Running head: CORPORA FOR COMPUTATIONAL LINGUISTICS

Corpora for computational linguistics

Constantin Orăsan, Le An Ha, Richard Evans, Laura Hasler and Ruslan Mitkov

Research Group in Computational Linguistics

University of Wolverhampton

United Kingdom

{C.Orasan, L.A.Ha, R.J.Evans, L.Hasler, R.Mitkov}@wlv.ac.uk

Abstract

Since the mid 90s corpora has become very important for computational linguistics. This paper offers a survey of how they are currently used in different fields of the discipline, with particular emphasis on anaphora and coreference resolution, automatic summarisation and term extraction. Their influence on other fields is also briefly discussed.

Corpora for computational linguistics

Introduction

Corpora play a major role in the development of new methods in many fields of computational linguistics, as well as the improvement of existing ones. Even though corpora were first employed in computational linguistics in the 1960s (Edmundson, 1969), they started to be used on a large scale only in the 1990s. There are two main reasons for this: first and foremost, advances in computer hardware that came to fruition in the 90s allowed researchers to process large corpora without their needing access to highly specialised computer hardware. As a direct result of these technological advances, the number of available corpora has also increased. The second reason for the increased interest in corpora was motivated by a paradigm shift which marked the field of artificial intelligence in the late 80s and early 90s, when researchers saw the limitations of knowledge-based systems, and started to use approaches such as statistical and machine learning methods which rely on empirical evidence (Russell & Norvig, 1995). As a result, researchers working in computational linguistics began to take an interest in corpus-based research once more.

This paper presents a survey of corpora for computational linguistics, with an emphasis on how corpora are used in anaphora and coreference resolution, automatic summarisation and term extraction. Their influence on other fields in computational linguistics is also briefly mentioned.

Brief introduction to corpora

In this section general concepts about corpora are described in order to facilitate an understanding of the rest of the paper by the reader. However, it should be pointed out

that this section will not try to give a comprehensive overview of corpora in general. Instead, it will mainly focus on those concepts which are relevant to corpora used in computational linguistics research. Comprehensive introductions to corpora and pointers for further readings can be found in (Biber, Conrad, & Reppen, 1998; McEnery, 2003).

McEnery (McEnery, 2003) defines a corpus as "a large body of evidence typically composed of attested language use". In addition it is generally agreed that in order to be useful, a corpus needs to be in machine readable format, so that it is possible to quickly extract relevant information from it. Research which uses corpora can be divided into *corpus-based approaches* where a hypothesis is checked against a corpus, and *corpus-driven approaches* where hypotheses are drawn from a corpus (Biber et al., 1998). In computational linguistics the former is usually associated with the *evaluation* procedure where the success of a method is computed by comparing its output with the corpus annotation, whereas the latter corresponds to the *training* stage in machine learning approaches employed in computational linguistics.

Depending on the type of texts contained in a corpus, it is possible to have *monolingual corpora* where all the texts are in the same language and *multilingual corpora* where the texts are in several languages. Multilingual corpora which contain texts about more or less the same topic are known as *comparable corpora*, whilst multilingual corpora containing texts which are translations of each other are called *parallel corpora*. For applications which deal only with texts in one language, monolingual corpora are usually enough, but there are monolingual applications which can benefit from parallel corpora such as the anaphora resolution method proposed by Mitkov and Barbu (2004). For multilingual applications parallel corpora are usually favoured, but given that quite often such corpora are not available, comparable corpora are used instead.

If a corpus contains only written materials it is called a *written corpus*, if it contains only spoken materials it is referred to as a *spoken corpus*, and if it contains both types of

texts it is known as a *mixed corpus*. In computational linguistics, the type of corpus is chosen depending on the intended application being developed. Spoken and written language have very different properties, so the corpus used has to be chosen accordingly. The size of a corpus is another parameter which is normally considered when talking about corpora. In general large corpora are more useful than smaller ones, but in some cases it is not feasible to build very large corpora (Leech, 1991). However, in computational linguistics very large corpora are preferred because smaller data sets do not usually provide enough examples to enable a method to learn from them (Church & Mercer, 1993).

Corpora can be enhanced with different types of linguistic information by annotating them. Computational linguistics has benefited from both raw and annotated corpora. Raw corpora are usually used to extract patterns and train unsupervised learning methods, whereas annotated corpora are generally employed to train and test supervised machine learning methods and to evaluate the results of many different kinds of methods. Given the large costs associated with the annotation process, in recent years an increasing number of methods which rely on unannotated corpora have been developed. Some of these methods are presented in the next section.

Raw corpora

Raw corpora are collections of texts which were not enhanced with any additional information. As a result, they are investigated using manual or semi-automatic approaches where examples are extracted from the corpus and then analysed. This approach is usually employed in language studies or lexicography, but an increasing number of researchers from computational linguistics have started to use raw corpora in their research. This section will briefly present some of the most important ways in which unannotated corpora can be used.

Banko and Brill (2001) argue that an easier and more successful way to develop applications based on machine learning is not to optimise existing machine learning algorithms, but to train them on larger corpora. They present a method to disambiguate confusion sets such as (*weather, whether*) which instead of trying to develop more advanced disambiguation methods, trains existing unsupervised methods on larger corpora. Evaluation of a system trained on several corpora, each an order of magnitude larger than the previous one, reveals a constant improvement of the results. Even for a 1 billion word corpus, the evaluation does not seem to suggest that the performance of unsupervised methods stops improving in the face of additional training data.

Wiebe and Riloff (2005) present a method which uses unannotated texts to classify sentences in texts as objective and subjective. In order to achieve this, high precision rule-based classifiers were developed and used to automatically annotate an initial training set. An extraction pattern learner was used on this initial training set to extract high confidence extraction patterns associated with objectivity/subjectivity. The extracted patterns were then used to improve the rule-based classifiers. Given the nature of the method, this process can be repeated several times in order to develop a self-training sentence classifier.

Unannotated corpora were also used in word sense disambiguation where they were mainly used to derive word senses (Schütze, 1998). The proposed method relies on clustering the contexts of words together. Another method which requires only raw corpora for word sense disambiguation was proposed by Resnik (1995). In this case the corpus is employed to compute scores for nodes in WordNet (Fellbaum, 1998) which are then used by a disambiguation algorithm. Unannotated corpora have also been successfully exploited to extract collocation patterns (Dagan & Itai, 1990) for anaphora resolution, identify hyponymy relations (Hearst, 1992), to develop part of speech taggers (Brill, 1995), and to develop named entity recognisers (Evans, 2003).

Annotated corpora

Corpus annotation is the process of enriching a corpus with "interpretative, linguistic information" (Leech, 1997). As a result of this process the linguistic material becomes "attached to, linked with, or interspersed with the electronic representation of the language material itself". Annotated corpora is an indispensable resource for most NLP tasks or applications since the data they provide are critical to the development, optimisation and evaluation of new approaches. The result of the annotation process is usually referred to as a *gold standard*, supposedly containing 100% correct information.

The annotation process is *interpretative* because in many cases the decision about how to annotate a text is highly subjective. In order to reduce this subjectivity, a set of annotation guidelines should be developed to provide detailed explanation of, motivation, and justification for the decisions that human annotators should take during the annotation process. For example the guidelines produced for the MUC annotation scheme (Hirschman, 1997) (discussed in the section dedicated to corpora for anaphora and coreference resolution) stipulate which noun phrases should be marked up as coreferential and when. As another example, annotation guidelines used in automatic summarisation indicate how to determine those sentences in a document which contain enough information to be considered important.

