergence), rank correlation and intersection to compare automatically acquired SCF distributions against gold standard distributions. Such evaluation can offer a good insight into the accuracy of corpus-based frequencies.

The methods listed so far are used for evaluating SCF acquisition in its own context. However, it is generally agreed that the ultimate demonstration of success is improved performance on an application task. Task-based evaluation may be done, for instance, by examining application performance with and without integrating the SCFs information, and seeing how much the integrated information improves performance.

3.3.2. State-of-the-art performance

When examining the performance of SCF acquisition systems, one should remember that they differ in various ways. Variation in the number of target SCFs, test verbs, gold standards, and the size of test data make direct comparison of different results difficult. However, sampling the results of systems is useful as it reveals the upper limits of performance.

Carroll and Rooth (1998) evaluated their system (targeting 15 SCFs) on a set of 100 test verbs using data from the BNC. In their evaluation against SCFs listed in a dictionary they obtained 79% type precision and 75% type recall.

Briscoe and Carroll (1997) evaluated their much more ambitious system (targeting 161 SCFs) on a set of 7 test verbs using data extracted from the Susanne corpus. They obtained 77% type precision and 43% type recall against a gold standard which was constructed by manually analysing SCFs in test sentences. The low type recall indicates that many of the additional SCFs in their classifier are low in frequency and hard to detect.

Korhonen (2002) evaluated the system of Briscoe and Carroll (1997) so that she replaced the BHT filter with the semantically-motivated smoothing approach for hypothesis selection combined with simple thresholding for filtering. She obtained 87.1% type precision and 71.2% type recall on a set of 45 test verbs against a gold standard obtained via manual analysis of corpus data. The improved recall indicates that the smoothing approach is effective in dealing with sparse data.

Preiss et al. (2007) evaluated their system targeting 168 verbal, 37 adjectival and 31 nominal frames on a set of 183 verbs, 30 nouns and 30 adjectives using data extracted from the BNC. The results when using simple thresholding method for filtering (i.e. not the smoothing method of Korhonen (2002)) were 81.8% type precision and 59.5% type recall for verbs, 91.2% type precision and 47.2% type recall for nouns, and were 95.5% type precision and 57.6% type recall for adjectives. The noun and adjective classifiers yield high precision compared to recall because many of their SCFs in their gold standard are extremely low in frequency.

From these approaches, only Briscoe and Carroll (1997) have evaluated token recall (81%) and reported task-based evaluation which shows that the acquired SCF frequencies improve parsing. SCF data extracted using the system of Briscoe and Carroll (or its later developments) has been subsequently shown to aid lexical classification, selectional preference acquisition and empirical linguistic research (McCarthy 2001, Buttery/Korhonen 2007, Sun/Korhonen 2009).

4. Future work

While the results achieved with current systems are generally encouraging, the accuracy of the resulting lexicons shows room for improvement. Errors arise in automatic SCF acquisition for several reasons. Due to ungrammaticalities of natural language, some noise already occurs in input data. Further errors arise when processing the data through different phases of hypothesis generation and selection.

With hypothesis generation, the most frequently reported error is the inability of a system properly to distinguish between arguments and adjuncts. This makes detection of SCFs involving PPs especially difficult. Although one can make simple assumptions, for instance, that arguments of specific verbs tend to occur with greater frequency in potential argument positions than adjuncts, problems arise when the judgments of argument-adjunct distinction require a deeper analysis. Many argument-adjunct tests cannot yet be exploited automatically since they rest on semantic judgments that cannot yet be made automatically. One example is the syntactic tests involving diathesis alternation possibilities which require recognition that the same argument occurs in different argument positions. Recognizing identical or similar arguments requires considerable quantities of lexical data or the ability to back-off to lexical semantic classes.

