

# Summarizing Scientific Articles — Experiments with Relevance and Rhetorical Status

Simone Teufel\*  
Columbia University

Marc Moens†  
University of Edinburgh

*In this paper we argue that scientific articles require a different summarization strategy than, for instance, news articles. We propose a strategy 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 with human judgements of rhetorical status and relevance. 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 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 can be used to generate task-oriented and user-tailored summaries designed to give users an overview of a scientific field.*

## 1 Introduction

Summarization 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 for users 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. 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

---

\* The work reported here was done at the University of Edinburgh, HCRC Language Technology Group.  
Author now at Columbia University, Computer Science Department, 1214 Amsterdam Avenue, New York, NY10025

† HCRC Language Technology Group, 2 Buccleuch Place, Edinburgh EH8 9LW, UK

source text. We show how this added contextual information can be used to make the end product more informative and more valuable than sentence extracts.

Most current summarization research is focused on news articles; other genres have received little attention in the summarization community. But summarization strategies are genre-dependent, and methods developed for other genres do not necessarily work well for scientific articles.

Our application domain is the summarization of scientific articles. Summarization of such texts requires a different approach from, e.g., the approach used in the summarization of news articles. For example, in their work on the summarization of news stories Barzilay et al. (1999) introduce the concept of *information fusion* which is based on the identification of recurrent descriptions of the same events. This is efficient because in the news domain news-worthy events are frequently repeated over a short period of time. However, in scientific writing, similar “events” are rare: new ideas are the main focus of articles, and these ideas are distinguished by their *uniqueness* and *difference* from other ideas.

Other approaches to the summarization of news articles make use of the typical journalistic writing style, for example the fact that the most news-worthy 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 own contribution of the paper is often given much later.

The length of scientific articles presents another problem. Let us assume that our overall summarization strategy is to first select relevant sentences or concepts, and to then synthesize summaries using this material. For a typical 10–20 sentence newswire story, a compression to 20 or 30% provides a reasonable input set for the second step. The Achilles heel of sentence extraction, its context insensitivity, does not hurt much in this case: extracted sentences are still “connected” enough with respect to cohesion and discourse structure, 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 context insensitivity problem 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 summarizing 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 the rhetorical status of each sentence, as exemplified by Figure 25. Such lists serve two purposes: in themselves, they already provide a better characterization of scientific articles than sentence extracts do, and in the longer run, they will serve as better input material for further processing.

The added rhetorical context allows for a new kind of summaries. While the actual construction of these summaries is outside the scope of this work, section 3 motivates what they could look like. They are summaries which can be tailored to users’ expertise and task; several articles can be summarized together, and their contrasts or complementarity can be brought out in the summaries; finally, the summaries can be displayed together with citations in a graphical form to help users navigate several related papers.

The aspects of document structure we model in this work are genre-specific. They

are discussed in detail in section 2.1, and can be summarized as follows:

- *Rhetorical status in terms of problems solving*: What is the goal and contribution of the paper? This type of information can be recognized by conventional patterns of presentation (section 2) and is often marked by *meta-discourse* (section 5.3).
- *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 recognized by following the “agent structure” of text, i.e., by looking at all grammatical subjects occurring in sequence (section 5.4).
- *Relatedness between articles*: What articles is this work similar to, and in what respect? This type of information can be found by examining citations (section 2.2), section headers, and fixed indicator phrases like “*in contrast to ...*” (section 5.3).

These features of rhetorical structure were arrived at partly through a study of the nature of scientific articles. But we also measured the degree to which human annotators agree on these features before adding them to our framework (section 4.2). In section 6 we present an overview of these intrinsic evaluation procedures.

## 2 Rhetorical Status, Citations and Relevance

It is important for our task to find the right definition of “rhetorical status” for scientific articles. The definition should both capture some true generalizations of the structure of the texts, and also provide the right kind of information for better summaries for final task. Another requirement is that the analysis should be applicable to research articles from different presentational traditions and subject matters. While our analysis is designed to be independent of subject matter, it is restricted to the genre of research articles. This allows us to use knowledge about the rhetorical devices typical in the genre.