An annotation strategy in the form of guidelines outlining what to annotate and when to annotate it, and recommending the best annotation practice, can be very helpful to the annotators, and should improve annotation consistency, replicability, and inter-annotator agreement. Inter-annotator agreement is computed using statistical measures such as the kappa statistic (Carletta et al., 1997), which compares the label assigned by annotators to the same span of text, and also takes into consideration other factors such as the probability that agreement would be achieved by chance. In cases where the kappa statistic cannot be applied, precision and recall can also be computed by

considering one of the annotations as the gold standard and comparing the rest of the annotations with it.

The annotation process requires the markup in text of items that have a special meaning for some purpose, using an *annotation scheme*. This scheme has a special meaning according to the purpose of the annotation and there are many ways in which it can be implemented. In the past, ad-hoc annotation schemes which were suitable for the researchers who implemented NLP systems were used, but in recent years the vast majority of annotation schemes have been based on XML[1] which enables researchers to share and reuse corpora much more easily. Initiatives such as the Corpus Encoding Standard (CES) and Corpus Encoding Standard for XML (XCES) contain guidelines which also ensure the consistency between annotations applied by different researchers.[2]

The annotation can usually be applied using any text editor, but in order to facilitate the process and minimise the number of errors introduced, annotation software is normally used. In general, this software hides information that is irrelevant or troublesome to human annotators in their task and ensures that the resulting annotation is valid. Annotation software can either be developed with a specific goal in mind for example CLinkA can be used only for annotating coreference (Orăsan, 2000), or it can be more general purpose, such as the Alembic Workbench (Day et al., 1998), MMAX (Müller & Strube, 2001) or PALinkA (Orăsan, 2003) tools.

Once a corpus has been annotated it can be used for both training and evaluation purposes. When used for training, methods are developed to learn from the annotation, in contrast to evaluation, where the output of a method used in NLP is compared with the annotation.

Corpora for anaphora and coreference resolution

The process of determining the antecedent of an anaphoric expression is called *anaphora resolution*, whilst *coreference resolution* tries to determine all coreferential chains from a document or set of documents (Mitkov, 2002). Since the early 1990s, research and development in anaphora and coreference resolution has benefited from the availability of corpora, both raw and annotated. The automatic training and evaluation of anaphora resolution algorithms requires that the annotation cover not only single pairs of anaphors and antecedents, but also anaphoric chains, since the resolution of a specific anaphor would be considered successful if any preceding element of the anaphoric chain associated with that anaphor were identified. For coreference resolution, full coreferential chains are also necessary. Unfortunately, anaphorically or coreferentially annotated corpora are not widely available and those that exist are not of a large size.

Annotation schemes

In recent years, a number of corpus annotation schemes for marking up anaphora have come into existence. Notable amongst these are the UCREL anaphora annotation scheme applied to newswire texts (Fligelstone, 1992; Garside & Rayson, 1997) and the SGML-based (MUC) annotation scheme used in the MUC-7 coreference task (Hirschman, 1997). Other well known schemes include those presented in de Rocha (1997) for annotating spoken Portuguese, Botley (1999) for annotating demonstrative pronouns, Bruneseaux and Romary's scheme (Bruneseaux & Romary, 1997), the DRAMA scheme (Passonneau & Litman, 1997), the annotation scheme for marking up definite noun phrases proposed by Poesio and Vieira (1998), and the MATE scheme for annotating coreference in dialogues proposed by Davies, Poesio, Bruneseaux, and Romary (1998).

The UCREL scheme allows the marking of a wide variety of cohesive features as well as the direction of reference (anaphoric or cataphoric), the type of cohesive relationship

involved, the antecedent of an anaphor, and various semantic features of anaphors and antecedents. It also allows for the marking of split antecedents and uncertain cases.

The SGML-based MUC annotation scheme (Hirschman, 1997) has been used in the production of evaluation data for the Message Understanding Conferences (MUCs) and also by a number of researchers to annotate coreferential links (Mitkov et al., 2000; Gaizauskas & Humphreys, 2000). Given an antecedent A and an anaphor B, where both A and B are strings in the text, the basic MUC coreference annotation has the form

```
<COREF ID="100"> A </COREF> ...
<COREF ID="101" TYPE=IDENT REF="100">B</COREF>
```

In the MUC scheme, the attribute ID contains a unique identifier for the annotated string, REF indicates the ID of an entity which is coreferential with the string, TYPE indicates the type of relationship between anaphor and antecedent and IDENT indicates the identity relationship between anaphor and antecedent. The MUC scheme only covers the identity (IDENT) relation for noun phrases and does not include other kinds of coreference relations such as set/subset, part/whole, etc. In addition to these attributes, the annotator can add two more, the first of which is MIN, used in the automatic evaluation of coreference resolution systems. The value of MIN represents the smallest continuous substring of the element that must be identified by a system in order to consider a resolution correct. Secondly, the attribute STATUS can be used and set to the value OPT. This information is used to express the fact that mark-up of some elements in corpora are optional.

van Deemter and Kibble (1999) have criticised the MUC coreference annotation scheme, stating that 'it goes beyond annotation of coreference as it is commonly understood' since it also marks non-referring elements such as quantifying NPs (e.g. every man, most computational linguists) as parts of coreferential chains. van Deemter and

Kibble also express their reservation regarding the marking of indefinite NPs and predicate NPs. However, despite its imperfections, the MUC scheme has the advantage of offering a standard format. Also, although it has been designed to mark only a small subset of anaphoric and coreferential relations, the SGML framework does provide a useful starting point for the standardisation of different anaphoric annotation schemes.

The DRAMA scheme (Passonneau & Litman, 1997)[3] identifies anaphors and antecedents in a text, and marks coreference relationships between them. Although similar to the MUC scheme, the DRAMA scheme classifies and marks different kinds of bridging relationships. DRAMA includes instructions for dealing with some difficult problems in identifying the *markable* entities in dialogues, and treats a wider set of markables than the MUC scheme does.

The scheme proposed by Bruneseaux and Romary (1997) identifies anaphors and antecedents in the text and marks the relationships between them, as is the case with other schemes such as MUC, DRAMA and UCREL. An innovation of this scheme is that it allows references to the visual context to be encoded. The scheme also allows the marking of deixis in the form of pointing and mouse-click gestures.

Poesio and Vieira's (1998) first scheme allows annotators to classify definite noun phrases and mark their textual relationships with other NPs rather than linking referential expressions as in the MUC and DRAMA schemes. As a result, the number of markable entities in this scheme is much more limited. In addition to allowing the classification of definite NPs, the second scheme presented in Poesio and Vieira (1998) allows the marking of the referential link between referring definite NPs and their antecedents.

The MATE scheme for annotating coreference in dialogues (Davies et al., 1998) draws on the MUC coreference scheme, adding mechanisms for marking-up further types of information about anaphoric relations as is done in the UCREL, DRAMA and Bruneseaux and Romary schemes. In particular, this scheme allows for the mark up of

anaphoric constructs typical in Romance languages, such as clitics, and of some typical dialogue phenomena. The scheme also provides for the mark up of ambiguities and misunderstandings in dialogue. The strength of the MATE scheme is that, while based on the widespread MUC scheme and adopting the popular SGML standard, as with the UCREL scheme, it covers a rich variety of anaphoric relations which makes it a promising general-purpose framework.

