Summarising Scientific Articles — Experiments with Relevance and Rhetorical Status

Simone Teufel*
Cambridge University

Marc Moens*
Rhetorical Systems
and University of Edinburgh

In this paper we propose a strategy for the summarisation of scientific articles which concentrates on the rhetorical status of statements in the article: material for summaries is selected in such a way that summaries can highlight the new contribution of the source paper and situate it with respect to earlier work.

We provide a gold standard for summaries of this kind consisting of a substantial corpus of conference articles in computational linguistics annotated with human judgements of the rhetorical status and relevance of each sentence in the articles. We present several experiments measuring our judges' agreement on these annotations.

We also present an algorithm which, on the basis of the annotated training material, selects content from unseen articles and classifies it into a fixed set of seven rhetorical categories. The output of this extraction and classification system can be viewed as a single-document summary in its own right; alternatively, it provides starting material for the generation of task-oriented and user-tailored summaries designed to give users an overview of a scientific field.

1 Introduction

Summarisation systems are often two-phased, consisting of a content selection step followed by a regeneration step. In the first step, text fragments (sentences or clauses) are assigned a score which reflects how important or contentful they are. The highest-ranking material can then be extracted and displayed verbatim as "extracts" (Luhn, 1958; Edmundson, 1969; Paice, 1990; Kupiec, Pedersen, and Chen, 1995). Extracts are often useful in an information retrieval environment since they give users an idea as to what the source document is about (Tombros, Sanderson, and Gray, 1998; Mani et al., 1999), but they are texts of relatively low quality. Because of this, it is generally accepted that some kind of post-processing should be performed to improve the final result, by shortening, fusing or otherwise revising the material (Grefenstette, 1998; Mani, Gates, and Bloedorn, 1999; Jing and McKeown, 2000; Barzilay et al., 2000; Knight and Marcu, 2000).

However, the extent to which it is possible to do post-processing is limited by the fact that contentful material is extracted without information about the general discourse context in which the material occurred in the source text. For instance, a sentence describing the solution to a scientific problem might give the main contribution of the paper, but it might also refer to a previous approach which the authors criticise. Depending on its rhetorical context, the same sentence should be treated very differently

* The work reported here was carried out while both authors were at the University of Edinburgh, HCRC Language Technology Group. Contact: Simone.Teufel@cl.cam.ac.uk

in a summary. We propose in this paper a method for sentence and content selection from source texts, which adds context in the form of information about the rhetorical role the extracted material plays in the source text. This added contextual information can then be used to make the end product more informative and more valuable than sentence extracts.

Our application domain is the summarisation of scientific articles. Summarisation of such texts requires a different approach from, e.g., the approach used in the summarisation of news articles. For example, Barzilay et al. (1999) introduce the concept of *information fusion* which is based on the identification of recurrent descriptions of the same events in news articles. This approach works well because in the news domain newsworthy events are frequently repeated over a short period of time. However, in scientific writing, similar "events" are rare: the main focus is on new scientific ideas, whose main characteristic is their uniqueness and difference from previous ideas.

Other approaches to the summarisation of news articles make use of the typical journalistic writing style, for example the fact that the most newsworthy information comes first; as a result, the first few sentences of a news article are good candidates for a summary (Brandow, Mitze, and Rau, 1995; Lin and Hovy, 1997). The structure of scientific articles does not reflect relevance this explicitly. Instead, the introduction often starts with general statements about the importance of the topic and its history in the field; the actual contribution of the paper itself is often given much later.

The length of scientific articles presents another problem. Let us assume that our overall summarisation strategy is first to select relevant sentences or concepts, and then to synthesise summaries using this material. For a typical 10–20 sentence newswire story, a compression to 20 or 30% of the source provides a reasonable input set for the second step. The extracted sentences are still thematically connected, and concepts in the sentences are not taken completely out of context. In scientific articles, however, the compression rates have to be much higher—shortening a 20-page journal article to a half-page summary requires a compression to 2.5% of the original. Here, the problematic fact that sentence selection is context insensitive does make a qualitative difference. If only one sentence per two pages is selected, all information about how the extracted sentences and their concepts relate to each other is lost; without additional information, it is difficult to use the selected sentences as input to the second stage.

We present an approach to summarising scientific articles which is based on the idea of restoring the discourse context of extracted material by adding the rhetorical status to each sentence in a document. The innovation of our approach is that it defines principles for content selection specifically for scientific articles, and that it combines sentence extraction with robust discourse analysis. The output of our system is a list of extracted sentences along with their rhetorical status (e.g. sentence **11** describes the scientific goal of the paper, and sentence **9** criticises previous work), as illustrated in figure 1. The example paper we use throughout the article is F. Pereira, N. Tishby, L. Lee's *Distributional Clustering of English Words* (ACL-1993, cmp_lg/9408011); it was chosen as it is the paper most often cited within our collection.

Such lists serve two purposes: in themselves, they already provide a better characterisation of scientific articles than sentence extracts do, and in the longer run, they will serve as better input material for further processing.

An extrinsic evaluation (Teufel, 2001; Teufel, In Preparation) shows that the output of our system is already a useful document surrogate in its own right. But post-processing could turn the rhetorical extracts into something even more valuable: the added rhetorical context allows for the creation of a new kind of summary. Consider for instance the user–oriented and task–tailored summaries shown in figures 2 and 3. Their composition was guided by fixed building plans for different tasks and different user

AIM	**10**	*Our research addresses some of the same questions and uses similar raw data, but we investigate how to factor word association tendencies into associations of words to certain hidden senses classes and associations between the classes themselves.*
	11	*While it may be worthwhile to base such a model on preexisting sense classes (Resnik, 1992), in the work described here we look at how to derive the classes directly from distributional data.*
	162	*We have demonstrated that a general divisive clustering procedure for probability distributions can be used to group words according to their participation in particular grammatical relations with other words.*
BASIS	**19**	*The corpus used in our first experiment was derived from newswire text automatically parsed by Hindle's parser Fidditch (Hindle, 1993).*
	113	*The analogy with statistical mechanics suggests a deterministic annealing procedure for clustering (Rose et al., 1990), in which the number of clusters is determined through a sequence of phase transitions by continuously increasing the parameter* **EQN** *following an annealing schedule.*
CONTRAST	**9**	*His notion of similarity seems to agree with our intuitions in many cases, but it is not clear how it can be used directly to construct word classes and corresponding models of association.*
	14	*Class construction is then combinatorially very demanding and depends on frequency counts for joint events involving particular words, a potentially unreliable source of information as we noted above.*

Figure 1
Extract of system output for example paper

models, whereby the building blocks are defined as sentences of a specific rhetorical status. In our example, most textual material is extracted verbatim (additional material is underlined in figures 2 and 3; the original sentences are given in figure 6). The first example is a short abstract generated for a non-expert user and for general information; its first two sentences give background information about the problem tackled. The second abstract is aimed at an expert; therefore, no background is given, and instead differences of this approach to similar ones are described.

The actual construction of these summaries is a complex process involving tasks such as sentence planning, lexical choice and syntactic realisation — tasks which are outside the scope of this paper. The important point is that it is the knowledge about the rhetorical status of the sentences which enables the tailoring of the summaries according to users' expertise and task. The rhetorical status allows for other kinds of applications too: several articles can be summarised together, contrasts or complementarity between articles can be expressed, and summaries can be displayed together with citation links to help users navigate several related papers.

The rest of this paper is structured as follows: section 2 describes the theoretical and empirical aspects of document structure we model in this work. These aspects include rhetorical status and relatedness:

- *Rhetorical status in terms of problem solving:* What is the goal and contribution of the paper? This type of information is often marked by meta-discourse and by conventional patterns of presentation (cf. section 2.1).
- *Rhetorical status in terms of intellectual attribution:* What information is claimed to be new, and which statements describe other work? This type of information can be recognised by following the "agent structure" of text, i.e., by looking at all grammatical subjects occurring in sequence (cf. section 2.2).
- *Relatedness between articles:* What articles is this work similar to, and in what respect? This type of information can be found by examining fixed indicator phrases like *"in contrast to . . . "*, section headers and citations (cf. section 2.3).

0 *This paper's topic is to automatically classify words according to their contexts of use.* **4** *The problem is that for large enough corpora the number of possible joint events is much larger than the number of event occurrences in the corpus, so many events are seen rarely or never, making their frequency counts unreliable estimates of their probabilities.* **162** *This paper's specific goal is to group words according to their participation in particular grammatical relations with other words,* **22** *more specifically to classify nouns according to their distribution as direct objects of verbs.*

Figure 2
Non-expert summary, general purpose

44 *This paper's goal is to organise a set of linguistic objects such as words according to the contexts in which they occur, for instance grammatical constructions or n-grams.* **22** *More specifically: the goal is to classify nouns according to their distribution as direct objects of verbs.* **5** *Unlike Hindle (1990),* **9** *this approach constructs word classes and corresponding models of association directly.* **14** *In comparison to Brown et al. (1992), the method is combinatorially less demanding and does not depend on frequency counts for joint events involving particular words, a potentially unreliable source of information.*

Figure 3
Expert summary, contrastive links

These aspects of rhetorical status are encoded in an annotation scheme which we present in section 2.4. Annotation of relevance is covered in section 2.5.

In section 3, we report on the construction of a gold standard for rhetorical status and relevance, and on the measurement of agreement amongst human annotators. We then describe the system we built which simulates the human annotation in section 4. Section 5 presents an overview of the intrinsic evaluation we performed, and section 6 closes with a summary of the contribution on this work, limitations and future work.

2 Rhetorical Status, Citations and Relevance

It is important for our task to find the right definition of "rhetorical status" to describe the content in scientific articles. The definition should both capture generalisations about the nature of scientific texts, and also provide the right kind of information to enable the construction of better summaries for a practical application. Another requirement is that the analysis should be applicable to research articles from different presentational traditions and subject matters.

For the development of our scheme, we used the chronologically first 80 articles in our corpus of conference articles in computational linguistics (articles presented at COLING, ANLP and (E)ACL conferences or workshops). Due to the interdisciplinarity of the field, the papers in this collection cover a challenging range of subject matters, such as logic programming, statistical language modelling, theoretical semantics, computational dialectology and computational psycholinguistics. The research methodology and tradition of presentation is very different in these fields (e.g., computer scientists write very different papers than theoretical linguists), and we expect our analysis to be equally applicable in a wider range of disciplines and sub-disciplines other than those named.

2.1 Rhetorical Status

Our model relies on the following dimensions of document structure in scientific articles:

Problem structure. Research is often described as a problem solving activity (Jordan, 1984; Trawinski, 1989; Zappen, 1983). Three information types can be expected to occur in any research article: problems (research goals), solutions (methods) and results. In many disciplines, particularly the experimental sciences, this problem-solution structure has been crystallised in a fixed presentation of the scientific material, as Introduction, Method, Result and Discussion (van Dijk, 1980). But many texts in computational linguistics do not adhere to this presentation, and our analysis therefore has to be based on the underlying logical (rhetorical) organisation, using textual representation only as an indication.

Intellectual attribution. Scientific texts should make clear what the new contribution is, as opposed to previous work (specific other researchers' approaches) and background material (generally accepted statements). We noticed that intellectual attribution has a segmental character. Statements in a segment *without* any explicit attribution are often interpreted as belonging to the most recent explicit attribution statement (e.g. *"Other researchers claim that"*). Our rhetorical scheme assumes that readers have no difficulty in understanding intellectual attribution, an assumption which we verified experimentally.

Scientific argumentation. In contrast to the view of science as a disinterested "fact factory", researchers like Swales (1990) have long claimed that there is a strong social aspect to science, because the success of a researcher is correlated with her ability to convince the field of the quality of her work and the validity of her arguments. Authors construct an argument which Myers (1992) calls the "rhetorical act of the paper": the statement that their work is a valid contribution to science. Swales breaks down this "rhetorical act" into single, non-hierarchical argumentative moves (i.e., rhetorically coherent pieces of text, which perform the same communicative function). His CARS model ("Constructing A Research Space") shows how patterns of these moves can be used to describe the rhetorical structure of introduction sections of physics articles. Importantly, Swales' moves describe the rhetorical status of a text segment with respect to the overall message of the document, and not with respect to adjacent text segments.

Attitude towards other people's work. We are interested in how authors include reference to other work into their argument. In the flow of the argument, each piece of other work was mentioned for a specific reason: it is portrayed as a rival approach, as a prior approach with a fault, or as an approach contributing parts of the own solution. In well-written papers, this relation is often expressed in an explicit way. The next section looks at the stylistic means available to the author to express the connection between previous approaches and their own work.

2.2 Meta-Discourse and Agentivity

Explicit meta-discourse is an integral aspect of scientific argumentation and a way of expressing attitude towards previous work. Examples for meta-discourse are phrases like *"we argue that"* and *"in contrast to common belief, we"*. Meta-discourse is ubiquitous in scientific writing: Hyland (1998) found a meta-discourse phrase on average after every 15 words in running text.