In fact, there is a limit to how far we can get with subcategorization acquisition merely by exploiting syntactic information. As Briscoe and Carroll (1997) point out, the ability to recognize that argument slots of different SCFs for the same predicate share selectional restrictions / preferences would assist recognition that the predicate undergoes specific diathesis alternations. This in turn would assist inferences about control, equi, and raising, enabling finer-grained SCF classifications and yielding a more comprehensive subcategorization dictionary (Boguraev/Briscoe 1987). In the end, any adequate subcategorization dictionary needs to be supplemented with information on semantic selectional preferences / restrictions and diathesis alternations to provide a full account of subcategorization and to be useful as a lexical resource. The obvious way to further develop automatic subcategorization acquisition is therefore to work on the acquisition of lexical-semantic information from corpora and to incorporate semantic acquisition to support syntactic acquisition (and vice versa).

5. A large automatically acquired lexicon

While research into further improving the systems will continue, the state of the art has already developed to the point where the best existing systems are capable of detecting

comprehensive SCF (frequency) information with accuracy high enough to benefit practical NLP tasks. Although no complete SCF acquisition tools are (currently) publicly available which would enable the creation of lexicons from scratch, Korhonen et al. (2006) has recently made the first automatically acquired large-scale subcategorization lexicon for English verbs available together with hypothesis selection software which can be used to create sub-lexicons for various use. This VALEX (http://www.cl.cam.ac.uk/~alk23/subcat/lexicon.html) lexicon was acquired from five corpora and the Web using Briscoe and Carroll's (1997) system enhanced with different hypothesis selection options. It provides SCF frequency information for 6,397 (American and British) English verbs.

The lexicon includes a lexical entry for each verb and SCF combination found in corpus data: 212,741 entries in total, and 33 per verb on average. A lexical entry specifies (at minimum) the verb and the SCF in question, the syntax of detected arguments, the raw and relative frequencies of the SCF given the verb, the POS tags of the verb tokens, the argument heads in different argument positions, and the frequency of possible lexical rules (e.g. the passive rule) applied during parsing. This information stored in the lexical entries has proved useful for various tasks, including parsing (Carroll et al. 1998), lexical classification (Sun/Korhonen 2009) and the acquisition of selectional preferences and diathesis alternations (McCarthy 2001).

The web distribution of VALEX provides the following materials:

- The description of the 163 SCF types in the lexicon
- The large automatically acquired (unfiltered, noisy) subcategorization lexicon
- Software which can be used to filter out noisy SCFs from the large lexicon, improve the quality of automatically acquired SCF distributions, and build sub-lexicons suitable for different purposes
- Four sub-lexicons created using the software which are more accurate than the basic noisy lexicon and which can be readily employed by users who prefer not to run the software themselves
- Documentation which explains the different sub-lexicon options provided by the software and evaluates their accuracy

6. Selected bibliography

Banerjee, S./Das, D./Bandyopadhyay, S. (2009): Bengali Verb Subcategorization Frame Acquisition − A Baseline Model. In: Proceedings of the 7th Workshop on Asian Language Resources, Suntec City, Singapore, 76−83.

Boguraev, B./Briscoe, E. J. (eds.) (1987): Computational Lexicography for Natural Language Processing. London.

Boguraev, B./Briscoe, E. J./Carroll, J./Carter, D./Grover, C. (1987): The derivation of a grammatically-indexed lexicon from the Longman Dictionary of Contemporary English. In: Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics, Stanford, CA, 193−200.

Brent, M. (1991): Automatic acquisition of subcategorization frames from untagged text. In: Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics. Berkeley, CA, 209−214.

Brent, M. (1993): From grammar to lexicon: unsupervised learning of lexical syntax. Computational Linguistics 19.3, 243−262.

Briscoe, E. J. (2000): Dictionary and system subcategorisation code mappings. Unpublished manuscript, University of Cambridge: Computer Laboratory.

Briscoe, E. J./Carroll, J. (1997): Automatic extraction of subcategorization from corpora. In: Proceedings of the 5th ACL Conference on Applied Natural Language Processing. Washington, D. C., 356−363.

Briscoe, E. J./Carroll, J. (2002): Robust accurate statistical annotation of general text. In: Proceedings of the 3rd International Conference on Language Resources and Evaluation. Gran Canaria, Spain, 1499−1504.

Buttery, P./Korhonen, A. (2007): I will shoot your shopping down and you can shoot all my tins − Automatic Lexical Acquisition from the CHILDES Database. In: Proceedings of the ACL 2007 Workshop on Cognitive Aspects of Computational Language Acquisition. Prague, Czech Republic, XX−XX.