The XML-based scheme described in Tutin et al. (2000) supports the annotation of a variety of anaphoric relations such as coreference, set membership, substitution, sentential anaphora and indefinite relations which include all cases not covered by the first four types, such as bound anaphora.[4] The annotation scheme encodes the boundaries of each expression, the link between two expressions, and the type of relationship between them. Tutin et al.'s scheme can also encode special cases such as identity-of-sense anaphora, ambiguity, and coordinated (split) antecedents.

Ge's annotation (Ge et al., 1998) covers five kinds of relationships involving pronouns. The author marks pronouns which have explicit nominal antecedents, pronouns with split antecedents, pronouns pointing to an action or event not represented by a single noun phrase as well as two types of pleonastic pronouns: those that are not specific enough and those that appear in cleft constructions.

de Rocha (1997) describes a detailed annotation scheme for marking anaphoric references in a corpus of spoken Portuguese dialogues and extracts from the London-Lund corpus. de Rocha's scheme explores the relationship between anaphora and the topic structure of discourse, by signalling discourse, segment and subsegment topics. In addition to being able to mark features of the discourse structure, this scheme can also mark different aspects of anaphors, such as the type of anaphor (e.g. subject pronoun or full noun phrase), the type of antecedent (implicit, non-surface or explicit, surface antecedent), the topicality status of the antecedent and the type of knowledge required for

the processing of the anaphor (such as syntactic, collocational or discourse knowledge). It allows for anaphora in spoken (and presumably written) texts to be analysed according to a rich variety of inter-related factors, in a way which extends beyond the descriptive analysis of Halliday and Hasan (which is largely implemented in the UCREL annotation scheme outlined above); however, it is very labour-intensive to apply.

Botley's scheme (Botley, 1999) describes the different functions of anaphoric demonstratives in written and spoken texts. Essentially, this scheme classifies demonstrative anaphors according to five distinctive features, each of which can have one of a series of values.

Corpora annotated with anaphoric or coreferential links

One of the few anaphorically annotated resources, the Lancaster Anaphoric Treebank is a 100,000 word sample of the Associated Press (AP) corpus (Leech & Garside, 1991), marked-up with the UCREL anaphora annotation scheme. The MUC coreference task (MUC-6 and MUC-7) gave rise to the production of texts annotated for coreferential links for training and evaluation purposes. The annotated data which complied with the MUC annotation scheme was mostly from the genre of newswire reports on subjects such as corporate buyouts, management takeovers, airline business and plane crashes. Approximately 65,000 words were annotated in total.[5]

A part of the Penn Treebank[6] was annotated to support a statistical pronoun resolution project at Brown University (Ge et al., 1998). The resulting corpus contains 93,931 words and 2,463 pronouns. In addition to providing information on coreference between pronouns and noun phrases, or generally between any two noun phrases, pleonastic pronouns were marked.

A corpus containing around 60,000 words, annotated with a scheme similar to the MUC annotation scheme using the annotation tool CLinkA (Orăsan, 2000), has been

produced at the University of Wolverhampton (Mitkov et al., 2000). The corpus features fully annotated coreferential chains and covers texts from different user manuals (printers, video-recorders etc.). Using a slightly modified version of the guidelines proposed by (Mitkov et al., 2000), (Hasler, Orăsan, & Naumann, 2006) presents a corpus containing over 50,000 words from the domain of terrorism and security news which was annotated with coreference information for NPs and events. A project conducted by members of the University of Stendahl, Grenoble and Xerox Research Centre Europe (Tutin et al., 2000) delivered a 1,000,000 word corpus annotated for anaphoric and cataphoric links. The annotation was limited to (anaphor, closest antecedent) pairs rather than full anaphoric chains. This limitation makes the corpus more suitable for theoretical linguistic research than for evaluation and testing anaphora resolution systems where full anaphoric or coreferential chains are needed. Texts annotated for coreferential links in French are also reported by Popescu-Belis (1998) who marked texts with MUC and Bruneseaux and Romary schemes.

As a consequence of the increasing number of projects in multilingual anaphora resolution, the need for parallel bilingual and multilingual corpora annotated for coreferential or anaphoric links has become obvious. At present these are limited to a small-size English-Romanian corpus developed for testing a bilingual coreference resolution system (Harabagiu & Maiorano, 2000), and a parallel English-French corpus of coreferentially annotated technical manuals at the University of Wolverhampton. The English part of this corpus contains 25,499 words and the French part 28,037 words (Mitkov & Barbu, 2004).

Annotation strategy and inter-annotator agreement

The annotation of anaphoric or coreferential relations is a notoriously difficult, time-consuming and labour-intensive task even when focusing on one single variety of the

phenomenon.[7] Fligelstone (1992) notes that "The nature of the task, with its heavy reliance on interpretation, suggests that it may prove impossible to achieve such a high degree of inter-analyst consistency as with the parsing scheme...". The complexity of the task imposes a restriction that the annotation process should not follow a detailed level of analysis (as in the case of the UCREL and MUC schemes) but focus instead on the identity relation (the MUC scheme). The experience with the MUC annotation scheme shows that even within the narrow domain of NP coreference it is not always easy to decide which NPs should be marked as coreferential. For related discussion on the complexity of anaphora and coreference see (van Deemter & Kibble, 1999).

Mitkov et al. (2000) take the view that in many cases having a more reliable annotation which ensures high agreement between annotators is preferable to a more complicated one covering a wider range of phenomena because the former is less prone to annotation errors.

Corpora for text summarisation

Automatic summarisation is the process which produces summaries from one or more source texts using fully automatic means (Hovy, 2003). When the summary is produced from one document, the field is known as *single document summarisation*, whilst *multi-document summarisation* produces summaries from several documents. A special case of multi-document summarisation is multilingual summarisation, where the input texts are written in different languages. When corpora are built for text summarisation, different issues need to be addressed for each of these types.

The way most summarisation methods use corpora assumes that the corpus is annotated with information which indicates the relevance of sentences to a document or to a specific topic. In this way, a summarisation method can extract a set of features considered important for each sentence, and then use the annotation to determine which

combinations of features indicate sentences to be included in the summary. When a corpus is used to evaluate a summarisation method, the sentences extracted by the method are compared with the ones marked as relevant in the corpus. This approach is very similar to that used in anaphora and coreference resolution.

Corpora for single document summarisation

Building annotated corpora for automatic summarisation has proved to be a daunting task because of the difficulty of defining what exactly an important unit (e.g. sentence, clause or paragraph) is. This mainly stems from the fact that the decision as to whether a sentence is important enough to be marked as such is a highly subjective one, and as a result, the agreement between different annotators is often quite low.

The annotation scheme used to label corpora for automatic summarisation is usually minimal in that it encodes information only about the importance of each sentence. In cases where a sentence is not important, no annotation is attached to the sentence. In some cases, additional information not normally used directly in the summarisation process, but which can have a beneficial impact on the quality of automatic summaries, can also be marked (Hasler, Orăsan, & Mitkov, 2003).

There are several ways to produce corpora for single document summarisation. The best established one is manual annotation, which requires a human annotator to read the whole text and mark important units. Given that manual annotation is difficult and time consuming, automatic annotation methods have been proposed as an alternative. An overview of both manual and automatic annotation methods is presented next.

Manually built corpora for automatic summarisation.