For the development of our scheme, we used 80 conference articles in the field of computational linguistics articles (articles presented at COLING, ANLP and (E)ACL conferences or workshops). Due to the interdisciplinarity of the field, the papers in our collection cover a challenging range of subject matters, such as logic programming, statistical language modelling, theoretical semantics, computational dialectology and computational psycholinguistics. Because the research methodology and tradition of presentation is very different in these fields (i.e., because computer scientists write very different papers than theoretical linguists), we would expect our analysis to be equally applicable in a range of 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; 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 crystallized in the typical presentation of the scientific material, e.g., the well-known, fixed Introduction/Method/Result/Discussion structure (van Dijk, 1980). But many texts in

computational linguistics do not adhere to the traditional writing style, so the definitions of our analysis are primarily based on the underlying logical (rhetorical) organization, and use 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). Authors have good reason to write their papers in such a way that the attribution of statements is clear; they are forced by the reward system in science to stake their own scientific claim. Our rhetorical scheme assumes that readers have no difficulty in understanding this distinction, an assumption which we verified experimentally (section 4).

*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 argument into single, non-hierarchical argumentative moves (i.e., rhetorically coherent pieces of text, which convey the same communicative function). His CARS model 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 piece with respect to the overall message of the document, and not with respect to adjacent text pieces.

*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, there is a reason why each piece of other work was mentioned: 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.

## 2.2 Citations and Relatedness

Traditionally, there are two ways of accessing scientific information — keyword based searches, and citation links; traditionally, these two ways are considered as diametrically opposed (Garfield, 1979). If one is primarily interested in relations between scientific work, then citations seem the most natural source of information.

Citation indexes are constructs which contain pointers between *cited* texts and *citing* texts, traditionally in printed form. Recently, citation indexes can also be induced from arbitrary electronic articles on the fly ("autonomously"), cf. Lawrence, Giles and Bollacker's (1999) *CiteSeer*. In this tool, citations are presented in their context in running text, and users can browse the citation contexts. Browsing each citation is time-consuming, but necessary: just knowing *that* an article cites another is not enough. One needs to read the context of the citation to understand the relation between the articles.

The field of content citation analysis researches the differences in how and why an author cites (Moravcsik and Murugesan, 1975; Weinstock, 1971; MacRoberts and MacRoberts, 1984). Citations may vary in in many dimensions; e.g., they can be central or perfunctory, positive or negational (i.e., criticizing); 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 researchers:

- The articles which cite a given article negatively or contrastively; and
- The articles which build on the article or cite it positively.

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 1**

Annotation Scheme for Rhetorical Status

We suggest that these rhetorical distinctions can be made manually and automatically for each citation; we use a large corpus of scientific papers with humans' judgments of this distinction to train a system to do so. The use we plan to make of this distinction will be described in section 3. The general idea is that there are advantages in incorporating rhetorically enriched citation information systematically into summaries and summary-like entities, rather than treating summary and citation information as diametrically opposed ways of information access.

### 2.3 The Rhetorical Annotation Scheme

Our Rhetorical Annotation Scheme (Figure 1) encodes those aspects of scientific argumentation and relatedness to other work which were described in the last two subsections. The categories are assigned to full sentences. Ideally, we would have liked to assign the rhetorical labels entities shorter than sentences (e.g. clauses), but as the analysis will be done by machine as well as by humans, this was not possible. We think that it is a very hard problem to computationally separate clauses in a sensible way, whereas adequate sentence boundary detectors exist.

The category semantics are defined by rhetorical status of the sentence in the global context of the paper. For instance, while the OTHER category describes all neutral descriptions of specific 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). This last category could have been further subdivided into solution/method, results, and further work, for instance; in the work reported here, this is not done.

The annotation scheme is designed in such a way that it provides information for different applications (different types of information access). Summaries and better citation indexes are just one possible application. One might imagine the indexing and previewing of the internal structure of the article as another application; therefore, our scheme contains the additional category TEXTUAL (Swales' move 3.3), which captures previews of section structure ("*section 2 describes our data . . .*"). This information makes it possible to label sections with the author's indication of their contents.

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 only one sentence long, but zones can contain many more sentences.

## 2.4 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 the annotation just discussed). Relevance in general is a sticky issue because it is *situational* to a unique occasion (Saracevic, 1975; Spärck Jones, 1990; Mizzaro, 1997). This means that humans perceive relevance differently from each other, and dependent on the current situation. Paice and Jones (1993) remark that in an informal sentence selection experiment where they used agriculture articles and experts in the field as subjects, the subjects were unduly influenced by their personal research interest.