A large proportion of scientific meta-discourse is conventionalised, particularly in

• *We employ Suzuki's algorithm to learn case frame patterns as dendroid distributions.* (9605013)

• *Our method combines similarity-based estimates with Katz's back-off scheme, which is widely used for language modeling in speech recognition.* (9405001)

• *Thus, we base our model on the work of Clark and Wilkes-Gibbs (1986), and Heeman and Hirst (1992) . . .* (9405013)

• *The starting point for this work was Scha and Polanyi's discourse grammar (Scha and Polanyi, 1988; Pruest et al., 1994).* (9502018)

• *We use the framework for the allocation and transfer of control of Whittaker and Stenton (1988).* (9504007)

• *Following Laur (1993), we consider simple prepositions (like "in") as well as prepositional phrases (like "in front of").* (9503007)

• *Our lexicon is based on a finite-state transducer lexicon (Karttunen et al., 1992).* (9503004)

• *Instead of . . . we will adopt a simpler, monostratal representation that is more closely related to those found in dependency grammars (e.g., Hudson (1984)).* (9408014)

Figure 4
Statements expressing research continuation, with source articles

the experimental sciences, and particularly in the methodology or result section (e.g., *"we present original work . . . "*, or *"An ANOVA analysis revealed a marginal interaction/a main effect of . . . "*). Swales (1990) lists many such fixed phrases as co-occurring with the moves of his CARS model (p.144;pp.154–158;pp.160–161). They are useful indicators of overall importance (Pollock and Zamora, 1975); also, they can be relatively easily recognised with information extraction techniques, e.g., regular expressions. Paice (1990) introduces grammars for pattern matching of indicator phrases, e.g., *"the aim/purpose of this paper/article/study"* and *"we conclude/propose"*.

Apart from this conventionalised meta-discourse, we noticed that our corpus contains a large number of meta-discourse statements which are less formalised: statements about aspects of the problem-solving process or the relation to other work. Figure 4, for instance, shows that there are many ways to say that one's research is based on somebody else's ("research continuation"). The sentences do not look similar on the surface: the syntactic subject can be the authors, the originators of the method or even the method itself. Also, the verbs are very different (*"base, be related, use, follow"*). Some sentences use metaphors of change and creation. The wide range of linguistic expression we observed presents a challenge for recognition and correct classification by standard information extraction patterns.

With respect to agents occurring in scientific meta-discourse, we make two suggestions: a) that scientific argumentation follows *prototypical* patterns and employs recurrent types of agents and actions, and b) that it is possible to recognise many of them automatically. Agents play fixed roles in the argumentation, and there are so few of these roles that they can be enumerated: as rivals, as contributors of part of the solution (*"they"*), as the entire research community in the field, or as the authors of the paper themselves (*"we"*). Note the similarity of agent roles to the three kinds of intellectual attribution mentioned above. We also propose prototypical actions frequently occurring in scientific discourse: the field might *"agree"*, a particular researcher can *"suggest"* something, and a certain solution could either *"fail"* or *"be successful"*. In section 4 we will describe the three features used in our implementation which recognize meta-discourse.

Another important construct which expresses relations to other researchers' work are formal citations, which we will turn to now.

2.3 Citations and Relatedness

Citation indexes are constructs which contain pointers between *cited* texts and *citing* texts (Garfield, 1979), traditionally in printed form. When done on-line (as in *CiteSeer*, cf. Lawrence et al. (1999) or in Nanba and Okumura's (1999) work), citations are presented in context for users to browse. Browsing each citation is time-consuming, but useful: just knowing *that* an article cites another is often not enough. One needs to read the context of the citation to understand the relation between the articles. Citations may vary in many dimensions; e.g., they can be central or perfunctory, positive or negative (i.e., critical); apart from scientific reasons, there is also a host of social reasons for citing ("Politeness, tradition, piety"; (Ziman, 1969).)

We concentrate on two citation contexts which are particularly important for the information needs of researchers:

- Contexts where an article is cited negatively or contrastively; and
- Contexts where an article is cited positively, or where the authors state that their own work originates from the cited work.

A distinction of these contexts would enable us to build more informative citation indexes. We suggest that this rhetorical distinction can be made manually and automatically for each citation; we use a large corpus of scientific papers with humans judgements of this distinction to train a system to do the same.

2.4 The Rhetorical Annotation Scheme

Our Rhetorical Annotation Scheme (cf. figure 5) encodes the aspects of scientific argumentation, meta-discourse and relatedness to other work described before. The categories are assigned to full sentences, but a similar scheme could be developed for clauses or phrases.

The annotation scheme is non-overlapping and non-hierarchical, and each sentence must be assigned to exactly one category. As adjacent sentences of the same status can be considered to form zones of the same rhetorical status, we call the units *Rhetorical Zones*. The shortest zones are one sentence long.

AIM	Specific research goal of the current paper
TEXTUAL	Statements about section structure
OWN	(Neutral) description of own work presented in current paper: Methodology, results, discussion
BACKGROUND	Generally accepted scientific background
CONTRAST	Statements of comparison with or contrast to other work; weaknesses of other work
BASIS	Statements of agreement with other work or continuation of other work
OTHER	(Neutral) description of other researchers' work

Figure 5
Annotation scheme for rhetorical status

The rhetorical status of a sentence is determined on the basis of the global context of the paper. For instance, while the OTHER category describes all neutral descriptions of other researchers' work, the categories BASIS and CONTRAST are applicable to sentences expressing a research continuation relationship or a contrast to other work. Generally accepted knowledge is classified as BACKGROUND, whereas own work is separated into the specific research goal (AIM), and all other statements about the own work (OWN).

The annotation scheme expresses important discourse and argumentation aspects of scientific articles, but with its 7 categories it is not designed to model the full complexity of scientific texts. The category OWN, for instance, could be further subdivided into method (solution), results, and further work, which is not done in the work reported here. There is a conflict between explanatory power and the simplicity necessary for reliable human and automatic classification, and we decided to restrict ourselves to the rhetorical distinctions which are most salient and potentially most useful for several information access applications. The user–tailored summaries and more informative citation indexes we mentioned before are just two such applications; another one is be the indexing and previewing of the internal structure of the article. To enable this, our scheme contains the additional category TEXTUAL, which captures previews of section structure (*"section 2 describes our data ... "*). This information would make it possible to label sections with the author's indication of their contents.

Our rhetorical analysis is non-hieararchical, in contrast to Rhetorical Structure Theory (Mann and Thompson, 1987; Marcu, 1999), and it concerns text pieces at a lower level of granularity. While we do agree with RST that the structure of text is hierarchical in many cases, it is our belief that the relevance and function of certain text pieces can be determined without analyzing the full hierarchical structure of the text. Another difference to RST is the fact that our analysis aims at capturing the rhetorical status of a piece of text in respect to the overall message, and not in relation to adjacent pieces of text.

2.5 Relevance

As our immediate goal is to select important content from a text, we also need a second set of gold standards, which are defined by relevance (as opposed to rhetorical status). Relevance is a difficult issue because it is *situational* to a unique occasion (Saracevic, 1975; Spärck Jones, 1990; Mizzaro, 1997): humans perceive relevance differently from each other, and differently in different situations. Paice and Jones (1993) report that they abandonned an informal sentence selection experiment in which they used agriculture articles and experts in the field as subjects, as the subjects were too strongly influenced by their personal research interest.

As a result of subjectivity, a number of human sentence extraction experiments over the years have resulted in low agreement figures. Rath, Resnick and Savage (1961) report that six subjects agreed only on 8% of 20 sentences they were asked to select out of short *Scientific American* texts, and that five agreed on 32% of the sentences. They found that after six weeks, subjects selected on average only 55% of the sentences they themselves selected previously. Edmundson et al. (1961) find similarly low human agreement for research articles. More recent experiments reporting more positive results all used news text (Jing et al., 1998; Zechner, 1995). As discussed above, the compression rates on these texts are far lower: there are fewer sentences to choose from, making it easier to agree which ones to select. Scientific texts also require more background knowledge, thus importing an even higher high level of subjectivity into sentence selection experiments.

Recently, researchers have been looking for more objective definitions of relevance.

Kupiec, Pedersen and Chen (1995) define relevance by abstract-similarity: a sentence in the document is considered relevant if it shows a high level of similarity to a sentence in the abstract. This definition of relevance has the advantage that it is fixed, i.e., the researchers have no influence over it. However, it relies on two assumptions: that the writing style is such that there is a high degree of overlap between sentences in the abstract and in the main text, and that the abstract is indeed the target output which is most adequate for the final task.

In our case, neither assumption holds. First, the experiments in Teufel and Moens (1997) showed that in our corpus only 45% of the abstract sentences appear elsewhere in the body of the document – either as a close variant or in identical form – whereas Kupiec et al. report a figure of 79%. We believe that the reason for the difference is that in our case the abstracts were produced by the document authors, and by professional abstractors in Kupiec et al.'s case. Author summaries tend to be less systematic (Rowley, 1982) and more "deep generated", while summaries by professional abstractors follow an internalised building plan (Liddy, 1991) and are often created by sentence extraction (Lancaster, 1998).

Second, and more importantly, the abstracts and improved citation indexes we intend to generate are not modelled on traditional summaries — traditional summaries do not provide the type of information needed for the applications we have in mind. Information about related work plays an important role in our strategy for summarisation and citation indexing, but such information is rarely found in abstracts. We empirically found that the rhetorical status of information occurring in the author abstracts is very limited, and mostly consists of information about the goal of the paper and specifics of the solution. Details of this analysis are given in section 3.2.2.

We thus decided to augment our corpus with an independent set of human judgements of relevance. We wanted to replace the vague definition of relevance often used in sentence extraction experiments with a more operational definition based on rhetorical status. For instance, a sentence is only considered relevant if it describes the research goal or states a difference with a rival approach. More details of the instructions we used are given in section 3.

Thus, we have two parallel human annotations in our corpus: Rhetorical Annotation and Relevance Selection. In both tasks, *each* sentence in the articles is classified: each sentence receives one rhetorical category and also the label "irrelevant" or "relevant". This strategy can create redundant material, e.g., when the same fact is expressed in a sentence in the introduction, a sentence in the conclusions and one in the middle of the document. But this redundancy also helps mitigate one of the main problems with sentence-based gold standards, namely the fact that there is no one single best extract for a document. In our annotation, *all* qualifying sentences in the document are identified and classified into the same group, which makes later comparisons with system performance fairer. Also, later steps can not only find redundancy in the intermediate result and remove it, but also use the redundancy as an indication of importance.

Figure 6 gives an example of the manual annotation. Relevant sentences of all rhetorical categories are shown. Our system creates a list like the one in figure 6 automatically; figure 23 shows the actual output of the system when run on the example paper. In the next section, we turn to the manual annotation step and the development of the gold standard used during system training and system evaluation.

Aim:

10	*Our research addresses some of the same questions and uses similar raw data, but we investigate how to factor word association tendencies into associations of words to certain hidden senses classes and associations between the classes themselves.*
22	*We will consider here only the problem of classifying nouns according to their distribution as direct objects of verbs; the converse problem is formally similar.*
25	*The problem we study is how to use the* **EQN** *to classify the* **EQN**.
44	*In general, we are interested on how to organise a set of linguistic objects such as words according to the contexts in which they occur, for instance grammatical constructions or n-grams.*
46	*Our problem can be seen as that of learning a joint distribution of pairs from a large sample of pairs.*
162	*We have demonstrated that a general divisive clustering procedure for probability distributions can be used to group words according to their participation in particular grammatical relations with other words.*

Background:

0	*Methods for automatically classifying words according to their contexts of use have both scientific and practical interest.*
4	*The problem is that for large enough corpora the number of possible joint events is much larger than the number of event occurrences in the corpus, so many events are seen rarely or never, making their frequency counts unreliable estimates of their probabilities.*

Own (Details of Solution):

66	*The first stage of an iteration is a maximum likelihood, or minimum distortion, estimation of the cluster centroids given fixed membership probabilities.*
140	*The evaluation described below was performed on the largest data set we have worked with so far, extracted from 44 million words of 1988 Associated Press newswire with the pattern matching techniques mentioned earlier.*
163	*The resulting clusters are intuitively informative, and can be used to construct class-based word coocurrence* [sic] *models with substantial predictive power.*

Contrast with Other Approaches/Weaknesses of Other Approaches:

9	*His notion of similarity seems to agree with our intuitions in many cases, but it is not clear how it can be used directly to construct word classes and corresponding models of association.*
14	*Class construction is then combinatorially very demanding and depends on frequency counts for joint events involving particular words, a potentially unreliable source of information as we noted above.*
41	*However, this is not very satisfactory because one of the goals of our work is precisely to avoid the problems of data sparseness by grouping words into classes.*

Basis (Imported Solutions):

65	*The combined entropy maximization entropy* [sic] *and distortion minimization is carried out by a two-stage iterative process similar to the EM method (Dempster et al., 1977).*
113	*The analogy with statistical mechanics suggests a deterministic annealing procedure for clustering (Rose et al., 1990), in which the number of clusters is determined through a sequence of phase transitions by continuously increasing the parameter* **EQN** *following an annealing schedule.*
153	*The data for this test was built from the training data for the previous one in the following way, based on a suggestion by Dagan et al. (1993).*

Figure 6
Example of manual annotation: relevant sentences with rhetorical status

3 Human Judgements: the Gold Standard

For any linguistic analysis which requires subjective interpretation and which is therefore not objectively true or false, it is important to show that humans share some intuitions about the analysis. This is typically done by showing that they can apply it independently of each other and that the variation they display is bounded, i.e. not arbitrarily high. The argument is strengthened if the judges are people other than the developers of the analysis, preferably "naïve" subjects, i.e. not computational linguists. Apart from the cognitive validation of our analysis, high agreement is essential if the annotated corpus is to be used as training material for a machine learning process, like the one we describe in section 4. Noisy and unreliably annotated training material will very likely deteriorate the classification performance.