Carroll, J./Minnen, G./Briscoe, E. J. (1998): Can subcategorisation probabilities help a statistical parser? In: Proceedings of the 6th ACL/SIGDAT Workshop on Very Large Corpora. Montreal, Canada, 118−126.

Carroll, G./Rooth, M. (1998): Valence induction with a head-lexicalized PCFG. In: Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing. Granada, Spain.

Chesley, P./Salmon-Alt, S. (2006): Automatic extraction of subcategorization frames for French. In: Proceedings of Language Resources and Evaluation Conference. Genoa, Italy, XX-Xx.

Debowski, L. (2009): Valence extraction using EM selection and co-occurrence matrices. In: Language Resources and Evaluation, 43(4), 301–327.

Dunning, T. (1993): Accurate methods for the statistics of surprise and coincidence. In: Computational Linguistics 19(1), 61–74.

Esteve Ferrer, E. (2004): Towards a semantic classification of Spanish verbs based on subcategorisation information. In: Proceedings of the ACL 2004 workshop on Student research. Barcelona, Spain, XX–XX.

Gahl, S. (1998): Automatic extraction of subcorpora based on subcategorization frames from a part-of-speech tagged corpus. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics. Montreal, Canada, 428–432.

Gamallo, P./Agustini, A./Lopes, G. P. (2003): Learning Subcategorisation Information to Model a Grammar with Co-Restrictions. In: Traitement Automatique de la Langue 44,1, 93–118.

Grishman, R./Macleod, C./Meyers, A. (1994): Complex syntax: Building a Computational Lexicon. In: Proceedings of the International Conference on Computational Linguistics. Kyoto, Japan, 268–272.

Han, X./Zhao, T. (2006): Two-Fold Filtering for Chinese Subcategorization Acquisition with Diathesis Alternations Used as Heuristic Information. In: Computational Linguistics and Chinese Language Processing 11,2, 101–114.

Hindle, D./Rooth, M. (1993): Structural ambiguity and lexical relations. In: Computational Linguistics 19(2), 103–120.

Ienco, D./ Villata, S./ Bosco, C. (2008): Automatic extraction of subcategorization frames for Italian. In: Proceedings of LREC, 2094–2100.

Kawahara, D./Kurohashi, S. (2002): Fertilization of case frame dictionary for robust Japanese case analysis. In: Proceedings of the 19th International Conference on Computational Linguistics. Taipei, Taiwan, 425–431.

Keller, F./Lapata, M./Ourioupina, O. (2002): Using the web to overcome data sparseness. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Philadelphia, PA, 230–237.

Korhonen, A. (2002): Subcategorization Acquisition. Ph.D. thesis, University of Cambridge, England.

Korhonen, A./Krymolowski, Y./Marx, Z. (2003): Clustering Polysemic Subcategorization Frame Distributions Semantically. In: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics. Sapporo, Japan, 64–71.

Korhonen, A./Krymolowski, Y./Briscoe, E. J. (2006): A Large Subcategorization Lexicon for Natural Language Processing Applications. In: Proceedings of the 5th international conference on Language Resources and Evaluation. Genova, Italy, XX–XX.

Lapata, M./Keller, F./Schulte im Walde, S. (2001): Verb frame frequency as a predictor of verb bias. In: Journal of Psycholinguistic Research 30(4), 419–435.

Leech, G. (1992): 100 million words of English: the British National Corpus. Language Research 28 (1), 1–13.

Levin, B. (1993): English Verb Classes and Alternations. Chicago University Press, Chicago.

Manning, C. (1993): Automatic acquisition of a large subcategorization dictionary from corpora. In: Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics. Columbus, Ohio, 235–242.

Maragoudakis, M./Kermanidis, K./Fakotakis, N./ Kokkinakis, G. (2001): Learning Automatic Acquisition of Subcategorization Frames using Bayesian Inference and Support Vector Machines. In: Proceedings of the IEEE International Conference on Data Mining. San José, California, XX–XX.