This is one of the reasons why human sentence extraction experiments typically report low agreement figures. Only few of these have been performed on scientific text. 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) report similarly low human agreement for research articles. More recent experiments reporting more positive results all used news text (Jing et al., 1998; Zechner, 1995). We think that part of the problem with scientific articles is the high compression and other specifics of the genre described in the introduction; part is the fact that large stretches of text in scientific articles can only be understood with considerable domain knowledge. All these problems import a high level of subjectivity into sentence selection experiments, particularly if the instructions are vague (“select relevant sentences”).

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 on it. However, it relies on two assumptions: a) there is a high degree of overlap between abstract and document sentences and b) the abstract is indeed the target output which is most adequate for the final task.

In our case, neither assumption a) nor assumption b) holds. First, the experiments in Teufel and Moens (1997) showed that in our corpus the overlap between abstract and document sentences is as low as 45%, whereas Kupiec et al. report an overlap of 79%. We believe that the reason for this is that the abstracts were produced by different populations: document authors in our case, and 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 internalized building plan (Liddy, 1991) and are often created by sentence extraction (Lancaster, 1998). Second, and more gravely, the abstracts we generate are not exactly modelled on traditional summaries, as we will explain in detail in the next section. Traditional summaries do not provide us with the type of information we need for our task. The *type* of information found in abstracts is normally restricted to few rhetorical categories, most prominently information about the goal of the paper and specifics of the solution. Information about related work is rarely found in abstracts, but in our strategy for summarization this information does play an important role (as will become clearer in section 3)—whether or not the author decided to include it in the abstract or not.

We thus decided to augment our corpus with a set of human judgements of relevance which is different from abstract similarity, namely human-selected sentences. 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 these instructions are given in section 4.

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 with a sentence in the introduction, a sentence in the conclusions and one in the middle of the document. But this type of redundancy 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 comparison with the 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 2 gives an example of the manual annotation. Relevant sentences of all rhetorical categories are shown. The example paper (F. Pereira, N. Tishby, L. Lee: *Distributional Clustering of English Words*, ACL-1993, cmp\_lg/9408011) was chosen from our collection as it is the most-cited paper within the collection. Our system creates a list like the one in Figure 2 automatically; indeed, Figure 25 shows the actual system output on the example paper. As an intermediate result, this type of output (extracted sentences with their rhetorical status) could be used directly as a better type of extract. Even though it contains some redundancy, it clearly provides more specific information than a mere sentence extract.

Such a list is at the same time the right kind of input material for a summary creation step. Based on this list, one could create user- and task-flexible single- and multiple document summaries, some of which could be graphically displayed in combination with citation information. The actual realization of these summaries is our long-term goal, a substantial research task which is outside the scope of this work. However, we describe the design of the summaries and the tasks they were created for, because the current work is heavily influenced by these ideas.

### 3 Summary Design and Summary Task

We describe in this section what types of summaries one could create from the lists given in Figures 2 and 25. When designing such summaries, it is important to keep in mind the task that the user wants to solve with the summaries. This section will show that the design of summaries is influenced by the type of task one assumes.

#### 3.1 Creating User-Tailored Summaries

One general advantage of automatic summaries is that they can be more flexible than human-written ones and more responsive to the users’ needs. Even though automatic summaries might be of a lower text quality when compared to human-crafted ones, we predict that they will support users in making a more informed decision on how well a paper fits their information needs. They don’t need to be self-contained: they can contain pointers (e.g., in the form of hyper-links) to certain passages in the full article, providing an interactive preview of the article’s contents. And they can show how articles are related and summarize their similarities and differences.

Figures 3 and 4 show the type of user-oriented and task-tailored abstracts we envisage. 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, describing differences of this approach to similar ones. Note the integral use of information about contrasted approaches in the second example.