In inherently subjective tasks, it is also common practice to consider human performance as an upper bound. The theoretically best performance of a system is reached if agreement amongst a pool of human annotators does not decrease when the system is added to the pool. This is so because an automatic process cannot do any better in this situation than to be indistinguishable from human performance.

3.1 Corpus

The annotated development corpus consists of 80 conference articles in computational linguistics (12,188 sentences; 285,934 words). It is part of a larger corpus of 260 articles (1.1 million words), which we collected from the CMP_LG archive (CMP_LG, 1994). The appendix lists the 80 articles (archive numbers, titles and authors) of our development corpus; it consists of the 80 chronologically oldest articles in the larger corpus, containing articles deposited between May 1994 and May 1996, whereas the entire corpus stretches until 2001.

Papers were included if they were presented at one of the following conferences (or associated workshops): *The Annual Meeting of the Association for Computational Linguistics* (ACL), *The Meeting of the European Chapter of the Association for Computational Linguistics* (EACL), the *Conference on Applied Natural Language Processing* (ANLP), the *International Joint Conference on Artificial Intelligence* (IJCAI) and the *International Conference on Computational Linguistics* (COLING). As mentioned above, a wide range of different subdomains of the field of computational linguistics are covered.

We added XML markup to the corpus: titles, authors, conference, date, abstract, sections, headlines, paragraphs and sentences were marked up. Equations, tables, images were removed and replaced by place holders. Bibliography lists were marked up and parsed. Citations and occurrences of author names in running text were recognised, and self citations were recognised and specifically marked up. (Linguistic) example sentences and example pseudo code were manually marked up, such that clean textual material (i.e., the running text of the article without interruptions) was isolated for automatic processing. The implementation uses the TTT software (Grover, Mikheev, and Matheson, 1999).

3.2 Annotation of Rhetorical Status

The annotation experiment described here (and in Teufel et al. (1999) in more detail) tests the rhetorical annotation scheme presented in section 2.3.

3.2.1 Rationale and Experimental Design

Annotators. Three task-trained annotators were used: Annotator A and B have degrees in Cognitive Science and Speech Therapy. They were paid for the experiment. Both are well-used to reading scientific articles for their studies, and roughly understand the contents of the articles they annotated because of the closeness of their fields to computational linguistics. Annotator C is the first author. We did not want to declare Annotator C the expert annotator; we believe that in subjective tasks like the one described here, there are no real experts.

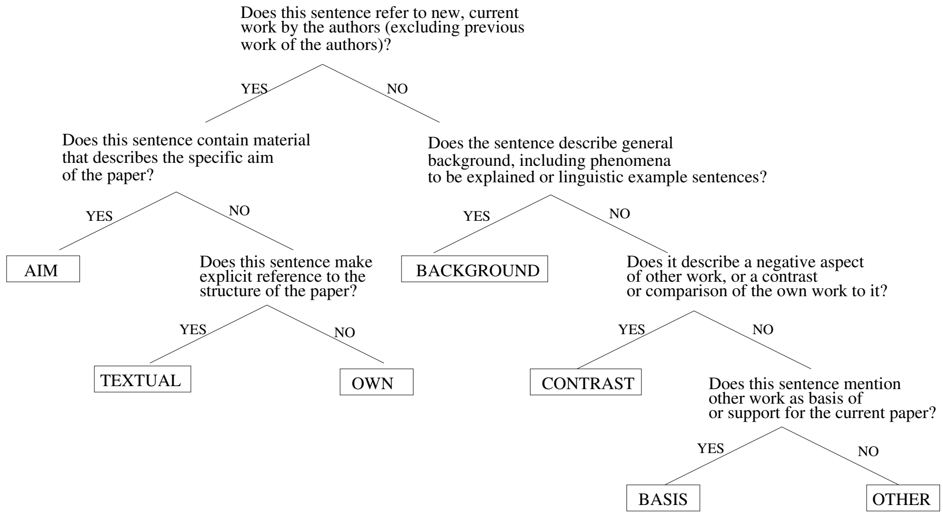

Figure 7
Decision tree for rhetorical annotation

Guidelines. Written guidelines (17 pages) describe the semantics of the categories, ambiguous cases and decision strategies. The guidelines also include the decision tree reproduced in figure 7.

Training. Annotators received a total of 20 hours of training. Training consisted of the presentation of annotation of six example papers and the annotation of 8 training articles under real conditions, i.e. independently. In subsequent training sessions, decision criteria for difficult cases encountered in the training articles were discussed. Obviously, the training articles were excluded from measurements of human agreement.

Materials and Procedure. 25 articles were used for annotation. As no annotation tool was available at the time, annotation took place on paper; the categories were later transferred into the electronic versions of the articles by hand. Skim-reading and annotation typically took between 20–30 minutes per article, but there were no time restrictions. No communication between the annotators was allowed during annotation. Six weeks after the initial annotation, annotators were asked to re-annotate 6 random articles out of the 25.

Evaluation measures. We measured two formal properties of the annotation: stability and reproducibility (Krippendorff, 1980). Stability, the extent to which one annotator will produce the same classifications at different times, is important because an instable annotation scheme can never be reproducible. Reproducibility, the extent to which different annotators will produce the same classifications, is important because it measures the consistency of shared understandings (or meaning) held between annotators.

We use the Kappa coefficient K (Siegel and Castellan, 1988) to measure stability and reproducibility, following Carletta (1996). The Kappa coefficient is defined as follows:

$$K = \frac{P(A)-P(E)}{1-P(E)}$$

where P(A) is pairwise agreement, and P(E) random agreement. K varies between 1 when agreement is perfect, and -1 when there is a perfect negative correlation. K=0 is defined as the level of agreement which would be reached by random annotation using the same distribution of categories as the real annotators did.

The main advantage of Kappa as an annotation measure is that it factors out random agreement by numbers of categories and by their distribution. As Kappa also abstracts over the number of annotators considered, it allows us to numerically compare agreement between a group of human annotators with the agreement between the system and one or more annotators (section 5), which we use as one of the performance measures of the system.

3.2.2 Results

The annotation experiments show that humans distinguish the seven rhetorical categories with a stability of K=.82, .81, .76 (N=1220; k=2, where K stands for the Kappa coefficient, N for the number of items (sentences) annotated and k for the number of annotators). This is equivalent to 93%, 92% and 90% agreement. Reproducibility was measured at K=.71 (N=4261, k=3), which is equivalent to 87% agreement. On Krippendorff's (1980) scale, agreement of K=.8 or above is considered as reliable, agreement of .67–.8 as marginally reliable and agreement of K<.67 as unreliable. On Landis and Koch's (1977) more forgiving scale, agreement of .0–.2 is considered as showing "slight" correlation, .21–.4 as "fair", .41–.6 as "moderate", .61–0.8 as "substantial", and .81 –1.0 as "almost perfect". According to these guidelines, our results can be considered reliable, substantial annotation.

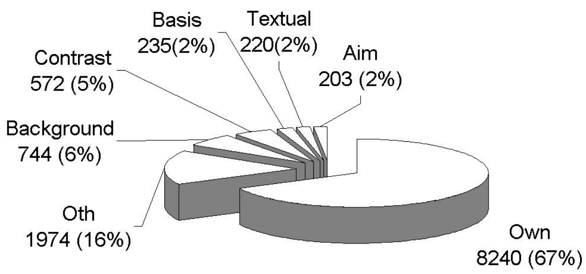

Figure 8
Distribution of rhetorical categories (entire document)

Figure 8 shows that the distribution of the seven categories is very skewed, with 67% of all sentences being classified as OWN. (The distribution was calculated using all three judgements per sentence, cf. the calculation of Kappa. The total number of items is then k·N, i.e. 12783 in this case.)

Figure 9 shows a confusion matrix between two annotators. The numbers represent absolute sentence numbers, and the diagonal (boldfaced numbers) are the counts of sentences that were identically classified by both annotators. We used Krippendorff's diagnostics to determine which particular categories humans had most problems with: for each category, agreement is measured with a new data set where all categories except for the category of interest are collapsed into one meta-category. Original agreement is compared to that measured on the new (artificial) data set; high values show that annotators can distinguish the given category well from all others. When compared to the overall reproducibility of K=.71, the annotators were good at distinguishing AIM (Krippendorff's diagnostics; K=.79) and TEXTUAL (K=.79). The high agreement in AIM sentences is a positive result which seems to be at odds with previous sentence extraction experiments. We take this as an indication that some types of rhetorical classification are easier for human minds to do than unqualified relevance decision. We also think that the positive results are partly due to the existence of the guidelines.

The annotators were less consistent at determining BASIS (K=.49) and CONTRAST

		ANNOTATOR B							
		AIM	CTR	TXT	OWN	BKG	BAS	OTH	Total
ANNOTATOR C	AIM	**35**	2	1	19	3		2	62
	CTR		**86**		31	16		23	156
	TXT			**31**	7			1	39
	OWN	10	62	5	**2298**	25	3	84	2487
	BKG		5		13	**115**		20	153
	BAS	2			18	1	**18**	14	53
	OTH	1	18	2	55	10	1	**412**	499
	Total	48	173	39	2441	170	22	556	3449

Figure 9
Confusion matrix between annotators B and C

(K=.59). This same picture emerges if we look at precision and recall of single categories between two annotators (cf. figure 10). Precision and recall for AIM and TEXTUAL are high at 72%/56% and 79%/79%, whereas they are lower for CONTRAST (50%/55%) and BASIS (82%/34%).

This contrast in agreement might have to do with the location of the rhetorical zones in the paper: AIM and TEXTUAL zones are usually found in fixed locations (beginning or end of the introduction section) and are explicitly marked with meta-discourse, whereas CONTRAST sentences, and even more so BASIS sentences, are usually interspersed within longer OWN zones. As a result, these categories are more exposed to lapses of attention during annotation.

	AIM	CTR	TXT	OWN	BKG	BAS	OTH
Precision	72%	50%	79%	94%	68%	82%	74%
Recall	56%	55%	79%	92%	75%	34%	83%

Figure 10
Annotator C's precision and recall per category if annotator B is gold standard

With respect to the longer, more neutral zones (intellectual attribution), annotators often had problems in distinguishing OTHER work from OWN work, particularly in cases where the authors did not express a clear distinction between *current, new work* and *previous own work* (which, according to our instructions, should be annotated as OTHER). Another persistently problematic distinction for our annotators was that between OWN and BACKGROUND. This could be a sign that some authors aimed their papers at an expert audience, and thus thought it unnecessary to signal clearly which statements are commonly agreed in the field, as opposed to their own new claims. If a paper is written in such a way, it can indeed only be understood with a considerable amount of domain knowledge, which our annotators did not have.

Because intellectual attribution (the distinction between OWN, OTHER and BACKGROUND material) is an important part of our annotation scheme, we conducted a second experiment measuring how well our annotators could distinguish just these three roles, using the same annotators and 22 different articles. We wrote new guidelines of 7 pages describing the semantics of the three categories. Results show higher stability compared to the full annotation scheme (K=.83, .79, .81; N=1248; k=2) and higher reproducibility (K=.78, N=4031, k=3), corresponding to 94%, 93% and 93% percentage agreement (stability) and 93% (reproducibility). It is most remarkable that agreement of annotation of intellectual attribution in the abstracts is almost perfect: K=.98 (N=89, k=3), corresponding to 99% agreement. This points to the fact that authors, when writing the abstracts, take care to make it clear who a certain statement is attributed to. This effect also holds for the annotation with the full scheme with all seven categories: again, reproducibility in the abstract is higher (K=.79) than in the entire document (K=.71), but the effect is much weaker.

Abstracts might be easier to annotate than the rest of the paper, but this does not necessarily make it possible to define a gold standard by only looking at the abstracts. As fore-shadowed in section 2.5, abstracts do not contain all types of rhetorical information. AIM and OWN sentences make up 74% of the sentences in abstracts, and only 5% of all Contrast sentences and 3% of all BASIS sentences occur in the abstract.

Abstracts in our corpus are also not structurally homogeneous. When we inspected the rhetorical structure of abstracts in terms of sequences of rhetorical zones, we found a high level of variation. Even though the sequence AIM–OWN is very common (contained in 73% of all abstracts), the 80 abstracts still contain 40 different rhetorical sequences, 28 of which are unique. This heterogeneity is in stark contrast to the systematic structures Liddy (1991) found to be produced by professional abstractors. Both observations, the lack of certain rhetorical types in the abstracts and their rhetorical heterogeneity, reassure us in our decision not to use human-written abstracts as a gold standard.

3.3 Annotation of Relevance

We collected two different kinds of relevance gold standards for the documents in our development corpus: abstract-similar document sentences, and additional manually selected sentences.

In order to establish alignment between summary and document sentences, we used a semi-automatic method, which relies on a simple surface similarity measure (longest common subsequence of content words, i.e. excluding words on a stop list). As in Kupiec et al.'s experiment, final alignment was decided by a human judge, where the criterion was semantic similarity of the two sentences. The following sentence pair illustrates a *direct match*:

> **Summary:** In understanding a reference, an agent determines his confidence in its adequacy as a means of identifying the referent.
>
> **Document:** An agent understands a reference once he is confident in the adequacy of its (inferred) plan as a means of identifying the referent.