The most common way to build corpora for single document summarisation requires human annotators to read the whole source text and manually mark each sentence for importance according to a set of guidelines. Manually annotated corpora were first used

to train and test summarisation methods by Edmundson (1969). Here, the most important sentences from a heterogenous corpus consisting of 200 documents in the fields of physics, life science, information science and humanities were annotated by human judges. In order to ensure consistency of annotation, the judges were asked to follow a set of guidelines and to select those sentences which indicate *what* is the subject area, *why* is the research necessary, *how* is the problem solved and *which* are the findings of the reported research. These rules broadly correspond to the moves in a scientific paper (Swales, 1990). In addition, the annotators were advised to choose those sentences that minimise redundancy and maximise coherence. Unfortunately, no inter-annotator agreement is reported in the paper.

Hasler, Orăsan, and Mitkov (2003) present an enhanced annotated corpus which differs from the majority of available resources in that it encodes more information. In addition to containing information about the importance of sentences, it indicates parts which can be removed from sentences marked as essential/important. A different label is also provided for those sentences which are not significant enough to be marked as important in their own right, but which have to be considered as they contain information necessary for the understanding of the content of other sentences marked as essential/important. This additional information was included in order to give users an insight into the conciseness and coherence of summaries, respectively. The CAST corpus consists of 163 annotated newswire and scientific texts totalling almost 150,000 words, with a number of texts annotated by two or three annotators. In order to ensure consistency, a detailed set of guidelines was given to the annotators who were asked to identify sentences containing the most important 15% of the text (essential sentences), and then an additional 15% which is the next most important (important sentences).[8] The inter-annotator agreement calculated on this corpus revealed very low values for the kappa statistic, which indicated little agreement among annotators. However, manual

investigation of the selected sentences showed that this low agreement value is caused by the fact that in many cases, the annotators marked different sentences which convey similar information. In light of this, cosine similarity (Salton & McGill, 1983) was used to compute the overlap in the information covered by the annotated sentences. Using only occurrence of words and not senses, the cosine similarity indicated a substantial overlap in the information present.

Marcu (1997) describes an experiment in which 13 independent judges were asked to rate each unit from 5 texts as very important, important and unimportant. In order to facilitate their decision, the texts to be annotated were already split into units. Comparison between the annotations showed that the judges were consistent when they were asked to mark very important and unimportant units, but less consistent in what they considered important. It was possible to apply simple majority voting (i.e. more than 7 judges chose the same category for a unit) in 87% of the cases to decide the importance of a unit. Statistical significance tests showed that the agreement between annotators is significant.

In a similar experiment to determine agreement between annotators, Tsou, Lin, Lai, and Chan (1998) asked 6 groups of evaluators, 3 from North China and 3 from Taipei, to mark with red the most important 10% units in a text, and with blue 15% of the next most important ones, without giving them exact instructions about how to identify a unit. The importance of each unit was computed using a weighted average measure called *perceived importance*. Comparison between propositions' perceived importance showed average overall inter-group consistency between North China and Taiwan, and high intra-group consistency. Therefore, the authors conclude that the background of the annotators plays an important role in selecting the important units.

Given that identification of important sentences is very subjective and difficult, Kupiec, Pederson, and Chen (1995) and Teufel and Moens (1997) took advantage of the

presence of human produced abstracts for texts, and asked annotators to align sentences from the document with sentences from the human produced abstracts. This set of aligned sentences from the document is considered to convey the information from the abstract best, and therefore can be used as a gold standard. Kupiec, Pederson, and Chen (1995) found that 79% of the sentences in the abstracts could be perfectly matched with sentences from the full text, whereas Teufel and Moens (1997) observed that only 31.7% of the sentences from the abstracts have a perfect match. The percent of matching clauses is even lower in the experiment presented by Marcu (1999). Reasons for these very dissimilar results could be the fact that the researchers worked with various types of documents, and also did not use a common definition of a perfect match.

Automatic building of corpora for summarisation.

Despite the fact that it is difficult for humans to align units from a summary with units from the whole document, particularly in cases where the documents are long and contain specialised language, several methods which automatically produce this alignment have been proposed. The underlying idea is that humans often produce a summary through copy-paste of at least parts of sentences from the whole document, so it should be possible to produce this alignment automatically (Jing & McKeown, 1999). The main advantage of such methods is that they can be used to produce large-scale corpora for summarisation with minimum effort.

Marcu (1999) proposes a greedy method which eliminates units from the full document that do not reduce the similarity between a human produced summary and the full document. When it is not possible to shorten the text further on the basis of the similarity measure, the rhetorical structure of the reduced document is used to eliminate more clauses. The method was evaluated on 10 randomly selected articles with an average length of 1,066 words from the Ziff-Davis corpus and it was found to be close to human performance. The method was subsequently used to create a corpus of 6,942 texts with

the important clauses annotated. A drawback of the method proposed by Marcu is the fact that it does not allow any control over the number of sentences which are identified as important. This means that it cannot be used to create corpora that can directly be used to train and evaluate methods which produce summaries of a predefined length. Orăsan (2005) adapted Marcu's method to address this problem, but concluded that for short summaries the method is not suitable. As a result, a genetic algorithm is proposed instead.

Another method which uses (full document, abstract) pairs to align sentences from the summary with sentences from the whole document was proposed by Jing and McKeown (1999). In this case, the abstract is seen as a sequence of words, some of which appear in the document. Therefore the problem of alignment is reformulated as a problem of finding the most likely position of the words from the abstract in the full document using a Hidden Markov Model. The method was also evaluated on the Ziff-Davis corpus and similar results to those reported by Marcu (1999) were obtained.

Corpora for multidocument summarisation

As mentioned above, multi-document summarisation is concerned with creating summaries from more than one source document. This means that in addition to the challenges in single document summarisation, issues such as redundancy, compression ratio, passage selection, the temporal dimension of texts and cross-document coreference need to be tackled (Goldstein, Kantrowitz, Mittal, & Carbonell, 1999).

As a result of these issues, the type of annotation produced in multi-document summarisation differs in several ways from that used in single document summarisation. The first and foremost difference is the fact that it is no longer necessary to annotate sentences which cover the general important information in the documents. Multi-document summarisation is usually employed to produce user-focused summaries, and therefore sentences are annotated with information about how important they are for

a given topic. The most common way to annotate a corpus for multi-document summarisation is to decide on the topics of interest, use a reliable search engine to retrieve documents relevant to the query, and then ask human judges to mark how relevant each sentence from the top retrieved documents is to the selected topic. Radev, Jing, and Budzikowska (2000) used *utility judgement* , which requires assigning a score from 0 to 10 to each sentence as to how relevant that sentence would be to the topic.

In some cases, where corpora are intended only to evaluate multi-document summarisation methods, humans are required to produce "ideal summaries" from clusters of documents relevant to a topic. In these cases, the automatic summaries are evaluated by manually or automatically comparing them with the "ideal summaries".

Given the issues involved in creating corpora for multi-document summarisation, many researchers in the area use readily available corpora which tend to come from well-established, large-scale evaluation conferences such as the Document Understanding Conferences (DUC)[9], the Topic Detection and Tracking (TDT)[10] initiative and the Text REtrieval Conferences (TREC)[11]. Some of these corpora are specifically designed for summarisation evaluation conferences, whilst others are developed for the evaluation of systems in related fields such as topic detection and tracking, or information retrieval. These corpora are useful because they are of a suitable size, and already contain the annotations or gold standard texts needed to exploit them to their full potential.

The SUMMAC corpora.