The process of actually generating the surface realization of the abstracts is not a

<p><b>Aim:</b></p> <p>10 <i>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.</i></p> <p>22 <i>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.</i></p> <p>25 <i>The problem we study is how to use the EQN to classify the EQN.</i></p> <p>44 <i>In general, we are interested on how to organize a set of linguistic objects such as words according to the contexts in which they occur, for instance grammatical constructions or n-grams.</i></p> <p>46 <i>Our problem can be seen as that of learning a joint distribution of pairs from a large sample of pairs.</i></p> <p>162 <i>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.</i></p>
<p><b>Background:</b></p> <p>0 <i>Methods for automatically classifying words according to their contexts of use have both scientific and practical interest.</i></p> <p>4 <i>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.</i></p>
<p><b>Own (Details of Solution):</b></p> <p>66 <i>The first stage of an iteration is a maximum likelihood, or minimum distortion, estimation of the cluster centroids given fixed membership probabilities.</i></p> <p>140 <i>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.</i></p> <p>163 <i>The resulting clusters are intuitively informative, and can be used to construct class-based word cooccurrence [sic] models with substantial predictive power.</i></p>
<p><b>Contrast with Other Approaches/Weaknesses of Other Approaches:</b></p> <p>9 <i>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.</i></p> <p>14 <i>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.</i></p> <p>41 <i>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.</i></p>
<p><b>Basis (Imported Solutions):</b></p> <p>65 <i>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).</i></p> <p>113 <i>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.</i></p> <p>153 <i>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).</i></p>

**Figure 2**

Example of Manual Annotation: Relevant Sentences with Rhetorical Status

focus of this paper; however, we give readers an impression of how we envisage this step. To select the best source sentences for each version, one might use lexical heuristics to identify general as opposed to expert terminology (“*events, nouns, words*” in the non-expert summary; “*n-grams, association, joint events*” in the expert summary).

Non-underlined material in the examples was extracted verbatim, and underlined

**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 3**  
 Non-Expert Summary, General Purpose

**44** *This paper's goal is to organize 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 4**  
 Expert Summary, Contrastive Links

sentence parts were added in the process, requiring further processing. Surface operations like the ones described in Jing and McKeown (2000) could be used to string sentences together; alternatively, template-based output like in Paice and Jones (1993) could be created.

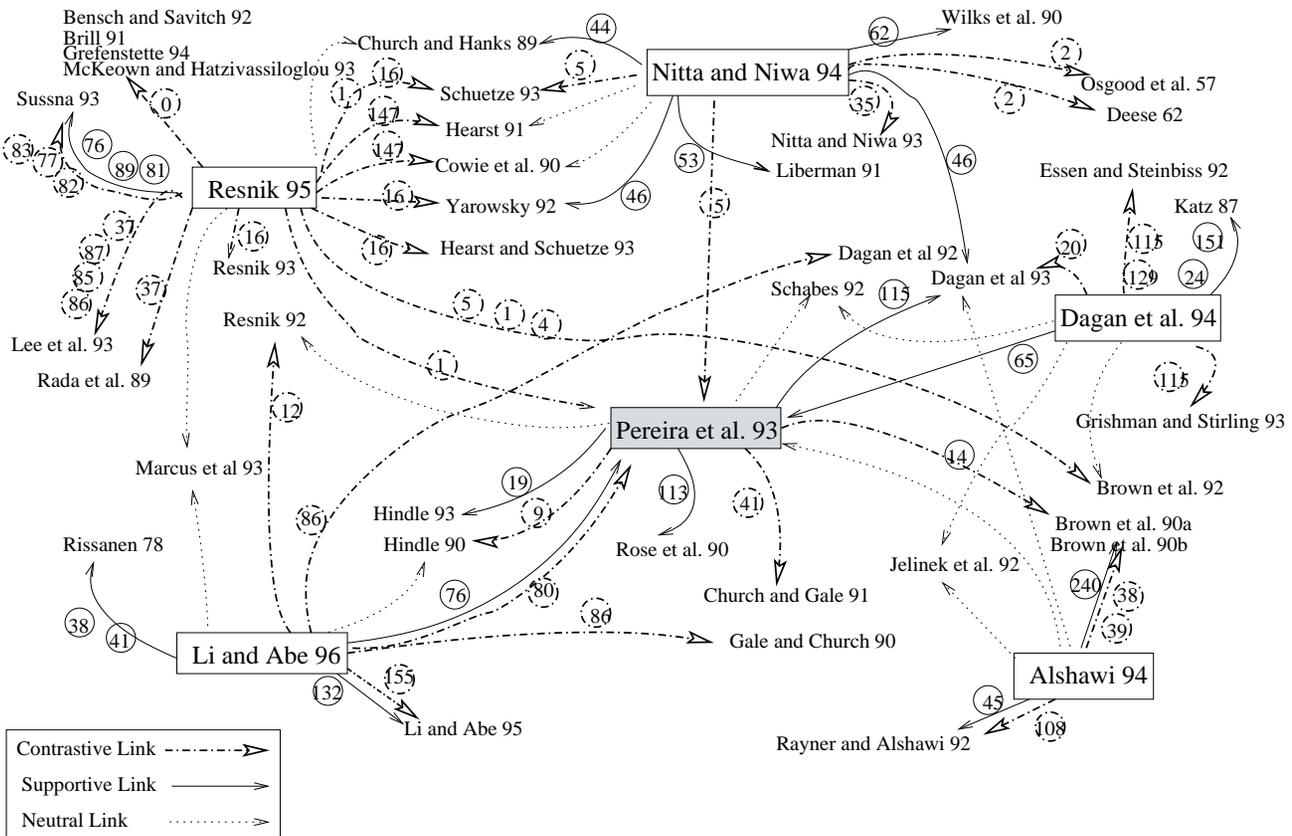
### 3.2 Incorporating Citation Information into Summaries