Of the 346 abstract sentences contained in the 80 documents, 156 (45%) could be aligned this way. Because of this low agreement, and the fact that certain rhetorical types are not present in the abstracts, we decided not to rely on abstract alignment as our only gold standard. Instead, we use manually selected sentences as an alternative gold standard, which is more informative, but also more subjective.

We wrote 8 pages of guidelines which describe relevance criteria; e.g., our definition prescribes to select neutral descriptions of other work only if the other work is an essential part of the solution presented, whereas *all* statements of criticism are to be included. The first author annotated all documents in the development corpus with relevance using the rhetorical zones and abstract similarity as aides in the relevance decision, and also skim-reading the whole paper before making the decision. This resulted in 5 to 28 sentences per paper and a total of 1183 sentences.

Implicitly, rhetorical classification of the extracted sentences was already given as each of these sentences already had a rhetorical status assigned to it. However, the rhetorical scheme we use for this task is slightly different. We excluded TEXTUAL, as this category was designed for document uses other than summarisation. If a selected sentence had the rhetorical class TEXTUAL, it was reclassified into one of the other six categories. Figure 11 shows the resulting category distribution amongst these 1183 sentences, which is far more evenly distributed than the one covering *all* sentences (cf. figure 8). CONTRAST and OWN are the two most frequent categories.

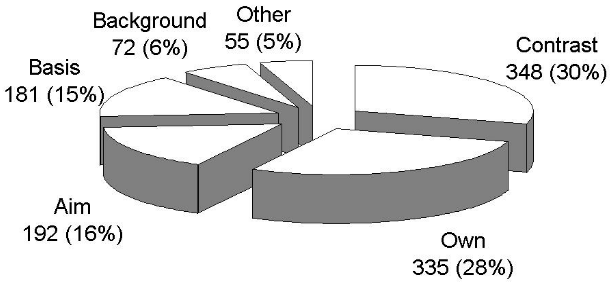

Figure 11
Distribution of rhetorical categories (relevant sentences)

We did not verify the relevance annotation with human experiments. We accept that the set of sentences chosen by the human annotator is only *one* possible gold standard. What is more important is that humans can agree on the rhetorical status of the relevant sentences. Liddy observed that agreement on rhetorical status was easier for professional abstractors than sentence selection: while they did not necessarily agree which individual sentences should go into an abstract, they did agree on the rhetorical information types that make up a good abstract.

We asked our trained annotators to classify a set of 200 sentences, randomly sampled from the 1183 sentences selected by the first author, into the six rhetorical categories. The sentences were presented in order of occurrence in the document, but without any context in terms of surrounding sentences. We measured stability at K=.9,.86,.83 (N=100, k=2) and reproducibility at K=.84 (N=200, k=3). These results are reassuring: they show that the rhetorical status for *important* sentences can be particularly well determined, better than rhetorical status for *all* sentences in the document (where reproducibility was K=.71, cf. section 3.2.2).

4 The System

We now describe an automatic system which can perform extraction and classification of rhetorical status on unseen text (cf. also a prior version of the system reported in Teufel and Moens (2000) and Teufel (1999)). We decided to use machine learning to do so, based on a variety of sentential features similar to the ones reported in the sentence extraction literature. Human annotation is used as training material such that the associations between these sentential features and the target sentences can be learned. It is also used as gold standard for intrinsic system evaluation.

A simpler machine learning approach using only word frequency information and no other features, as typically used in tasks like text classification, could have been employed (and indeed Nanba and Okumura (1999) do so for classifying citation contexts). To test if such a simple approach would be enough, we performed a text categorisation experiment, using the Rainbow implementation of a naïve Bayes *TF*IDF* method (McCallum, 1997), and considering each sentence as a "document". The result was a classification performance of K=.30; the classifier nearly almost chooses OWN and OTHER segments. The rare but important categories AIM, BACKGROUND, CONTRAST and BASIS could only be retrieved with low precision and recall. Therefore, text classification methods do not provide a solution to our problem. This is not surprising, given that the definition of our task has little to do with the distribution of "content-bearing" words and phrases, much less so than the related task of topic segmentation (Morris and Hirst, 1991; Hearst, 1997; Choi, 2000), or Saggion and Lapalme's (2000) approach to the summarisation of scientific articles, which relies on scientific concepts and their relations. Instead, we predict that other indicators apart from the simple words contained in the sentence could provide strong evidence for the modelling of rhetorical status. Also, the relatively small amount of training material we have at our disposal requires a machine learning method which makes optimal use of as many different kinds of features as possible. We predicted that this would increase precision and recall on the categories we are interested in. The text classification experiment is still useful as it provides a non-trivial baseline for comparison with our intrinsic system evaluation presented in section 5.

4.1 Classifiers

We use a naïve Bayesian model as in Kupiec et al.'s (1995) experiment, cf. figure 12.

$$P(C|F_0,\ldots,F_{n-1}) \approx \; P(C)\ \frac{\prod_{j=0}^{n-1} P(F_j|C)}{\prod_{j=0}^{n-1} P(F_j)}$$

$P(C\|F_0,\ldots,F_{n-1})$:	Probability that a sentence has target category C, given its feature values $F_0,\ldots,F_{n-1}$;
$P(C)$:	(Overall) probability of category C;
$P(F_j\|C)$:	Probability of feature-value pair F_j, given that the sentence is of target category C;
$P(F_j)$:	Probability of feature value F_j;

Figure 12
Naïve Bayesian Classifier

Sentential features are collected for each sentence (figure 13 gives an overview of the features we used). Learning is supervised: in the training phase, associations between these features and human-provided target-categories are learned. The target categories

Type	Name	Feature description	Feature values
Absolute Location	`Loc`	Position of sentence in relation to 10 segments	A-J
Explicit Structure	`Section Struct`	Relative and absolute position of sentence within section (e.g., first sentence in section or somewhere in second third)	7 values
	`Para Struct`	Relative position of sentence within a paragraph	Initial, Medial, Final
	`Headline`	Type of headline of current section	15 prototypical headlines or *Non-Prototypical*
Sentence length	`Length`	Is the sentence longer than a certain threshold, measured in words?	Yes or No
Content Features	`Title`	Does the sentence contain words also occurring in the title or headlines?	Yes or No
	`TF*IDF`	Does the sentence contain "significant terms" as determined by the *TF*IDF* measure?	Yes or No
Verb Syntax	`Voice`	Voice (of first finite verb in sentence)	Active or Passive or NoVerb
	`Tense`	Tense (of first finite verb in sentence)	9 simple and complex tenses or NoVerb
	`Modal`	Is the first finite verb modified by modal auxiliary?	Modal or no Modal or NoVerb
Citations	`Cit`	Does the sentence contain a citation or the name of an author contained in the reference list? If it contains a citation, is it a self citation? Whereabouts in the sentence does the citation occur?	{Citation (self), Citation (other), Author Name, or None} X {Beginning, Middle, End}
History	`History`	Most probable previous category	7 Target Categories + "BEGIN"
Meta-discourse	`Formulaic`	Type of formulaic expression occurring in sentence	18 Types of Formulaic Expressions + 9 Agent Types or None
	`Agent`	Type of Agent	9 Agent Types or None
	`SegAgent`	Type of Agent	9 Agent Types or None
	`Action`	Type of Action, with or without Negation	27 Action Types or None

Figure 13
Overview of feature pool

are the 7 categories in the Rhetorical Annotation experiment, and relevant/non-relevant in the Relevance Selection experiment. In the testing phase, the trained model provides the probability of each target category for each sentence of unseen text, on the basis of the sentential features identified for the sentence.

4.2 Features

Some of the features in our feature pool are unique to our approach, for instance the meta-discourse features. Others are borrowed from the text extraction literature (Paice, 1990) or related tasks and adapted to the problem of determining rhetorical status.

Absolute location of a sentence: In the news domain, sentence location is the single most important feature for sentence selection (Brandow, Mitze, and Rau, 1995); in our domain, location information, while less dominant, can still give a useful indication. Rhetorical zones appear in typical positions in the article, as scientific argumentation follows certain patterns (Swales, 1990). For example, limitations of the author's own method can be expected to be found towards the end of the article, whereas limitations of *other* researchers' work often occur in the introduction. We observed that the size of rhetorical zones depends on location, with smaller rhetorical zones occurring towards the beginning and the end of the article. We model this by assigning location values in the following fashion: the article is divided into 20 equal parts, counting sentences. Sentences occurring in parts 1, 2, 3, 4, 19, 20 receive the values A, B, C, D, I, J respectively. Parts 5 and 6 are pooled, and sentences occurring in them are given the value E; the same procedure is applied to parts 15 and 16 (value G) and 17 and 18 (value H). The remaining sentences in the middle (parts 7–14) all receive the value F, cf. figure 14.

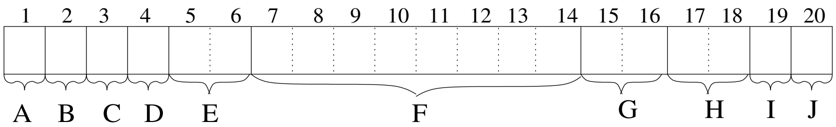

Figure 14
Values of location feature

Section structure: Sections can have an internal structuring; for instance, sentences towards the beginning of a section often have a summarising function. The section location feature divides each section into three parts and assigns 7 values: first sentence, last sentence, second or third sentence, second-last or third-last sentence, or else either somewhere in the first, second or last third of the section.

Paragraph structure: In many genres, paragraphs also have internal structure (Wiebe, 1994), with high-level or summarising sentences occurring more often at the periphery of paragraphs. In this feature, sentences are distinguished into those leading or ending a paragraph, and all others.

Headlines: Prototypical headlines can be an important predictor of the rhetorical status of sentences occurring in the given section; however, not all texts in our collection use such headlines. Whenever a prototypical headline is recognised (using a set of regular expressions), it is classified into one of the following 15 classes: *Introduction, Implementation, Example, Conclusion, Result, Evaluation, Solution, Experiment, Discussion, Method, Problems, Related Work, Data, Further Work, Problem Statement*. If none of the patterns match, the value *Non-Prototypical* is assigned.

Sentence length: Kupiec et al. (1995) report sentence length as a useful feature for text extraction. In our implementation, sentences are divided into long or short sentences, by comparison to a fixed threshold (12 words).

Title word contents: Sentences containing many "content-bearing" words have been hypothesised to be good candidates for text extraction. Baxendale (1958) extracted all words except those on the stop-list from the title and the headlines and determined for each sentence if it contained these words or not. We received better results by excluding headline words and only using title words.

*TF*IDF word contents:* How content-bearing a word is can also be measured with frequency counts (Salton and McGill, 1983). The *TF*IDF* formula assigns high values to words which occur frequently in one document, but rarely in the overall collection of documents. We use the 18 highest scoring *TF*IDF* words, and classify sentences into those that contain one or more of these words, and those that do not.

Verb syntax: Linguistic features like tense and voice often correlate with rhetorical zones; Biber (1995) and Riley (1991) show correlation of tense and voice with prototypical section structure ("method", "introduction"). In addition, the presence or absence of a modal auxiliary might be relevant to detect the phenomenon of "hedging" (i.e., statements in which an author distances herself from her claims or signals low certainty *"these results might indicate that ... possibly ..."* (Hyland, 1998)). For each sentence, we use part of speech-based heuristics to determine tense, voice and presence of modal auxiliaries. This algorithm is shared with the meta-discourse features, and the details are described below (cf. p.23).

Citation: There are many connections between citation behaviour and relevance or rhetorical status. First, if a sentence contains a formal citation or the name of another author mentioned in the bibliography, it is far more likely to talk about other work than about own work. Second, if it contains a self citation, it is far more likely to contain a direct statement of continuation (25%) than a criticism (3%). Third, the importance of a citation has been related to the distinction between authorial and parenthetical citations. Citations are called authorial if they form a syntactically integral part of the sentence, or parenthetical if they do not (Swales, 1990). In most cases, authorial citations are used as the subject of a sentence, and parenthetical ones appear towards the middle or the end of the sentence.

We automatically recognise formal citations. We also parse the reference list at the end of the article, determine if a citation is a self-citation (i.e., if there is an overlap between the names of the cited researchers and the authors of the current paper), and we additionally find occurrences of authors' names in running text, but outside of formal citation contexts (e.g., *"Chomsky also claims that ..."*). The citation feature reports if a sentence contains an author name, a citation or nothing. If it contains a citation, the value records whether it is a self-citation, and also records the location of the citation in the sentence (in the beginning, the middle, or the end). This last distinction is a heuristic for the authorial/parenthetical distinction. We also experimented with including the *number* of different citations in a sentence, but this did not improve results.

History: As there are typical patterns in the rhetorical zones (e.g., AIM sentences tend to follow CONTRAST sentences), we wanted to include the category assigned to the previous sentence as one of the features. However, in unseen text, the previous target is unknown at training time (it is determined during testing). It can, however, be calculated as a second pass process during training. In order to avoid a full Viterbi search of all possibilities, we perform a beam search with width of 3 amongst the candidates of the previous sentence, following Barzilay et al. (2000).