SUMMAC was the first evaluation conference organised in the field of automatic summarisation (Mani et al., 1998). It was part of the Phase III of the TIPSTER Text Program which finished in 1998. Given that the main purpose of this conference was to explore the evaluation methods available for text summarisation, a corpus was not created specially for the conference. The corpora used was derived from the TREC data and was slightly different from one task to another.

In the *adhoc* task, which concentrated on indicative user-focused summaries, 20 topics, each with 50 documents, were extracted from the TREC data. The annotation available for these texts was the relevance of each document to a query available from the TREC corpus. For the classification task, which required judges to classify document in predefined classes on the basis of their summaries, only 10 topics, each with 100 documents, were selected. The topics of each document were again taken from the TREC corpus. It should be pointed out that SUMMAC evaluated both single and multi-document summaries, and therefore the resources can be used for both types of summarisation.

The DUC Corpora.

The increasing importance of summarisation was acknowledged by the research community through the organisation of the DUC evaluation conferences every year. The purpose of these conferences is to evaluate summarisation systems on different tasks using corpora distributed specifically for this purpose. Some of the tasks in DUC over the years have included automatic summarisation of single documents and of multiple documents on the same topic, creation of a short (100 word) summary by viewpoint and also in response to a question, and creation of a very short (10 word) summary for cross-lingual single documents. In order to evaluate the performance of the participating systems on these tasks, assessors select topics of interest from the datasets and produce summaries according to each task. Hence, the DUC corpora consist not only of collections of documents about the same topic (for multi-document summarisation) but also of human-produced responses to the tasks for evaluation purposes.

As an example, DUC2003 used 30 document clusters of around 10 texts each from each of the TDT (Topic Detection and Tracking) and TREC (Text REtrieval Conference) Novelty track collections. In DUC2004, this collection was used for training the participating systems which then used a corpus of 50 TDT English clusters, 25 TDT

Arabic clusters translated into English by fully automatic Machine Translation systems and 50 TREC English clusters. The sentences in the corpus were not annotated for their importance to a topic, instead summaries of different lengths were produced by humans and used as a gold standard. In addition, topics were attached to each news cluster.

Corpora from the Text Summarization Challenge (TSC).

The developments in the field of automatic summarisation as a result of the above evaluation conferences prompted the research community to organise such conferences for languages other than English. The TSC is a summarisation evaluation programme in Japan initially based on the SUMMAC evaluation. This series of evaluation conferences aims to develop resources for Japanese and to investigate evaluation methods. The participants have to complete various tasks, such as single document summarisation and user-focused summarisation of multiple documents at different compression rates, and their systems are evaluated at sentence-level against a corpus of human gold standard summaries (Okumura, Fukusima, & Nanba, 2003). For the single document tasks in the evaluation, human annotators were required to annotate important sentences in each article at 10%, 30%, and 50% compression rates, and to produce free abstracts at 20% and 40% compression rates. For the multi-document tasks, human judges were asked to produce free summaries from clusters of documents. All three TSC evaluations to date have used articles from the Mainichi Newspaper Database as their text collections, along with other news articles from the Web.

Corpora from the TDT initiative.

One of the most used corpora which was not specifically designed for summarisation evaluation is the TDT corpora. The TDT initiative is part of DARPA's TIDES programme and is designed to evaluate systems which perform the task of topic detection and tracking, i.e. identifying the first mention of a topic in a text collection and then

following its development. This can help to determine which documents are topically related, identify new documents which can be classified as topically related to an existing set, build clusters of texts that discuss the same topic and detect changes between topically cohesive sections. These tasks are clearly related to issues in multi-document summarisation, which is why annotated TDT corpora prove such a popular choice for the development and evaluation of systems that produce summaries of multiple documents.

The TDT corpora consist of newswire, broadcast radio, broadcast TV and website documents from various sources in English, Mandarin and Arabic, although most data are in English. The corpora are split into training, development and testing sets, and are annotated by human judges for topics which fall into several categories (Allan, Carbonell, Doddington, Yamron, & Yang, 1998). The corpora from TDT1 through TDT5 are available from the LDC and are often used as corpora for evaluations in other conferences such as DUC and ACE (Automatic Content Extraction).

CSTBank.

Radev, Otterbacher, and Zhang (2003) developed CSTBank, a corpus annotated for Cross Structure Theory (Radev et al., 2000), which could be useful for multi-document summarisation as it provides a theoretical model for issues that arise when trying to summarise multiple texts. Radev, Jing, and Budzikowska (2000) detail Cross Structure Theory (CST), a theory describing relationships between two or more sentences from different source documents related to the same topic. CST is related to Rhetorical Structure Theory (RST) (Mann & Thompson, 1988) but takes into account the features of multi-document structure and does not have an underlying tree representation or assume writers' intentions. There are 18 domain-independent relations such as identity, equivalence, subsumption, contradiction, overlap, fulfilment and elaboration, between texts spans. Radev, Jing, and Budzikowska (2000) argue that being aware of these relations during multi-document summarisation could help to minimise redundancy or

include contradictions from different sources and therefore improve the quality of the summary. CSTBank contains different clusters of documents arranged in families based on source texts and clustering methods, and is created from a number of text sources, including other existing corpora.

Radev, Otterbacher, and Zhang (2004) describe the annotation process for Phase I of CSTBank, in which they used 8 human judges to manually annotate the first 5 of 6 clusters of related texts using CST relations. Before annotation, the judges attended a training session, and received annotation guidelines which contained 15 practice pairs of sentences in each section for the annotators to complete to ensure sufficient understanding of the task and the relations. However, due to the fact that more than one CST relation can be allocated as the relations are not mutually exclusive, it was difficult to measure the inter-annotator agreement and this had to be based on the existence of relations rather than the relation type.

Multilingual summarisation

A recent trend in the field of automatic summarisation is to develop multilingual automatic summarisation methods. As a result of the fact that the corpus contains documents in several languages, building corpora for multilingual summarisation poses additional challenges.

SUMMBank is an English-Chinese parallel corpus annotated for single- and multi-document summarisation which was developed as part of the Summer 2001 Johns Hopkins Workshop (Radev et al., 2002). The corpus is based on the Hong Kong Newspaper Corpus, a parallel corpus distributed by LDC which contains translations and near translations of English and Chinese news articles, local administration announcements and descriptions of municipal events. SUMMBank consists of a mixture of automatic summaries, human summaries, and documents and summaries relevant to 20

queries. The automatic summaries were produced using four automatic summarisation methods and two baseline methods. These six methods were run at different compression rates to produce over 100 million summaries. For the manual summaries, three human annotators from the LDC used utility judgement to assess the importance of each sentence to a query. Using the scores assigned by humans to the sentences, over 10,000 abstracts and extracts have been produced. Summaries of 5%, 10%, 20%, 30% and 40% compression rates were produced for single documents.

Corpora and terminology extraction

Terminology extraction is an important part of the terminological analysis process required to produce a useful terminology. Terminologies are vital for any activity dealing with specialised languages because they constitute vocabularies for these languages. Terms are "linguistic labels of specialised abstractions, i.e. concepts, ideas, methods, etc." (ISO, 1990; Sager, 1990). The fact that terms are specialised has two implications. Firstly, it means that they are domain dependent (i.e. a term in a field does not have the same meaning in another field, or may not even be a term in another field). Secondly, this also means that the study of terms requires an interdisciplinary approach with linguists and domain specialists working together. Both theoretical and practical views of terms indicate the need to use corpora to perform terminological analysis, or terminology processing. Theoretically, if terms are specialised lexical units, studying terms would require certain knowledge of the specialised field. To acquire this knowledge, a linguist can either decide to study the whole subject, which may prove too time-consuming, or to study only a textual snapshot of the subject, represented in the form of a corpus of texts in the field. The latter option seems more appropriate, as it should take less time, and corpus investigation is more familiar to a linguist than having to study a new subject in its entirety. Practically speaking, corpora are the only place where the terminological

analysis process can take place.