Apart from textual summaries, one can also envisage different summary-like entities, e.g., the graphical display of citation information (“citation map”) in Figure 5. In this figure, the example paper is shown in the center. The other citations in boxes are those papers in our full-text collection (a corpus of 80 papers) which directly cite the example paper. Other papers — papers outside our corpus — which are cited by the example paper or by the citing papers are displayed without boxes. As we have no access to the bibliographies of these papers, the citation chain ends here; the papers cited by these papers cannot be displayed.

The map was augmented with rhetorical information about the type of citation links. Citation links are shown in different line styles, depending on the rhetorical context of citations (supportive, contrastive or neutral).

Additional information could be displayed in citation maps upon request, e.g., the goal statements of each paper, and the sentence expressing the authors’ stance towards the cited paper (in Figure 5 the sentences are represented by their numbers next to the citation links). Seeing the actual sentence can be very important to the understanding of the relation (*In which respect is the paper criticized, or where does the contrast lie? In which respect is the solution incorporated?*)

Shum (1998) argues that as researchers are concerned with relations between facts or documents, not just with facts themselves, their information needs are particularly complex. He mentions, for example, the need to find criticisms of a certain approach in



**Figure 5**  
Citation Map for Example Paper (Pereira et al., 1993)

the field, to find differences between rival approaches, to find schools of thought, i.e., to find out which approaches have been evolved out of which other ones. Citation maps support such searches efficiently.

But they also cater for searchers who are new to a scientific field; Kircz (1991) calls them “partially informed readers”. Such users have particular problems with information access in an information retrieval environment (Fenichel, 1981). They cannot use keyword searches efficiently, as they have not yet acquired the scientific jargon in the field. A tool that creates citation maps from articles could help them bootstrap an overview of the field—in contrast to keyword searches, they could use the tool without any prior domain knowledge. Bazerman (1985), for instance, argues that experienced researchers in a field have organized their domain knowledge in a kind of linked representation centered around research goals, methodologies, researcher names, research groups and schools (which he calls *research maps*). It is this kind of meta-information that citation maps are trying to mimic.

We now turn to more traditional summaries. One idea for multi-document summarization of scientific articles is to use information in the *citing* articles to summarize the *cited* article. Figure 6 shows in which respect the papers in our development corpus contrast themselves to Pereira et al. The criticizing sentences are determined automatically by our method (category CONTRAST). If criticizing sentences were collected over a large corpus of articles, frequent criticisms (“trends”) can be detected by similarity and reported in one sentence (“*Later approaches find fault with the large corpus size required in this method.*”) In our approach, the same could be done for similar (supportive) articles, which are possibly further developments of this approach (cf. Figure 7). Nanba and Okumura’s (1999) system does something similar: it displays clusters of documents which are classified as having the same rhetorical connection to a given paper. However, the only rhetorical relation covered is criticism, and the semantics of these clusters is given by a list of keywords instead of a summary.