Formulaic expressions: We now turn to the last three features in our feature pool, the meta-discourse features, which are more sophisticated than the other features. The first meta-discourse feature models formulaic expressions like the ones described by Swales, as they are semantic indicators which we expect to be helpful for rhetorical classification. We use a list of phrases described by regular expressions, similar to Paice's (1990) grammar. Our list is divided into 18 semantic classes (cf. figure 15), comprising a total of 640 patterns. The fact that phrases are clustered is a simple way of dealing with data sparseness. In fact, our experiments in section 5.1.2 will show the usefulness of our (manual) semantic clusters: the clustered list performs much better than the unclustered list (i.e., when the string itself is used as a value instead of its semantic class).

Agent: Agents and actions are more challenging to recognize. We use a mechanism which, dependent on the voice of a sentence recognises agents (subjects or prepositional

Indicator Type	Example	No
GAP_INTRODUCTION	*to our knowledge*	3
GENERAL_FORMULAIC	*in traditional approaches*	10
DEIXIS	*in this paper*	11
SIMILARITY	*similar to*	56
COMPARISON	*when compared to our*	204
CONTRAST	*however*	6
DETAIL	*this paper has also*	4
METHOD	*a novel method for* ***VERB****–ing*	33
PREVIOUS_CONTEXT	*elsewhere, we have*	25
FUTURE	*avenue for improvement*	16
AFFECT	*hopefully*	4
CONTINUATION	*following the argument in*	19
IN_ORDER_TO	*in order to*	1
POSITIVE_ADJECTIVE	*appealing*	68
NEGATIVE_ADJECTIVE	*unsatisfactory*	119
THEM_FORMULAIC	*along the lines of*	6
TEXTSTRUCTURE	*in section 3*	16
NO_TEXTSTRUCTURE	*described in the last section*	43
Total of 18 classes		640

Figure 15
Formulaic expression lexicon

phrases headed by "*by*") and their predicates ("actions"). Classification of agents and actions relies on a manually created lexicon of manual classes. Like in the `Formulaic` feature, similar agents and actions are generalised and clustered together to avoid data sparseness.

The lexicon for agent patterns (cf. figure 16) contains 13 types of agents and a total of 167 patterns. These 167 patterns expand to many more strings as we use a replace mechanism (e.g., the place holder **WORK_NOUN** in the 6th row of figure 16 can be replaced by a set of 37 nouns including *"theory, method, prototype, algorithm"*).

The main three agent types we distinguish are US_AGENT, THEM_AGENT and GENERAL_AGENT, following the types of intellectual attribution discussed above. A fourth type is US_PREVIOUS_AGENT (the authors, but in a *previous* paper).

Agent Type	Example	No	Removed
US_AGENT	*we*	22	
THEM_AGENT	*his approach*	21	
GENERAL_AGENT	*traditional methods*	20	X
US_PREVIOUS_AGENT	*the approach in* **SELFCITE**	7	
OUR_AIM_AGENT	*the point of this study*	23	
REF_US_AGENT	*this method (this* **WORK_NOUN***)*	6	
REF_AGENT	*the paper*	11	
THEM_PRONOUN_AGENT	*they*	1	X
AIM_REF_AGENT	*its goal*	8	
GAP_AGENT	*none of these papers*	8	
PROBLEM_AGENT	*these drawbacks*	3	X
SOLUTION_AGENT	*a way out of this dilemma*	4	X
TEXTSTRUCTURE_AGENT	*the concluding chapter*	33	
Total of 13 classes		167	

Figure 16
Agent lexicon

Additional agent types include non-personal agents like aims, problems, solutions, absence of solution, or textual segments. There are four equivalence classes of agents

Action Type	Example	No	Removed
AFFECT	*we hope to improve our results*	9	X
ARGUMENTATION	*we argue against a model of*	19	X
AWARENESS	*we are not aware of attempts*	5	+
BETTER_SOLUTION	*our system outperforms ...*	9	−
CHANGE	*we extend* **CITE**'*s algorithm*	23	
COMPARISON	*we tested our system against ...*	4	
CONTINUATION	*we follow* **CITE** ...	13	
CONTRAST	*our approach differs from ...*	12	−
FUTURE_INTEREST	*we intend to improve ...*	4	X
INTEREST	*we are concerned with ...*	28	
NEED	*this approach, however, lacks ...*	8	X
PRESENTATION	*we present here a method for ...*	19	−
PROBLEM	*this approach fails ...*	61	−
RESEARCH	*we collected our data from ...*	54	
SIMILAR	*our approach resembles that of*	13	
SOLUTION	*we solve this problem by ...*	64	
TEXTSTRUCTURE	*the paper is organized ...*	13	
USE	*we employ* **CITE**'*s method ...*	5	
COPULA	*our goal is to ...*	1	
POSSESSION	*we have three goals ...*	1	
Total of 20 classes		365	

Figure 17
Action lexicon

with ambiguous reference (*"this system"*), namely REF_AGENT, REF_US_AGENT, THEM_PRONOUN_AGENT, and AIM_REF_AGENT.

Agent classes were created based on intuition, but subsequently each class was tested with corpus statistics, to determine if a certain agent class should be removed or not. We wanted to find and exclude such classes which had a distribution very similar to the overall distribution of the target categories, as such features are not distinctive. We measured associations using the log-likelihood measure (Dunning, 1993) for each combination of target category and semantic class by converting each cell of the contingency into a 2X2 contingency table. We kept only classes of verbs where at least one category showed a high association (gscore > 5.0), as that means that in these cases the distribution was significantly different from the overall distribution. The last column in figure 16 shows that the classes THEM_PRONOUN, GENERAL, SOLUTION, PROBLEM and REF were removed; removal improved the performance of the `Agent` feature.

SegAgent: This is a variant of the `Agent` feature which keeps track of previously recognised agents; unmarked sentences receive these previous agents as a value (in the `Agent` feature, they would have received the value *None*).

Action: We use a manually created action lexicon containing 365 verbs (cf. figure 17). The verbs are clustered into 20 classes based on semantic concepts such as similarity, contrast, competition, presentation, argumentation and textual structure. For example, PRESENTATION_ACTIONs include communication verbs like *"present"*, *"report"*, *"state"* (Myers, 1992; Thompson and Yiyun, 1991), RESEARCH_ACTIONS include *"analyze"*, *"conduct"*, *"define"* and *"observe"*, and ARGUMENTATION_ACTIONS *"argue"*, *"disagree"*, *"object to"*. Domain-specific actions are contained in the classes indicating a problem (*"fail"*, *"degrade"*, *"waste"*, *"overestimate"*), and solution-contributing actions (*"circumvent"*, *solve"*, *"mitigate"*). The recognition of negation is essential; the semantics of *"not solving"* is closer to *"being problematic"* than it is to *"solving"*.

The following classes were removed by the gscore test described above, because their distribution was too similar to the overall distribution: FUTURE_INTEREST, NEED, ARGUMENTATION, AFFECT both in negative and positive contexts (X in last column of figure 17), and AWARE only in positive context ("+" in last column). The following classes had few occurrences in a negative context (<10 occurrences in the whole verb class) and were thus also removed: BETTER_SOLUTION, CONTRAST, PRESENT, PROBLEM ("–" in last column). Again, the removal improved the performance of the `Action` feature.

The algorithm for determining agents and actions relies on finite state patterns over part of speech (POS) tags. Starting from each finite verb, the algorithm collects chains of auxiliaries belonging to the associated finite clause, and thus determines the clause's tense and voice. Other finite verbs and commas are assumed to be clause boundaries. Once the semantic verb is found, its stem is looked up in the action lexicon. Negation is determined if one of 32 fixed negation words is present in a 6 word window to the right of the finite verb.

As our classifier requires one unique value for each classified item for each feature, we had to choose one value for sentences containing more than one finite clause. We return the following values for the action and agents feature: the first agent/action pair, if both are non-zero, otherwise the first agent without an action, otherwise the first action without an agent, if available.

In order to determine the level of correctness of agent and action recognition, we had to first manually evaluate the error level of the POS-Tagging of finite verbs, as our algorithm crucially relies on finite verbs. In a random sample of 100 sentences from our development corpus which contain finite verbs at all (they happened to contain a total of 184 finite verbs), the tagger (which is part of the TTT software) showed a recall of 95% and a precision of 93%.

We found that for the 174 correctly determined finite verbs, the heuristics for negation and presence of modal auxiliaries worked without any errors (100% accuracy, 8 negated sentences). The correct semantic verb was determined with 96% accuracy; most errors are due to misrecognition of clause boundaries. Action Type lookup was fully correct (100% accuracy), even in the case of phrasal verbs and longer idiomatic expressions ("*have to*" is a NEED_ACTION; "*be inspired by*" is a CONTINUE_ACTION). There were 7 voice errors, 2 of which were due to POS-tagging errors (past participle misrecognised). The remaining 5 voice errors correspond to 98% accuracy.

Correctness of Agent Type determination was tested on a random sample of 100 sentences containing at least one agent, resulting in 111 agents. No agent pattern that should have been identified was missed (100% recall). Of the 111 agents, 105 cases were correct (precision of 95%). Therefore, we consider the two features to be adequately robust to serve as sentential features in our system.

After detailing the features and classifiers of the machine learning system we use, we will now turn to an intrinsic evaluation of its performance.

5 Intrinsic System Evaluation

Our task is to perform content selection from scientific articles, which we do by classifying sentences into seven rhetorical categories. The summaries based on this classification use some of these sentences directly, namely sentences which express the contribution of a particular article (AIM), sentences expressing contrasts with other work (CONTRAST) and sentences stating imported solutions from other work (BASIS). Other, more frequent rhetorical categories, namely OTHER, OWN and BACKGROUND, might also be extracted into the summary.

Because the task is a mixture of extraction and classification, we report system performance as follows:

- We first report precision and recall values for all categories, in comparison to human performance and the text categorisation baseline, as we are primarily interested in good performance on the categories AIM, CONTRAST, BASIS and BACKGROUND.
- We are also interested in good overall classification performance, which we report using Kappa and Macro-F as our metric. We also discuss how well each single features does in the classification.
- We then compare the extracted sentences to our human gold standard for *relevance*, and report the agreement in precision and agreement per category.

5.1 Determination of Rhetorical Status

The results of stochastic classification were compiled with a 10-fold cross-validation on our 80-paper corpus. As we do not have much annotated material, cross-validation is a practical way to test as it can make use of the full development corpus for training, without ever using the same data for training and testing.

	AIM			CONTR.			TEXTUAL			OWN			BACKGR.			BASIS			OTHER		
	f	p	r	f	p	r	f	p	r	f	p	r	f	p	r	f	p	r	f	p	r
System	52	44	65	26	34	20	61	57	66	86	84	88	45	40	50	38	37	40	44	52	39
Baseline	11	30	7	17	31	12	23	56	15	83	78	90	22	32	17	7	15	5	44	47	42
Humans	63	72	56	52	50	55	79	79	79	93	94	92	71	68	75	48	82	34	78	74	83

Figure 18
Performance per category: F-measure, precision and recall

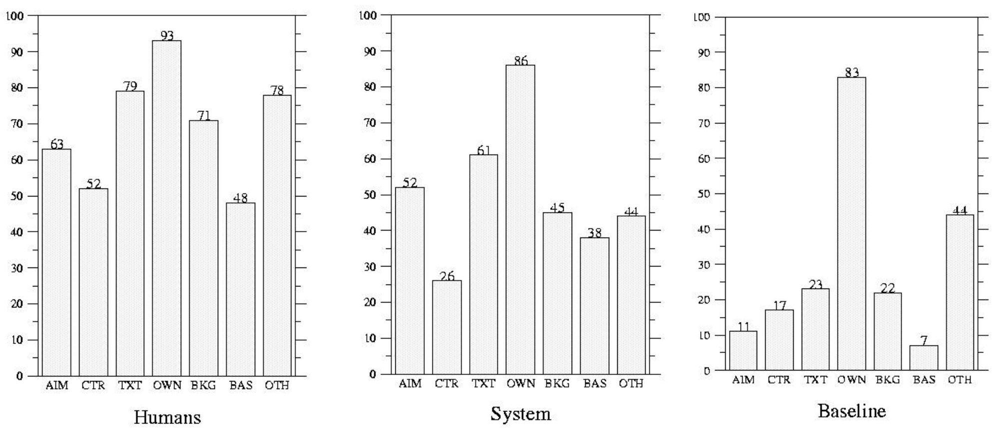

Figure 19
Performance per category: F-measure

5.1.1 Overall Results

Figure 18 and 19 show that the stochastic model obtains substantial improvement over the baseline in terms of precision and recall of the important categories AIM, BACKGROUND, CONTRAST and BASIS. We use the F-measure, defined by van Rijsbergen (1979) as $\frac{2PR}{P+R}$, as a convenient way of reporting precision (P) and recall (R) in one value. F-measures for our categories range from .61 (TEXTUAL), .52 (AIM) to .45 (BACKGROUND), .38 (BASIS) and .26 (CONTRAST). The recall for some categories is relatively low. As our gold standard is designed to contain a lot of redundant information for the same category, this is not too worrying. However, low precision in some categories (e.g., 34% for CONTRAST, in contrast to human precision of 55%) could potentially present a problem for later steps in the document summarisation process.

Overall, we find these results encouraging, particularly in view of the subjective nature of the task and the high compression achieved (2% for AIM, BASIS and TEXTUAL sentences, 5% for CONTRAST sentences and 6% for BACKGROUND sentences). No direct comparison with Kupiec et al.'s results is possible as different data sets are used and as their relevant sentences do not directly map into one of our categories. Assuming, however, that their relevant sentences are probably most comparable to our AIM sentences, our precision and recall of 44% and 65% compare favourably to theirs (42% and 42%).