A terminological analysis process can have various outcomes depending on its applications. For example, for the purpose of indexing, a linear or hierarchical list of terms can be produced to serve as entries in the index. For the purpose of producing a specialised monolingual dictionary or a glossary, not only terms, but also a short definition/description of them will be be required. A translation service required to translate technical or specialised documents may need to build bilingual or multilingual terminologies, and the outcome of the process should not only be the list of terms, but also their translations. Although these outcomes appear to be very different, they share the property that the output of the terminology extraction process should contain: 1) a list of terms; 2) a list of certain relations between them; and 3) definitive knowledge about them (term descriptions). In order to achieve these outcomes, corpora have to be used. In this section, we will limit ourselves to terminology extraction, which is a field in computational linguistics that extracts terminology from texts (semi)automatically. Terminology extraction can operate directly on corpora as well as on individual texts. Terminology extraction can provide most of the output required by a terminological analysis process.

In contrast to the fields presented in the previous two sections, the literature on automatic terminology processing shows that unannotated corpora are preferred, mainly because annotated corpora are rarely available or difficult to produce. One of the main reasons for this is the high cost associated with the production of such corpora, due to the fact that it requires domain experts. Furthermore, in contrast to other fields in computational linguistics, annotated corpora can rarely be reused in other domains, or sometimes even in other applications on the same domain. If a corpus is fully terminologically annotated, there is no need to perform terminological analysis. One instance in which a small, annotated corpus can prove useful in term extraction is when it is used to bootstrap a term extraction method that works in the same domain.

The extraction of the list of terms is usually achieved by using statistical measures. The hypothesis behind the use of these measures is that the appearance of terms in corpora should be statistically different from that of other lexical units. The foundation for this difference can be traced back to the nature of terms. For example, because terms are specialised lexical units, one should expect that they appear more frequently in a specialised corpus than in a general corpus. In addition, for terms which are constituted from several words, stronger co-occurrence patterns reflected in higher values for statistical measures such as mutual information are observed. More about terminology extraction approaches using statistical measures can be found in (Smadja, 1993; Daille, 1996).

Statistical measures are not the only way to identify terms; terms often have distinct lexical-syntactic features making them different from other units in texts. By discovering those lexical-syntactic features and using them, it should be possible to separate terms from other lexical units. These lexical-syntactic features can be divided into two groups. The first group uses only word, lemma, and part-of-speech information (for example LEXTER (Bourigault, 1994), JUS (Justeson & Katz, 1996)), whereas the second group uses information provided by a shallow parser to identify terms (Arppre, 1995; Hulth, 2003). The identification of these patterns is achieved using standard corpus linguistic approaches which rely on frequency lists of lexical-semantic patterns and direct corpus investigation.

A list of terms may be sufficient for certain applications, such as indexing, but this is not always the case. In addition to this list, additional information such as a short description of each term or relations between them would be desirable in some cases.

Relations between terms allow us to observe terms not as independent units, but as components of a coherent system. Knowing the relations between the term in question and other terms will facilitate better understanding of the term itself as well as the whole terminology. The extraction of relations among terms from corpora relies on a premise

that those relations are explicitly expressed in the text, and can be extracted using a wide range of NLP techniques, which would vary from simple pattern matching to deep parsing.

The first step in extracting terms' relations requires the identification of the important relations in a domain. Once these relations have been identified, it is possible to perform the extraction of terms' relations from corpora. One of the most common methods employed in this process relies on *knowledge patterns*. A knowledge pattern is "a linguistic pattern, which is repetitive and expresses domain knowledge about the terms" (Meyer, 2001).

Two steps have to be performed in order to extract term relations. In step 1, input from experts or resources which are rich in such patterns (e.g. glossaries[12]) are used to identify important knowledge patterns. Using pattern heuristics, a list of patterns which have the greatest statistical significance are extracted and serve as important knowledge patterns (Riloff & Jones, 1999; Ha, 2004). Using this procedure, patterns such as "X contain Y", "X produce Y" are extracted from a chemistry glossary, and "X cause Y", "X is-symptom-of Y" can be extracted from a cancer related information glossary.

Knowledge patterns can also be extracted from the corpus itself under a bootstrapping setting. Initially, a set of seed terms and knowledge patterns are introduced in the first round. Then in each round, term candidates which appear around knowledge patterns, and patterns appearing in the context of terms extracted in the previous round are added to the pool. By this method, both terms and knowledge patterns can be extracted from a corpus. The process can be controlled using glossaries as a source of "verified" knowledge patterns.

The second step in the relation extraction process uses the knowledge patterns to identify term relations using a corpus. Each time a pattern is found in the corpus, a relation between its arguments, if they are terms, will be established. For example, when the system processes a sentence such as "Pleural mesothelioma commonly causes

breathlessness or difficulty with breathing .", which contains the pattern "cause", a relation between the two terms "pleural mesothelioma" and "breathlessness" will be established. It should be noted that the current state-of-the-art in NLP still does not achieve very accurate identification of term relations, so the outputs from such methods should always be subjected to manual post editing.

Term descriptions can also be extracted from corpora. Terminologists have already exploited concordancers in the task of compiling term descriptions, as they show the contexts in which terms appear. However, this process is tedious and time consuming. Fortunately, NLP techniques can be used to filter out the vast majority of irrelevant information, offering considerable assistance to terminologists in their task. A term description can be considered as a summary of information about the term, and therefore automatic summarisation methods can be utilised for the task.

Parallel or comparable corpora can be used to build bilingual terminologies in two steps: 1. *acqusition steps* where the terms are identified in each language, and 2. *alignment step* where links are established between terms (Gaussier, 1998).

Conclusions

Computational linguistics is one discipline which has benefitted to a large extent from the developments in corpus linguistics. The advent of corpora in computational linguistics marked the move away from processing made up examples to processing real texts, enabling researchers to develop applications which can deal with real texts more successfully. An important influence on this development was progress in computer hardware, which has allowed researchers to build and investigate increasingly larger corpora, therefore improving the quality of the systems developed in computational linguistics.

This paper presented an overview of how corpora can be used in three highly

researched areas in computational linguistics: anaphora and coreference resolution, automatic summarisation and terminology extraction. For anaphora and coreference resolution and automatic summarisation, annotated corpora have proved ideal for the development and evaluation of new methods. Terminology extraction usually employs unannotated corpora due to its domain-specificity and the fact that annotated corpora often cannot be used in other domains, or even for different applications within the same domain.

In the past few years annotated, corpora have proved extremely useful in a wide range of applications, but recently it has been noticed that existing annotated corpora are not large enough to ensure notable advances in some areas of computational linguistics. To this end, researchers have started to focus their attention on very large unannotated corpora (over a billion words). In the not too distant future, as a result of further advances in hardware, software and management and handling of data, major advances are forseeable in computational linguistics as a result of using corpora.

References

Allan, J., Carbonell, J., Doddington, G., Yamron, J., & Yang, Y. (1998). Topic detection and tracking pilot study: Final report. In *Proceedings of the broadcast news transcription and understanding workshop* (pp. 194 – 218).