Contrasting paper	Contrast/Criticism
Nitta and Niwa (1994)	<i>However, using the co-occurrence statistics requires a huge corpus that covers even most rare words.</i> (S-5, 9503025)
Resnik (1995)	<i>However, for many tasks, one is interested in relationships among word senses, not words.</i> (S-1, 9511006)
Li and Abe (1996)	<i>Here, we restrict our attention on ‘hard clustering’ (i.e., each word must belong to exactly one class), in part because we are interested in comparing the thesauri constructed by our method with existing hand-made thesauri.</i> (S-80, 9605014)
Li and Yamanishi (1997)	<i>Although the finite mixture model has already been used elsewhere in NLP (Jelinek and Mercer, 1980; Pereira et al., 1993), this is the first work, to our knowledge, that uses it in the context of document classification.</i> (S-13, 9705005)

**Figure 6**  
Contrasting and Criticizing Citations to Pereira et al. (1993) in Other Articles

Similar paper	Similarity
Dagan et al. (1994)	<i>Following Pereira et al. (1993) we measure word similarity by relative entropy, or KL distance ...</i> (S-63, 9405001)
Li and Abe (1996)	<i>Perhaps the method proposed by Pereira et al. (1993) is the most relevant in our context.</i> (S-73, 9605014)

**Figure 7**  
Supportive and Continuing Citations to Pereira et al. (1993) in Other Articles

In the previous examples (Figures 5, 6 and 7), we used manual annotation instead

of the system output. The reason for this is that our system, while annotating sentences rhetorically, does not associate formal citations with the statements expressing contrast or continuation. Both Lawrence et al. (1999) and Nanba and Okumura (1999) assume implicitly that the sentence containing the formal citation will also contain the statement of stance towards the cited work, or at least enough context for the reader to infer it. However, we found that citations and stance statements often do not occur in the same sentence, particularly not for CONTRAST sentences. For instance, the sentences given in Figure 6 are the best sentences expressing the contrast, but only one out of four of these sentences contains a formal citation. Indeed, the more important a particular criticized work is to the current approach, the more space it will be afforded in the article, because the approach is first identified and described (and, as MacRoberts and MacRoberts (1984) found, the approach is praised first in order to “soften” the criticism); only then is the criticism itself given. This textual distance between the criticising statement and the formal citation seems to necessitate the task of automatically finding these associations. Once this association has been performed, our system with its categorial distinction between neutral work (OTHER) and stance towards work (BASIS, CONTRAST) buys an advantage over simpler citation analysis tools, as the context displayed to the user does not necessarily have to contain the citation itself and can therefore be concise.

### 3.3 Conclusion

This look at summary design and summary tasks has shown that the rhetorical categories produced by our system can create useful summaries which are fundamentally different from those generated by other systems. For instance, AIM, CONTRAST and BASIS sentences are direct input into many different types of summaries, and it is thus important to ensure that the system recognizes material that falls into those categories.

This task-orientation helped us develop the list of relevant rhetorical features. But we also wanted to make sure human annotators could distinguish between the different rhetorical features with great reliability, since that would help in the development of the gold standard used during system training and system evaluation. It is the development of this gold standard we turn to next.

## 4 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 this interpretation. This is typically done by showing that they can apply the analysis 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 “naive” subjects, i.e., not computational linguists. Apart from the cognitive validation of our analysis, high agreement is also the standard way to check if the annotated material can in principle be used as a training corpus for a statistical classification (described in section 5); noisy and unreliably annotated training material can be expected to deteriorate the classification performance.

In tasks where humans only agree up to a certain point, it is also useful to measure human performance and consider this as an upper bound. As the best possible performance an automatic process could have in such tasks is to be indistinguishable from human performance, one can compare the agreement of a pool of human annotators before and after a machine is added to the pool. The upper bound is reached if agreement in the new pool does not decrease.

## 4.1 Corpus

The annotated development corpus covers 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 corpus contains articles deposited there between 1994 and 1997.