		MACHINE							
		AIM	CTR	TXT	OWN	BKG	BAS	OTH	Total
HUMAN	AIM	**127**	6	13	23	19	5	10	203
	CTR	21	**112**	4	204	87	18	126	572
	TXT	14	1	**145**	46	6	2	6	220
	OWN	100	108	84	**7231**	222	71	424	8240
	BKG	14	31	1	222	**370**	5	101	744
	BAS	17	7	7	60	8	**97**	39	235
	OTH	6	70	10	828	215	72	**773**	1974
	Total	299	335	264	8614	927	270	1479	12188

Figure 20
Confusion matrix: human v. automatic annotation

Figure 20 shows a confusion matrix between one annotator and the system. The system is likely to confuse AIM and OWN sentences (e.g. 100 out of 172 sentences incorrectly classified as AIM by the system turned out to be indeed OWN sentences). It also shows a tendency to confuse OTHER and OWN sentences. The system also fails to distinguish categories involving other people's work, e.g. OTHER, BASIS and CONTRAST. Overall, these tendencies mirror human errors, as can be seen from a comparison with figure 9.

Figure 21 shows the results in terms of three overall measures: Kappa, percentage accuracy, and Macro-F (following Lewis (1991)). Macro-F is the mean of the F-measures of all seven categories. One reason for using Macro-F and Kappa is that we want to measure success particularly on the rare categories which are needed for our final task, i.e. AIM, BASIS and CONTRAST. Micro-averaging techniques like traditional accuracy tend to overestimate the contribution of frequent categories in skewed distributions like ours; this is undesirable, as OWN is the least interesting category for our purposes. This situation has parallels in information retrieval, where precision and recall are used because accuracy overestimates the performance on irrelevant items.

In the case of Macro-F, each category is treated as one unit, independent of the number of items contained in it. Therefore, the classification success of the individual items in rare categories is given more importance than classification success of frequent category items. However, when looking at the numerical values one should keep in mind that macro-averaging results are in general lower numerically (Yang and Liu, 1999). This is due to the fact that there are fewer training cases for the rare categories, which therefore perform worse with most classifiers.

In the case of Kappa, classifications which incorrectly favour frequent categories are punished due to a high random agreement. This effect can be shown most easily when considering the baselines. The most ambitious baseline we use is the output of a text categorisation system, as described in section 4. Other possible baselines, which are all easier to beat, include classification by the most-frequent category. This baseline turns out to be trivial, as it does not extract *any* of the rare rhetorical categories we are particularly interested in, and therefore receives a low Kappa value at K=–.12. Possible chance baselines include random annotation with uniform distribution (K=–.10; accuracy of 14%) and random annotation with observed distribution. The latter baseline is built into the definition of Kappa (K=0; accuracy of 48%).

While our system outperforms an ambitious baseline (Macro-F shows that our system performs roughly 20% better than text classification), and also performs much above chance, there is still a big gap in performance between humans and machine. Macro-F shows a 20% difference between our system and human performance. Kappa shows that if the system is put into a pool of annotators for the 25 articles for which 3-way human judgement exists, agreement drops from K=.71 to K=.59, which is a clear indication that the system's annotation is still distinguishably different from human annotation.

	System/Baseline Compared with One Human Annotator					3 Humans
	System	Text Class.	Random	Random (Distr.)	Most Freq.	
Kappa	.45	.30	–.10	0	–.13	.71
Accuracy	.73	.72	.14	.48	.67	.87
Macro-F	.50	.30	.09	.14	.11	.69

Figure 21
Overall classification results

5.1.2 Feature Impact

The previous results were compiled using *all* features, which is the optimal feature combination (as determined by an exhaustive search in the space of feature combinations). The most distinctive single feature is `Location` (achieving an agreement of K=.22 against one annotator, if this feature is used as the sole feature), followed by `SegAgent` (K=.19), `Citations` (K=.18), `Headlines` (K=.17), `Agent` (K=.08) and `Formulaic` (K=.07). In each case, the unclustered versions of `Agent`, `SegAgent` and `Formulaic` performed much worse than the clustered versions; they did not improve final results when added into the feature pool.

`Action` performs slightly better at K=-.11 than the baseline by most frequent category, but far worse than random by observed distribution. The following features on their own classify each sentence as OWN (and therefore achieve K=–.12): `Relative Location`, `Paragraphs`, `TF*IDF`, `Title`, `Sentence Length`, `Modality`, `Tense`, `Voice`. `History` performs very badly on its own at K=–.51; it classifies almost all sentences as BACKGROUND. This is so because the probability of the first sentence being a BACKGROUND sentence is almost 1, and, if no other information is available, it is it very likely that another BACKGROUND sentence will follow after a BACKGROUND sentence.

However, each of these features still contributes to the final result: if either of them is taken out of the feature pool, classification performance decreases. How can this be, given that the individual features perform worse than chance? As the classifier derives the posterior probability by multiplying evidence from each feature, even slight evidence coming from one feature can direct the decision into the right direction. A feature which contributes little evidence on its own (too little to break the prior probability, which is strongly biased towards OWN), can thus, in combination with others, still help disambiguating. For the naïve Bayesian classification method, indeed, it is most important that the features be as independent of each other as possible. This property cannot be assessed by looking at the feature's isolated performance, but only in combination with others.

Features	Precision/Recall per Category (in %)						
	AIM	CONTR.	TXT.	OWN	BACKG.	BASIS	OTHER
`SegAgent` alone	—	17/0	—	74/94	53/16	—	46/33
`Agent` alone	—	—	—	71/93	—	—	36/23
`Location` alone	—	—	—	74/97	40/36	—	28/9
`Headlines` alone	—	—	—	75/95	—	—	29/25
`Citation` alone	—	—	—	73/96	—	—	43/30
`Formulaic` alone	40/2	45/2	75/39	71/98	—	40/1	47/13
`Action` alone	—	43/1	—	68/99	—	—	—
`History` alone	—	—	—-	70/8	16/99	—	—

Figure 22
Precision and recall of rhetorical classification, individual features

It is also interesting to see that certain categories are disambiguated particularly well by certain features (cf. figure 22). The `Formulaic` feature, which is by no means the strongest feature, is nevertheless the most diverse, as it contributes to the disambiguation of six categories directly. This is due to the fact that many different rhetorical categories have typical cue phrases associated with them (whereas not all categories might have a preferred location in the document). Not surprisingly, `Location` and `History` are the features particularly useful for detecting BACKGROUND sentences, and `SegAgent` additionally contributes towards the determination of BACKGROUND zones (along with the `Formulaic` and the `Absolute Location` feature). The `Agent` and `Action` features also prove their worth as they manage to disambiguate categories which many of the other 12 features alone cannot disambiguate (e.g., CONTRAST).

5.1.3 System Output: the Example Paper

In order to give the reader an impression of how the figures reported in the previous section translate into real output, we present in figure 23 the output of the system when run on the example paper (all AIM, CONTRAST and BASIS sentences). The second column shows whether the human judge agrees with the systems decision (a tick for correct decisions, and the human's preferred category for incorrect decisions). 10 out of the 15 extracted sentences have been classified correctly.

The example also shows that the determination of rhetorical status is not always straightforward. For example, whereas the first AIM sentence which the system proposes (sentence **8**) is clearly wrong, all other "incorrect" AIM sentences carry important information about research goals of the paper: Sentence **41** states the goal in explicit

System	Human		
AIM	(OTH)	**8**	*In Hindle's proposal, words are similar if we have strong statistical evidence that they tend to participate in the same events.*
	√	*** 10**	*Our research addresses some of the same questions and uses similar raw data, but we investigate how to factor word association tendencies into associations of words to certain hidden senses classes and associations between the classes themselves.*
	√	**11**	*While it may be worthwhile to base such a model on preexisting sense classes (Resnik, 1992), in the work described here we look at how to derive the classes directly from distributional data.*
	(OWN)	**12**	*More specifically, we model senses as probabilistic concepts or clusters c with corresponding cluster membership probabilities* **EQN** *for each word w.*
	√	*** 22**	*We will consider here only the problem of classifying nouns according to their distribution as direct objects of verbs; the converse problem is formally similar.*
	(CTR)	**41**	*However, this is not very satisfactory because one of the goals of our work is precisely to avoid the problems of data sparseness by grouping words into classes.*
	(OWN)	**150**	*We also evaluated asymmetric cluster models on a verb decision task closer to possible applications to disambiguation in language analysis.*
	√	*** 162**	*We have demonstrated that a general divisive clustering procedure for probability distributions can be used to group words according to their participation in particular grammatical relations with other words.*
BAS	√	**19**	*The corpus used in our first experiment was derived from newswire text automatically parsed by Hindle's parser Fidditch (Hindle, 1993).*
	√	**20**	*More recently, we have constructed similar tables with the help of a statistical part-of-speech tagger (Church, 1988) and of tools for regular expression pattern matching on tagged corpora (Yarowsky, 1992).*
	√	*** 113**	*The analogy with statistical mechanics suggests a deterministic annealing procedure for clustering (Rose et al., 1990), in which the number of clusters is determined through a sequence of phase transitions by continuously increasing the parameter* **EQN** *following an annealing schedule.*
CTR	√	*** 9**	*His notion of similarity seems to agree with our intuitions in many cases, but it is not clear how it can be used directly to construct word classes and corresponding models of association.*
	√	*** 14**	*Class construction is then combinatorially very demanding and depends on frequency counts for joint events involving particular words, a potentially unreliable source of information as we noted above.*
	(OWN)	**21**	*We have not yet compared the accuracy and coverage of the two methods, or what systematic biases they might introduce, although we took care to filter out certain systematic errors, for instance the misparsing of the subject of a complement clause as the direct object of a main verb for report verbs like "say".*
	√	**43**	*This is a useful advantage of our method compared with agglomerative clustering techniques that need to compare individual objects being considered for grouping.*

Figure 23
System output for example paper

terms, but it also contains a contrastive statement, which the annotator decided to rate higher than the goal statement. Both sentences **12** and **150** give high-level descriptions of the work which might pass as a goal statement. Similarly, in sentence **21** the agent and action features detected that the first part of the sentence has something to do with comparing methods, and the system then (plausibly but incorrectly) decided to classify the sentence as CONTRAST. All in all, we feel that the extracted material conveys the rhetorical status adequately; an extrinsic evaluation additionally shows that the end result provides considerable added value when compared to sentence extracts.

5.2 Relevance Determination

The classifier for rhetorical status which we evaluated in the previous section is an important first step in our implementation; the next step is the determination of relevant sentences in the text. One simple solution for relevance decision would be to use *all* AIM, BASIS and CONTRAST sentences, as these categories are rare overall. The classifier we use has the nice property that it roughly keeps the distribution of target categories, so that we end up with a sensible number of these sentences.

The strategy of using all AIM, CONTRAST and BASIS sentences can be evaluated in a similar vein to the previous experiment. In terms of relevance, the asterisk in figure 23 marks sentences which the human judge found particularly relevant in the overall context (cf. the full set in figure 6). 6 out of all 15 sentences, and 6 out of the 10 sentences which received the correct rhetorical status, were judged relevant in the example.

	AIM		CONTR.		BASIS		BACKGROUND without classifier		BACKGROUND with classifier	
	p	r	p	r	p	r	p	r	p	r
System	96.2	69.8	70.1	23.8	70.5	39.4	16.0	83.3	38.4	88.2
Baseline	26.1	6.4	23.5	14.4	6.94	2.7	0.0	0.0	0.0	0.0

Figure 24
Relevance by human selection: precision and recall

Figure 24 reports the figure for the entire corpus by comparing the system's output of correctly classified rhetorical categories to human judgement. In all cases, the results are far above the non-trivial baseline. On AIM, CONTRAST and BASIS sentences, we achieve very high precision values of 96%, 70%, and 71%. Recall is lower at 70%, 24% and 39%, but low recall is less of a problem in our final task. Therefore, the main bottleneck is correct rhetorical classification. Once that is accomplished, the selected categories show high agreement with human judgement and should therefore represent good material for further processing steps.

However, if one is also interested in selecting BACKGROUND sentences, as we are, simply choosing all BACKGROUND sentences would result in low precision of 16% (albeit with a high recall of 83%), which does not seem to be the optimal solution. We therefore use a second classifier for finding the most relevant sentences independently, which was trained on the relevance gold standard. Our best classifier operates at a precision of 46.5% and recall of 45.2% (using the features `Location, Section Struct, Paragraph Struct, Title, TF*IDF, Formulaic` and `Citation` for classification). The second classifier (cf. rightmost columns in figure 24) raises precision for BACKGROUND sentences from 16% to 38%, while keeping recall high at 88%. This example shows that the right procedure concerning relevance determination changes from category to category, and also depends on the final task one is trying to accomplish.

6 Discussion

6.1 Contribution

We have presented a new method for content selection from scientific articles. The analysis is genre-specific; it is based on rhetorical phenomena specific to academic writing, such as problem-solution structure, explicit intellectual attribution and statements of relatedness to other work. The goal of the analysis is to identify the contribution of an article in relation to background material and to other specific current work.

Our methodology is situated between text extraction methods and fact extraction (template filling) methods: while our analysis has the advantage of being more context-sensitive than text extraction methods, it retains the robustness of this approach towards different subdomains, presentational traditions and writing styles.