Arppre, A. (1995). *Term extraction from unrestricted text.* (Tech. Rep.). Lingsoft website, http://www.lingsoft.com.

Banko, M., & Brill, E. (2001, July 9 – 11). Scaling to very very large corpora for natural language disambiguation. In *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics* (pp. 26 – 33). Toulouse, France.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.

Botley, S. (1999). *Corpora and discourse anaphora: using corpus evidence to test theoretical claims*. Unpublished doctoral dissertation, University of Lancaster, Lancaster, UK.

Bourigault, D. (1994). *Lexter, un logiciel d'extraction de terminologie des connaissances a partir de textes*. Unpublished doctoral dissertation, Ecole des Hautes Etudes en Sciences Sociales.

Brill, E. (1995, June). Unsupervised learning of disambiguation rules for part-of-speech tagging. In D. Yarowsky & K. Church (Eds.), *Third Workshop on Very Large Corpora* (pp. 1–13). Massachusetts Institute of Technology, Cambridge, Massachusetts, USA.

Bruneseaux, F., & Romary, L. (1997, June). Codage des références et coréférences dans les dialogues homme-machine. In *Proceedings of Joint International Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing* (pp. 15 – 17). Ontario, Canada.

Carletta, J., Isard, A., Isard, S., Jowtko, J. C., Doherty-Sneddon, G., & Anderson, A. H. (1997). The reliability of a dialogue structure coding scheme. *Computational Linguistics*, *23*(1), 13 – 32.

Church, K. W., & Mercer, R. L. (1993). Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics*, *19*(1), 1-24.

Dagan, I., & Itai, A. (1990). Automatic processing of large corpora for the resolution of anaphora references. In *Proceedings of the 13th International Conference on Computational Linguistics (COLING'90)* (Vol. III, pp. 1 – 3). Helsinki, Finland.

Daille, B. (1996). Study and implementation of combined techniques for automatic extraction of terminology. In J. L. K. Resnik & Philip (Eds.), *The balancing act: Combining symbolic and statistical approaches to language* (p. 49-66). MA MIT Press.

Davies, S., Poesio, M., Bruneseaux, F., & Romary, L. (1998). *Annotating coreference in dialogues: proposal for a scheme for MATE* (Tech. Rep.).

Day, D., Aberdeen, J., Caskey, S., Hirschman, L., Robinson, P., & Vilain, M. (1998). Alembic workbench corpus development tool. In *Proceedings of the First International Conference on Language Resource & Evaluation* (pp. 1021 – 1028).

de Rocha, M. (1997). Supporting anaphor resolution in dialogues with a corpus-based probabilistic model. In *Proceedings of the ACL'97/EACL'97 workshop on Operational factors in practical, robust anaphora resolution* (pp. 54 – 61). Madrid, Spain.

Edmundson, H. P. (1969, April). New methods in automatic extracting. *Journal of the Association for Computing Machinery*, *16*(2), 264 – 285.

Evans, R. (2003, September). A framework for named entity recognition in the open

domain. In *Proceedings of Recent Advances in Natural Language Processing (RANLP-2003)* (pp. 137 – 144). Borovetz, Bulgaria.

Fellbaum, C. (Ed.). (1998). *Wordnet: An eletronic lexical database.* The MIT Press.

Fligelstone, S. (1992). Developing a scheme for annotating text to show anaphoric relations. In G. Leitner (Ed.), *New Directions in English Language Corpora: Methodology, Results, Software Developments* (pp. 152 – 170). Mouton de Gruyter.

Gaizauskas, R., & Humphreys, K. (2000). Quantitative evaluation of coreference algorithms in an information extraction system. In S. Botley & A. M. McEnery (Eds.), *Corpus-based and computational approaches to discourse anaphora* (pp. 145 – 169). John Benjamins Publishing Company.

Garside, R., & Rayson, P. (1997). Higher-level annotation tools. In R. Garside, G. Leech, & A. McEnery (Eds.), *Corpus annotation: Linguistic information from computer text corpora* (pp. 179 – 193). Addison Wesley Longman.

Gaussier, F. (1998). Flow network models for word alignment and terminology extraction from bilingual corpora. In *Proceedings of the 17th international conference on Computational linguistics* (pp. 444 – 450). Montreal, Quebec, Canada.

Ge, N., Hale, J., & Charniak, E. (1998). A statistical approach to anaphora resolution. In *Proceedings of the Sixth Workshop on Very Large Corpora, COLING-ACL '98* (pp. 161 – 170). Montreal, Canada.

Goldstein, J., Kantrowitz, M., Mittal, V., & Carbonell, J. (1999). Summarizing text documents: Sentence selection and evaluation metrics. In *Proceedings of SIGIR '99* (pp. 121 – 128). Berkeley, California.

Ha, L. A. (2004). Co-training applied in automatic term extraction: an experiment. In *Proceedings of the 7th CLUK Colloquium.* Birmingham, UK.

Harabagiu, S., & Maiorano, S. (2000). Multilingual coreference resolution. In *Proceedings of ANLP-NAACL2000* (pp. 142 – 149). Seattle, Washington, US.

Hasler, L., Orăsan, C., & Mitkov, R. (2003, March). Building better corpora for summarisation. In *Proceedings of Corpus Linguistics 2003* (pp. 309 – 319). Lancaster, UK.

Hasler, L., Orăsan, C., & Naumann, K. (2006, 24 - 26 May). NPs for Events: Experiments in Coreference Annotation. In *Proceedings of the 5th edition of the International Conference on Language Resources and Evaluation (LREC2006)* (p. 1167 - 1172). Genoa, Italy.

Hearst, M. (1992, July). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING1992).* Nantes, France.

Hirschman, L. (1997). *MUC-7 coreference task definition.* http://www.muc.saic.com/proceedings/co_ task.pdf.

Hovy, E. (2003). Text summarisation. In R. Mitkov (Ed.), *The Oxford Handbook of computational linguistics* (pp. 583 – 598). Oxford University Press.

Hulth, A. (2003). Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of RANLP 2003.*

ISO. (1990). *ISO 1087 Vocabulary of terminology* (Tech. Rep.). International Organization for Standardization.

Jing, H., & McKeown, K. R. (1999, August). The decomposition of human-written summary sentences. In *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval (SIGIR'99)* (pp. 129 – 136). University of Berkeley, CA.

Justeson, J. S., & Katz, S. L. (1996). Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, *3*(2), 259-289.

Kupiec, J., Pederson, J., & Chen, F. (1995). A trainable document summarizer. In *Proceedings of the 18th ACM/SIGIR Annual Conference on Research and Development in Information Retrieval* (pp. 68 – 73). Seattle.

Leech, G. (1991). The state of the art in corpus linguistics. In K. Aijmer & B. Altenberg (Eds.), *English corpus linguistics* (pp. 8 – 29). Longman.

Leech, G. (1997). Introducing corpus annotation. In R. Garside, G. Leech, & A. McEnery (Eds.), *Corpus annotation: Linguistic information from computer text corpora* (pp. 1 –19). Addison Wesley Longman.

Leech, G., & Garside, R. (1991). Running a grammar factory: the production of syntactically analysed corpora and treebanks. In S. Johannsson & A. Stenstrom (Eds.), *English computer corpora: selected papers and research guide* (pp. 15 – 32). Berlin: Mouton De Gruyter.

Mani, I., Firmin, T., House, D., Chrzanowski, M., Klein, G., Hirshman, L., et al. (1998). *The TIPSTER SUMMAC text summarisation evaluation: Final report* (Tech. Rep. No. MTR 98W0000138). The MITRE Corporation.