McCarthy, D. (2001): Lexical Acquisition at the Syntax-Semantics Interface: Diathesis Alternations, Subcategorization Frames and Selectional Preferences. PhD thesis, University of Sussex, England.

Messient, C. (2008): A Subcategorization System for French Verbs. In: Proceedings of Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Student Research Workshop. Place, 55–60.

O'Donovan, R./Burke, M./Cahill, A./van Genabith, J./Way, A. (2005): Large-Scale Induction and Evaluation of Lexical Resources from the Penn-II and Penn-III Treebanks. In: Computational Linguistics 31(3), 328–365.

Pereira, F./Tishby, N./Lee, L. (1993): Distributional clustering of English words. In: Proceedings of the 31st Annual Meeting of the Association of Computational Linguistics. Columbus, Ohio, 183–190.

Preiss, J./Briscoe, E. J./Korhonen, A. (2007): A System for Large-scale Acquisition of Verbal, Nominal and Adjectival Subcategorization Frames

from Corpora. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics. Prague, Czech Republic, XX–XX.

Resnik, P. (1993): Selection and Information: A Class-Based Approach to Lexical Relationships. PhD thesis, University of Pennsylvania, PA.

Sag, I. A./Baldwin, T./Bond, F./Copestake, A./ Flickinger, D. (2002):

Multiword Expressions: A Pain in the Neck for NLP. In: Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics. Mexico City, Mexico, XX–XX.

Sarkar, A./Zeman, D. (2000): Automatic extraction of subcategorization frames for Czech. In: Proceedings of the 19th International Conference on Computational Linguistics. Saarbrücken, Germany, 691–697.

Schulte im Walde, S. (2006): Experiments on the automatic induction of german semantic verb classes. In: Computational Linguistics 32(2), 159–194.

Sun, L./Korhonen, A. (2009): Improving Verb Clustering with Automatically Acquired Selectional Preferences. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Suntec. Singapore, 638–647.

Surdeanu, M./Harabagiu, S./Williams, J./Aarseth, P. (2003): Using predicate-argument structures for information extraction. In: Proceedings of the 41st Annual Meeting of ACL. Sapporo, Japan, XX–XX.

Ushioda, A./Evans, D./Gibson, T./Waibel, A. (1993): The automatic acquisition of frequencies of verb subcategorization frames from tagged corpora. In: Boguraev, B./Pustejovsky, J. (eds.), SIGLEX ACL Workshop on the Acquisition of Lexical Knowledge from Text. Columbus, OH, 95–106.

Zernik, U. (1989): Lexicon Acquisition: Learning from Corpus by Capitalizing on Lexical Categories. In: Proceedings of Lexical Categories. Place, 1556–1562.

Anna Korhonen, Cambridge (UK)

105. Tools for lexicographic use of parallel and comparable corpora

1. Introduction

As already mentioned in Teubert (1996): "The primary goal of multilingual lexicography is translation". Hence, it seems obvious that corpora consisting of previous translations are valuable resources for creating bilingual lexicons. However, the use of parallel translation corpora is limited because of several reasons:

- Size and quality are essential for reliable lexicographic work. Even though parallel corpora are becoming more widely available for more language pairs their contents is still lacking in representativity, completeness and linguistic annotation. Most parallel corpora include documents in a specialized domain with only a limited coverage of the languages involved and are not systematically checked. One of the reasons for the limited availability of parallel corpora is related to copyright issues. High standard translations are usually published and not publicly available. This problem is less severe for monolingual corpora for which a richer variety of data is available. Professional translation, however, is expensive and, therefore, mainly done for commercially interesting purposes.
- Available parallel corpora currently do not include appropriate linguistic annotation. There are automatic annotation efforts producing linguistic features that are essential for efficient lexicographic work. However, these annotations are limited to the accuracy of the tools involved and are usually not checked manually. Hence, reliable annotation can not be expected, often even not with the accuracy reported for the tools applied because they have been trained on other domains and other text types. Furthermore, robust wide-coverage tools are only available for a few languages appropriate.
- Translation data cannot be taken as ground truth for bilingual lexicography. Not only are translation errors and sloppiness problematic but the general representativity of both languages also cannot be guaranteed. Translations "cannot but give a distorted picture of the lan-