Like fact-extraction methods (e.g., Radev and McKeown (1998)), our method also uses a "template" whose slots are being filled during analysis. The slots of our template are defined as rhetorical categories (like "Contrast") rather than by domain-specific categories (like "Perpetrator"). This makes it possible for our approach to deal with texts of different domains and unexpected topics.

Spärck Jones (1999) argues that it is crucial for a summarisation strategy to relate the large scale document structure of texts to reader's tasks in the real world, i.e., to the proposed use of the summaries. We feel that incorporating a robust analysis of discourse structure into a document summariser is one step along this way.

Our practical contributions are twofold. First, we present a scheme for the annotation of sentences with rhetorical status, and we have shown that the annotation is stable (K=.82, .81, .76) and reproducible (K=.71). Since these results indicate that the annotation is reliable, we use it as our gold standard for evaluation and training.

Second, we present a machine learning system for the classification of sentences by relevance and by rhetorical status. The contribution here is not the statistical classifier, which is well-known and has been used in a similar task by Kupiec et al. (1995), but instead the features we use. We have adapted 13 sentential features in such a way that they work robustly for our task, i.e. for unrestricted, real-world text. We also present three new features which detect scientific meta-discourse in a novel way. The results of an intrinsic system evaluation show that the system can identify sentences expressing the specific goal of a paper with 57% precision and 79% recall, sentences expressing criticism or contrast with 57% precision and 42% recall, and sentences expressing a continuation relationship to other work with 62% precision and 43% recall. This substantially improves a baseline of text classification which uses only a *TF*IDF* model over words. The agreement of correctly identified rhetorical roles with human relevance judgements is even higher (96% precision and 70% recall for goal statements, 70% precision and 24% recall for contrast, 71% precision and 39% recall for continuation). We see these results as an indication that shallow discourse processing with a well-designed set of surface-based indicators is possible.

6.2 Limitations and Future Work

The meta-discourse features, one focus of our work, currently depend on manual resources. The experiments reported here explore whether meta-discourse information is useful for the automatic determination of rhetorical status (as opposed to more shallow features), and this is clearly the case. However, the next step should be the automatic creation of such resources. For the task of dialogue act disambiguation, Samuel et al. (1999) suggest a method of automatically finding cue phrases for disambiguation. It may be possible to apply this or a similar method to our data and to compare the performance of automatically gained resources with manual ones.

Further work can be done on the semantic verb clusters described in section 4.2. Klavans and Kan (1998), who use verb clusters for document classification according to genre, observe that verb information is rarely used in current practical natural language applications. Most tasks such as information extraction and document classification identify and use nominal constructs instead (e.g., noun phrases, *TF*IDF* words and phrases).

The verb clusters we use were created using our intuition of which type of verb similarity would be useful in the genre and for the task. There are good reasons for using such a hand-crafted, genre-specific verb lexicon instead of a general resource such as WordNet or Levin's (1993) classes: many verbs used in the domain of scientific argumentation have assumed a specialised meaning, which our lexicon readily encodes. Klavans and Kan's classes, which are based on Levin's classes, are also manually created. Resnik and Diab (2000) present yet other measures of verb similarity, which could be used to arrive at a more data-driven definition of verb classes. We are currently comparing our verb clusterings to Klavans and Kan's, and to bottom–up clusters of verb similarities generated from our annotated data.

The recognition of agents, which is already the second best feature in the pool, could be further improved by including named entity recognition and anaphora resolution. Named entity recognition would help in cases like the following,

> *LHIP provides a processing method which allows selected portions of the input to be ignored or handled differently.* (S-5, 9408006)

where "*LHIP*" is the name of the authors' approach and should thus be tagged as US_AGENT; however, to do so, one would need to recognise it as a named approach, which is associated with the authors. It is very likely that such a treatment, which would have to include information from elsewhere in the text, would improve results, particularly as named approaches are frequent in the computational linguistics domain. Information about named approaches in themselves would also be an important aspect to include in summaries or citation indexes.

Anaphora resolution helps in cases where the agent is syntactically ambiguous between own and other approaches (e.g., *"this system"*). To test if and how much performance would improve, we manually simulated anaphora resolution on the 632 occurrences of REF_AGENT in the development corpus. (In the experiments in section 5 these occurrences had been excluded from the `agent` feature by giving them the value *None*; we include them now in their disambiguated state). 436 (69%) of the 632 REF_AGENTs were classified as US_AGENT, 175 (28%) as THEM_AGENT, and 20 (3%) as GENERAL_AGENT. As a result of this manual disambiguation, the performance of the `Agent` feature increased dramatically from K=.08 to K=.14, for `SegAgent` from K=.19 to K=.22. This is a clear indication of the potential added value of anaphora resolution for our task.

As far as the statistical classification is concerned, our results are still far from perfect. Obvious ways of improving performance are the use of a more complicated statistical classifier and of providing more training material. We have experimented with a Maximum Entropy model, RIPPER and decision trees; preliminary results do not show significant improvement over the naïve Bayesian model. One problem is that 4% of the sentences in our current annotated material are ambiguous: they receive the same feature representation, but are classified differently by the annotator. A possible solution is to find better and more distinctive features; we believe that robust, higher-level features like actions and agents are a step in the right direction. We also suspect that a big improvement could be achieved with smaller annotation units. Many errors come from the fact that one half of a sentence serves one rhetorical purpose, the other another, as in the following example:

> *The current paper shows how to implement this general notion, without following Krifka's analysis in detail.* (S-10, 9411019)

Here, the first part describes the paper's research goal, whereas the second expresses a contrast. Currently, *one* target category needs to be associated with the whole sentence (according to a rule in the guidelines, AIM is given preference over CONTRAST). As an undesired side effect the CONTRAST-like textual parts (and the features associated with this text piece, e.g., the presence of an author's name) are wrongly associated with the AIM target category. If we allowed for a smaller annotation unit, e.g., at the clause level, this systematic noise in the training data could be removed.

Another improvement in classification accuracy might be achieved by performing the classification in a cascading way. The system could first perform a classification into OWN-like classes (OWN, AIM, and TEXTUAL pooled), OTHER-like categories (OTHER, CONTRAST and BASIS pooled), and BACKGROUND, similar to the way human annotation proceeds. Subclassification amongst these classes would then lead to the final 7-way classification.

Acknowledgments
The authors would like to thank Jean Carletta for her help with the experimental design, Chris Brew for many helpful discussions, Claire Grover and Andrei Mikheev for advice on the XML implementation, and the annotators, Vasilis Karaiskos and Anne Wilson, for their meticulous work and criticism, which lead to several improvements in the annotation scheme. Thanks also to Byron Georgantopolous, who helped to collect the first version of the corpus, and to the four anonymous reviewers.

Appendix: List of articles in CL development corpus

	No.	Title, Conference, Authors
0	9405001	Similarity-Based Estimation of Word Cooccurrence Probabilities (ACL94), I.Dagan et al.
1	9405002	Temporal Relations: Reference or Discourse Coherence? (ACL94 Student), A.Kehler
2	9405004	Syntactic-Head-Driven Generation (COLING94), E.Koenig
3	9405010	Common Topics and Coherent Situations: Interpreting Ellipsis in the Context of Discourse Inference (ACL94), A.Kehler
4	9405013	Collaboration on Reference to Objects that are not Mutually Known (COLING94), P.Edmonds
5	9405022	Grammar Specialization through Entropy Thresholds (ACL94), C.Samuelsson
6	9405023	An Integrated Heuristic Scheme for Partial Parse Evaluation (ACL94 Student), A.Lavie
7	9405028	Semantics of Complex Sentences in Japanese (COLING94), H.Nakagawa, S.Nishizawa
8	9405033	Relating Complexity to Practical Performance in Parsing with Wide-Coverage Unification Grammars (ACL94), J.Carroll
9	9405035	Dual-Coding Theory and Connectionist Lexical Selection (ACL94 Student), Y.Wang
10	9407011	Discourse Obligations in Dialogue Processing (ACL94), D.Traum, J.Allen
11	9408003	Typed Feature Structures as Descriptions (COLING94 Reserve), P.King
12	9408004	Parsing with Principles and Probabilities (ACL94 Workshop), A.Fordham, M.Crocker
13	9408006	LHIP: Extended DCGs for Configurable Robust Parsing (COLING94), A.Ballim, G.Russell
14	9408011	Distributional Clustering of English Words (ACL93), F.Pereira et al.
15	9408014	Qualitative and Quantitative Models of Speech Translation (ACL94 Workshop), H.Alshawi
16	9409004	An Experiment on Learning Appropriate Selectional Restrictions from a Parsed Corpus (COLING94), F.Ribas
17	9410001	Improving Language Models by Clustering Training Sentences (ANLP94), D.Carter
18	9410005	A Centering Approach to Pronouns (ACL87), S.Brennan et al.
19	9410006	Evaluating Discourse Processing Algorithms (ACL89), M.Walker
20	9410008	Recognizing Text Genres with Simple Metrics Using Discriminant Analysis (COLING94), J.Karlgren, D.Cutting
21	9410009	Reserve Lexical Functions and Machine Translation (COLING94), D.Heylen et al.
22	9410012	Does Baum-Welch Re-estimation Help Taggers? (ANLP94), D.Elworthy
23	9410022	Automated Tone Transcription (ACL94 SIG), S.Bird
24	9410032	Planning Argumentative Texts (COLING94), X.Huang
25	9410033	Default Handling in Incremental Generation (COLING94), K.Harbusch et al.
26	9411019	Focus on “only” and “not” (COLING94), A.Ramsay
27	9411021	Free-ordered CUG on Chemical Abstract Machine (COLING94), S.Tojo
28	9411023	Abstract Generation Based on Rhetorical Structure Extraction (COLING94), K.Ono et al.
29	9412005	Segmenting Speech without a Lexicon: the Roles of Phonotactics and Speech Source (ACL94 SIG), T.Cartwright, M.Brent
30	9412008	Analysis of Japanese Compound Nouns using Collocational Information (COLING94), Y.Kobayasi et al.
31	9502004	Bottom-Up Earley Deduction (COLING94), G.Erbach
32	9502005	Off-line Optimization for Earley-style HPSG Processing (EACL95), G.Minnen et al.
33	9502006	Rapid Development of Morphological Descriptions for Full Language Processing Systems (EACL95), D.Carter
34	9502009	On Learning More Appropriate Selectional Restrictions (EACL95), F.Ribas
35	9502014	Ellipsis and Quantification: A Substitutional Approach (EACL95), R.Crouch
36	9502015	The Semantics of Resource Sharing in Lexical-Functional Grammar (EACL95), A.Kehler et al.
37	9502018	Algorithms for Analysing the Temporal Structure of Discourse (EACL95), J.Hitzeman et al.
38	9502021	A Tractable Extension of Linear Indexed Grammars (EACL95), B.Keller, D.Weir
39	9502022	Stochastic HPSG (EACL95), C.Brew
40	9502023	Splitting the Reference Time: Temporal Anaphora and Quantification in DRT (EACL95), R.Nelken, N.Francez
41	9502024	A Robust Parser Based on Syntactic Information (EACL95), K.Lee et al.
42	9502031	Cooperative Error Handling and Shallow Processing (EACL95 Student), T.Bowden
43	9502033	An Algorithm to Co-Ordinate Anaphora Resolution and PPS Disambiguation Process (EACL95 Student), S.Azzam
44	9502035	Incorporating “Unconscious Reanalysis” into an Incremental, Monotonic Parser (EACL95 Student), P.Sturt
45	9502037	A State-Transition Grammar for Data-Oriented Parsing (EACL95 Student), D.Tugwell