Mann, W. C., & Thompson, S. A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organisation. *Text*, *3*(8), 234–281.

Marcu, D. (1997). From discourse structures to text summaries. In I. Mani & M. Maybury (Eds.), *Proceedings of the ACL/EACL '97 Workshop on Intelligent Scalable Text Summarization* (pp. 82 – 88). Madrid, Spain.

Marcu, D. (1999, August). The automatic construction of large-scale corpora for summarization research. In *The 22nd International ACM SIGIR Conference on*

Research and Development in Information Retrieval (SIGIR'99) (p. 137-144). Berkeley, CA.

McEnery, T. (2003). Corpus linguistics. In R. Mitkov (Ed.), *The Oxford Handbook of Computational Linguistics* (pp. 448 – 463). Oxford University Press.

Meyer, I. (2001). Extracting knowledge-rich contexts for terminography. In C. J. D. Bourigault & M. C. L'Homme (Eds.), *Recent advances in computational terminology.* Amsterdam: John Benjamins.

Mitkov, R. (2002). *Anaphora resolution.* Longman.

Mitkov, R., & Barbu, C. (2004). Using corpora to improve pronoun resolution. *Languages in context*, *2*(4), 201 – 211.

Mitkov, R., Evans, R., Orăsan, C., Barbu, C., Jones, L., & Sotirova, V. (2000). Coreference and anaphora: developing annotating tools, annotated resources and annotation strategies. In *Proceedings of the Discourse, Anaphora and Reference Resolution Conference (DAARC2000)* (pp. 49–58). Lancaster, UK.

Müller, C., & Strube, M. (2001, 5th August). MMAX: A tool for the annotation of multi-modal corpora. In *Proceedings of the 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems* (pp. 45 – 50). Seattle, Washington.

Okumura, M., Fukusima, T., & Nanba, H. (2003, May 31 – June 1). Text summarization challenge 2: Text summarization evaluation at NTCIR workshop 3. In *Proceeding of the HLT-NAACL 2003 Workshop on Text summarization* (pp. 49 – 56). Edmonton, Alberta, Canada.

Orăsan, C. (2000). CLinkA a coreferential links annotator. In *Proceedings of LREC'2000* (pp. 491 – 496). Athens, Greece.

Orăsan, C. (2003, July, 5 - 6). PALinkA: a highly customizable tool for discourse

annotation. In *Proceedings of the 4th sigdial workshop on discourse and dialog* (p. 39 - 43). Sapporo, Japan.

Orăsan, C. (2005, February). Automatic annotation of Corpora for Text Summarisation: A Comparative Study. In *Proceedings of 6th International Conference, CICLing2005* (pp. 670 – 681). Mexico City, Mexico: Springer-Verlag.

Passonneau, R. J., & Litman, D. J. (1997). Discourse segmentation by human and automated means. *Computational Linguistics*, *23*(1), 103 – 139.

Poesio, M., & Vieira, R. (1998). A corpus-based investigation of definite description use. *Computational Linguistics*, *24*(2), 183 – 216.

Popescu-Belis, A. (1998, 28 - 30 May 1998). How corpora with annotated coreference links improve reference resolution. In *Proceedings of the First International Conference on Language Resources and Evaluation* (Vol. 1, pp. 567 – 571).

Radev, D., Jing, H., & Budzikowska, M. (2000, 30 April). Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation and user studies. In *Proceedings of the NAACL/ANLP Workshop on Automatic Summarization* (pp. 21 – 29). Seattle, WA, USA.

Radev, D., Otterbacher, J., & Zhang, Z. (2003). *CSTBank: Cross-document Structure Theory Bank.* http://tangra.si.umich.edu/clair/CSTBank.

Radev, D., Otterbacher, J., & Zhang, Z. (2004, May). CSTBank: A corpus for the study of cross-document structural relationship. In *Proceedings of LREC.* Lisbon, Portugal.

Radev, D., Teufel, S., Saggion, H., Lam, W., Blitzer, J., Çelebi, A., et al. (2002, June). *Evaluation of Text Summarization in a Cross-lingual Information Retrieval Framework* (Tech. Rep.). Baltimore, MD: Center for Language and Speech Processing, Johns Hopkins University.

Resnik, P. (1995). Disambiguating noun groupings with respect to Wordnet senses. In D. Yarovsky & K. Church (Eds.), *Proceedings of the third workshop on very large corpora* (pp. 54–68). Somerset, New Jersey: Association for Computational Linguistics.

Riloff, E., & Jones, R. (1999). Learning dictionaries for information extraction using multi-level bootstrapping. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence.* The AAAI Press/MIT Press.

Russell, S., & Norvig, P. (1995). *Artificial intelligence: a modern approach.* Prentice-Hall.

Sager, J. C. (1990). *A practical course in terminology processing.* Amsterdam: John Benjamins.

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval.* New York: McGraw-Hill.

Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, *24*(1), 97 – 124.

Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational Linguistics*, *19*(1), 143-177.

Swales, J. M. (1990). *Genre analysis: English in academic and research settings.* Cambridge University Press.

Teufel, S., & Moens, M. (1997, July 11). Sentence extraction as a classification task. In *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scallable Text Summarization* (pp. 58 – 59). Madrid, Spain.

Tsou, B. K., Lin, H.-L., Lai, T. B. Y., & Chan, S. W. K. (1998). Human judgment as a basis for evaluation of discourse-connective-based full-text abstraction in chinese. *Computational Linguistics and Chinese Language Processing*, *3*(1), 101 – 116.

Tutin, A., Trouilleux, F., Clouzot, C., Gaussier, E., Zaenen, A., Rayot, S., et al. (2000, 16th – 18th November). Annotating a large corpus with anaphoric links. In *Proceedings of the Discourse Anaphora and Reference Resolution Conference (DAARC2000)* (pp. 28 – 38). Lancaster, UK.

van Deemter, K., & Kibble, R. (1999, June). What is coreference, and what should coreference annotation be? In *Proceedings of the ACL99 Workshop on Coreference and its Applications* (pp. 90 – 96). College Park, Maryland, USA.

Wiebe, J., & Riloff, E. (2005, February). Creating subjective and objective sentence classifiers from unannotated texts. In A. Gelbukh (Ed.), *Proccedings of CICLing* (pp. 486 – 497). Mexico City, Mexico.

Footnotes

[1]XML stands for Extensible Markup Language and comprehensive information about it can be found at http://www.w3.org/XML/

[2]More information about CES and XCES can be found at: http://www.cs.vassar.edu/CES/

[3]DRAMA stands for Discourse Reference Annotation for Multiple Applications scheme (Passonneau & Litman, 1997)

[4]The scheme does not cover lexical noun phrase anaphora.

[5]This figure is based on data/information kindly provided to us by Nancy Chinchor.

[6]The Penn Treebank is a corpus of manually parsed texts from the Wall Street Journal.

[7]For example in the case of demonstrative anaphora it is well known that when the antecedent is a text segment longer than a sentence, it is often difficult to decide exactly which text portion represents the antecedent.

[8]The CAST corpus is available at http://clg.wlv.ac.uk/projects/CAST/

[9]http://duc.nist.gov/

[10]http://www.nist.gov/speech/tests/tdt/

[11]http://trec.nist.gov/

[12]A glossary contains terms and short descriptions of them, thus constituting sources that are rich in knowledge patterns, a fact which facilitates the extraction of patterns from them.