	No.	Title, Conference, Authors
46	9502038	Implementation and evaluation of a German HMM for POS disambiguation (EACL95 Workshop), H.Feldweg
47	9502039	Multilingual Sentence Categorization according to Language (EACL95 Workshop), E.Giguet
48	9503002	Computational Dialectology in Irish Gaelic (EACL95), B.Kessler
49	9503004	Creating a Tagset, Lexicon and Guesser for a French tagger (EACL95 Workshop), J.Chanod, P.Tapanainen
50	9503005	A Specification Language for Lexical Functional Grammars (EACL95), P.Blackburn, C.Gardent
51	9503007	The Semantics of Motion (EACL95), P.Sablayrolles
52	9503009	Distributional Part-of-Speech Tagging (EACL95), H.Schuetze
53	9503013	Incremental Interpretation: Applications, Theory, and Relationship to Dynamic Semantics (COLING95), D.Milward, R.Cooper
54	9503014	Non-Constituent Coordination: Theory and Practice (COLING94), D.Milward
55	9503015	Incremental Interpretation of Categorial Grammar (EACL95), D.Milward
56	9503017	Redundancy in Collaborative Dialogue (COLING92), M.Walker
57	9503018	Discourse and Deliberation: Testing a Collaborative Strategy (COLING94), M.Walker
58	9503023	A Fast Partial Parse of Natural Language Sentences Using a Connectionist Method (EACL95), C.Lyon, B.Dickerson
59	9503025	Occurrence Vectors from Corpora vs. Distance Vectors from Dictionaries (COLING94), Y.Niwa, Y.Nitta
60	9504002	Tagset Design and Inflected Languages (EACL95 Workshop), D.Elworthy
61	9504006	Cues and Control in Expert-Client Dialogues (ACL88), S.Whittaker, P.Stenton
62	9504007	Mixed Initiative in Dialogue: An Investigation into Discourse Segmentation (ACL90), M.Walker, S.Whittaker
63	9504017	A Uniform Treatment of Pragmatic Inferences in Simple and Complex Utterances and Sequences of Utterances (ACL95), D.Marcu, G.Hirst
64	9504024	A Morphographemic Model for Error Correction in Nonconcatenative Strings (ACL95), T.Bowden, G.Kiraz
65	9504026	The Intersection of Finite State Automata and Definite Clause Grammars (ACL95), G.van Noord
66	9504027	An Efficient Generation Algorithm for Lexicalist MT (ACL95),V.Poznanski et al.
67	9504030	Statistical Decision-Tree Models for Parsing (ACL95), D.Magerman
68	9504033	Corpus Statistics Meet the Noun Compound: Some Empirical Results (ACL95), M.Lauer
79	9504034	Bayesian Grammar Induction for Language Modeling (ACL95), S.Chen
70	9505001	Response Generation in Collaborative Negotiation (ACL95), J.Chu-Carroll, S.Carberry
71	9506004	Using Higher-Order Logic Programming for Semantic Interpretation of Coordinate Constructs (ACL95), S.Kulick
72	9511001	Countability and Number in Japanese-to-English Machine Translation (COLING94), F.Bond et al.
73	9511006	Disambiguating Noun Groupings with Respect to WordNet Senses (ACL95 Workshop), P.Resnik
74	9601004	Similarity between Words Computed by Spreading Activation on an English Dictionary (EACL93), H.Kozima, T.Furugori
75	9604019	Magic for Filter Optimization in Dynamic Bottom-up Processing (ACL96), G.Minnen
76	9604022	Unsupervised Learning of Word-Category Guessing Rules (ACL96), A.Mikheev
77	9605013	Learning Dependencies between Case Frame Slots (COLING96), H.Li, N.Abe
78	9605014	Clustering Words with the MDL Principle (COLING96), H.Li, N.Abe
79	9605016	Parsing for Semidirectional Lambek Grammar is NP-Complete (ACL96), J.Doerre

References

Barzilay, Regina, Michael Collins, Julia Hirschberg, and Steve Whittaker. 2000. The rules behind roles. In *Proceedings of AAAI-00*.

Barzilay, Regina, Kathleen R. McKeown, and Michael Elhadad. 1999. Information fusion in the context of multi-document summarization. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*, pages 550–557.

Baxendale, Phyllis B. 1958. Man-made index for technical literature—an experiment. *IBM Journal of Research and Development*, 2(4):354–361.

Biber, Douglas. 1995. *Dimensions of Register Variation: A Cross-linguistic Comparison*. Cambridge University Press, Cambridge, England.

Brandow, Ronald, Karl Mitze, and Lisa F. Rau. 1995. Automatic condensation of electronic publications by sentence selection. *Information Processing and Management*, 31(5):675–685.

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

Choi, Freddy Y. Y. 2000. Advances in domain independent linear text segmentation. In *Proceedings of the 6th Applied Natural Language Conference(ANLP-00) and the 1st Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-00)*, pages 26–33.

CMP_LG. 1994. The Computation and Language E-Print Archive, `http://xxx.lanl.gov/cmp-lg`.

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

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

Edmundson, H. P. et al. 1961. *Final Report on the Study for Automatic Abstracting*. Thompson Ramo Wooldridge, Canoga Park, CA.

Garfield, Eugene. 1979. *Citation Indexing: Its Theory and Application in Science, Technology and Humanities*. J. Wiley, New York, NY.

Grefenstette, Gregory. 1998. Producing intelligent telegraphic text reduction to provide an audio scanning service for the blind. In Radev and Hovy (Radev and Hovy, 1998), pages 111–117.

Grover, Claire, Andrei Mikheev, and Colin Matheson. 1999. LT TTT version 1.0: Text tokenisation software. Technical report, Human Communication Research Centre, University of Edinburgh. `http://www.ltg.ed.ac.uk/software/ttt/`.

Hearst, Marti A. 1997. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

Hyland, Ken. 1998. Persuasion and context: The pragmatics of academic metadiscourse. *Journal of Pragmatics*, 30(4):437–455.

Jing, Hongyan, Regina Barzilay, Kathleen R. McKeown, and Michael Elhadad. 1998. Summarization evaluation methods: Experiments and analysis. In Radev and Hovy (Radev and Hovy, 1998), pages 60–68.

Jing, Hongyan and Kathleen R. McKeown. 2000. Cut and paste based summarization. In *Proceedings of the 6th Applied Natural Language Conference(ANLP-00) and the 1st Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-00)*, pages 178–185.

Jordan, Michael P. 1984. *Rhetoric of Everyday English Texts*. George Allen and Unwin, London, UK.

Klavans, Judith L. and Min-Yen Kan. 1998. Role of verbs in document analysis. In *Proceedings of 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (ACL/COLING-98)*, pages 680–686.

Knight, Kevin and Daniel Marcu. 2000. Statistics-based summarization — step one: Sentence compression. In *Proceeding of The 17th National Conference of the American Association for Artificial Intelligence (AAAI-2000)*, pages 703–710.

Krippendorff, Klaus. 1980. *Content Analysis: An Introduction to its Methodology*. Sage Publications, Beverly Hills, CA.

Kupiec, Julian, Jan O. Pedersen, and Francine Chen. 1995. A trainable document summarizer. In *Proceedings of the 18th Annual International Conference on Research and Development in Information Retrieval (SIGIR-95)*, pages 68–73.

Lancaster, Frederick Wilfrid. 1998. *Indexing and Abstracting in Theory and Practice*. Library Association, London, UK.

Landis, J. Richard and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.

Lawrence, Steve, C. Lee Giles, and Kurt Bollacker. 1999. Digital libraries and autonomous citation indexing. *IEEE Computer*, 32(6):67–71.

Levin, Beth. 1993. *English Verb Classes and Alternations*. University of Chicago Press, Chicago, IL.

Lewis, David D. 1991. Evaluating text categorisation. In *Speech and Natural Language: Proceedings of the ARPA Workshop of Human Language Technology*.

Liddy, Elizabeth DuRoss. 1991. The discourse-level structure of empirical abstracts: An exploratory study. *Information Processing and Management*, 27(1):55–81.

Lin, Chin-Yew and Eduard H. Hovy. 1997. Identifying topics by position. In *Proceedings of the 5th Applied Natural Language Conference (ANLP-97)*, pages 283–290.

Luhn, Hans Peter. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165.

Mani, Inderjeet, Therese Firmin, David House, Gary Klein, Beth Sundheim, and Lynette Hirschman. 1999. The TIPSTER Summac Text Summarization Evaluation. In *Proceedings of the 9th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-99)*, pages 77–85.

Mani, Inderjeet, Barbara Gates, and Eric Bloedorn. 1999. Improving summaries by revising them. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*, pages 558–565.

Mani, Inderjeet and Mark T. Maybury, editors. 1999. *Advances in Automatic Text Summarization*. MIT Press, Cambridge, MA.

Mann, William C. and Sandra A. Thompson. 1987. Rhetorical Structure Theory: Description and construction of text structures. In Gerard Kempen, editor, *Natural Language Generation: New Results in Artificial Intelligence, Psychology, and Linguistics*. Marinus Nijhoff Publishers, Dordrecht, NL, pages 85–95.

Marcu, Daniel. 1999. Discourse trees are good indicators of importance in text. In Mani and Maybury (Mani and Maybury, 1999), pages 123–136.

McCallum, Andrew. 1997. Training algorithms for linear text classifiers. In *Proceedings of the 19th Annual International Conference on Research and Development in Information Retrieval (SIGIR-97)*.

Mizzaro, Stefano. 1997. Relevance: The whole history. *Journal of the American Society for Information Science*, 48(9):810–832.

Morris, Jane and Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17:21–48.

Myers, Greg. 1992. In this paper we report...—speech acts and scientific facts. *Journal of Pragmatics*, 17(4):295–313.

Nanba, Hidetsugu and Manabu Okumura. 1999. Towards multi-paper summarization using reference information. In *Proceedings of IJCAI-99*, pages 926–931.

Paice, Chris D. 1990. Constructing literature abstracts by computer: Techniques and prospects. *Information Processing and Management*, 26:171–186.

Paice, Chris D. and A. Paul Jones. 1993. The identification of important concepts in highly structured technical papers. In *Proceedings of the 16th Annual International Conference on Research and Development in Information Retrieval (SIGIR-93)*, pages 69–78.

Pollock, Joseph J. and Antonio Zamora. 1975. Automatic abstracting research at the chemical abstracts service. *Journal of Chemical Information and Computer Sciences*, 15(4):226–232.

Radev, Dragomir R. and Eduard H. Hovy, editors. 1998. *Working Notes of the AAAI Spring Symposium on Intelligent Text Summarization*.

Radev, Dragomir R. and Kathleen R. McKeown. 1998. Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, 24(3):469–500.

Rath, G.J, A. Resnick, and T. R. Savage. 1961. The formation of abstracts by the selection of sentences. *American Documentation*, 12(2):139–143.

Resnik, Philip and Mona Diab. 2000. Measuring verb similarity. In *Twenty Second Annual Meeting of the Cognitive Science Society (COGSCI2000)*.

Riley, Kathryn. 1991. Passive voice and rhetorical role in scientific writing. *Journal of Technical Writing and Communication*, 21(3):239–257.

Rowley, Jennifer. 1982. *Abstracting and Indexing*. Bingley, London, UK.

Saggion, Horacio and Guy Lapalme. 2000. Selective analysis for automatic abstracting: Evaluating ndicativeness and acceptability. In *Proceedings of*

Content-Based Multimedia Information Access (RIAO), pages 747–764.

Salton, Gerard and Michael J. McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill, Tokyo.

Samuel, Ken, Sandra Carberry, and K. Vijay-Shanker. 1999. Automatically selecting useful phrases for dialogue act tagging. In *Proceedings of the Pacific Association for Computational Linguistics (PACLING-99)*.

Saracevic, Tefko. 1975. Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, 26(6):321–343.

Siegel, Sidney and N. John Jr. Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, Berkeley, CA, 2nd edition.

Spärck Jones, Karen. 1990. What sort of thing is an AI experiment? In D. Partridge and Yorick Wilks, editors, *The Foundations of Artificial Intelligence: A Sourcebook*. Cambridge University Press, Cambridge, UK.

Spärck Jones, Karen. 1999. Automatic summarising: Factors and directions. In Mani and Maybury (Mani and Maybury, 1999), pages 1–12.

Swales, John, 1990. *Genre Analysis: English in Academic and Research Settings. Chapter 7: Research articles in English*, pages 110–176. Cambridge University Press, Cambridge, UK.

Teufel, Simone. 1999. *Argumentative Zoning: Information Extraction from Scientific Text*. Ph.D. thesis, School of Cognitive Science, University of Edinburgh, Edinburgh, UK.

Teufel, Simone. 2001. Task-based evaluation of summary quality: Describing relationships between scientific papers. In *Proceedings of NAACL-01 Workshop "Automatic Text Summarization"*.

Teufel, Simone. In Preparation. Extrinsic evaluation of different document surrogates.

Teufel, Simone, Jean Carletta, and Marc Moens. 1999. An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of the 8th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-99)*, pages 110–117.

Teufel, Simone and Marc Moens. 1997. Sentence extraction as a classification task. In *Proceedings of the ACL/EACL-97 Workshop on Intelligent Scalable Text Summarization*, pages 58–65.

Teufel, Simone and Marc Moens. 2000. What's yours and what's mine: Determining intellectual attribution in scientific text. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

Thompson, Geoff and Ye Yiyun. 1991. Evaluation in the reporting verbs used in academic papers. *Applied Linguistics*, 12(4):365–382.

Tombros, Anastasios, Mark Sanderson, and Phil Gray. 1998. Advantages of query biased summaries in information retrieval. In Radev and Hovy (Radev and Hovy, 1998).

Trawinski, Bogdan. 1989. A methodology for writing problem-structured abstracts. *Information Processing and Management*, 25(6):693–702.

van Dijk, Teun A. 1980. *Macrostructures: An Interdisciplinary Study of Global Structures in Discourse, Interaction and Cognition*. Lawrence Erlbaum, Hillsdale, NJ.

van Rijsbergen, Cornelis Joost. 1979. *Information Retrieval*. Butterworth, London, UK, 2nd edition.

Wiebe, Janyce. 1994. Tracking point of view in narrative. *Computational Linguistics*, 20(2):223–287.

Yang, Yiming and Xin Liu. 1999. A re-examination of text categorization methods. In *Proceedings of 22nd Annual International Conference on Research and Development in Information Retrieval (SIGIR-99)*, pages 42–49.

Zappen, James P. 1983. A rhetoric for research in sciences and technologies. In Paul V. Anderson, R. John Brockman, and Carolyn R. Miller, editors, *New Essays in Technical and Scientific Communication Research Theory Practice*. Baywood Publishing Company, Inc., Farmingdale, NY, pages 123–138.

Zechner, Klaus. 1995. Automatic text abstracting by selecting relevant passages. Master's thesis, Centre for Cognitive Science, University of Edinburgh, Edinburgh, UK.

Ziman, John M. 1969. Information, communication, knowledge. *Nature*, 224:318–324.