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 rich XML markup to the corpus: titles, authors, conference, date, abstract, sections, headlines, paragraphs and sentences are marked up. Equations, tables, images are removed and replaced by place holders. Bibliography lists are marked and parsed. Citations and occurrences of author names in running text are recognized. Self citations are recognized and marked as such. (Linguistic) example sentences and example pseudocode are manually marked as such, such that clean textual material, i.e., the scientific text without interruptions, is isolated for automatic processing. The implementation uses the TTT software (Grover, Mikheev, and Matheson, 1999).

## 4.2 Annotation of Rhetorical Status

**4.2.1 Rationale and Experimental Design** We describe here the annotation experiment with the rhetorical annotation scheme presented in section 2.2 (cf. Teufel et al. (1999) for more detail). We use three task-trained annotators: 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.

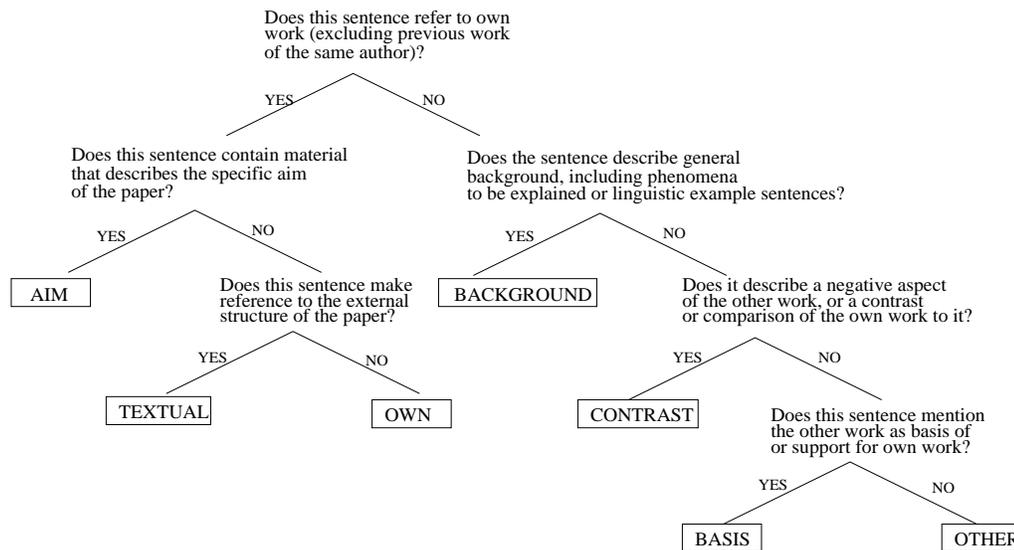
Annotators received a total of 20 hours of training. Written guidelines (17 pages) describe the semantics of the categories, ambiguous cases and decision strategies. Part of the instructions include the decision tree reproduced in Figure 8. Twenty-five articles were used for annotation. As at the time, no annotation tool was available, annotation took place on paper; the categories were later transferred into the electronic versions of the articles by hand. Skim-reading and annotation took typically between 20–30 minutes per article, but there were time restrictions. No communication between the annotators was allowed during annotation. Six weeks after initial annotation, annotators were asked to re-annotate 6 random articles out of the 25.

We measure 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



**Figure 8**  
Decision Tree for Rhetorical Annotation

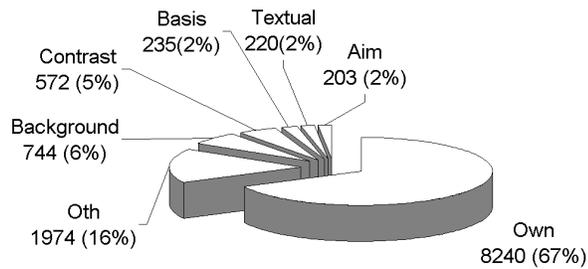
when agreement is perfect, and -1 when there is a perfect negative correlation.  $K=0$  is defined as the level of agreement which would be reached by random annotation using the same distribution of categories as the real annotators did.

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 6); we use this agreement as one of the performance measures of the system.

**4.2.2 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%, 90% agreement. Reproducibility was measured at  $K=.71$  ( $N=4261$ ,  $k=3$ ), which is equivalent to 87% agreement. How should the Kappa values be interpreted? 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",  $.61-.8$  as "substantial", and  $.81-1.0$  as "almost perfect". According to these guidelines, our results can be considered reliable, substantial annotation.

Figure 9 shows that the distribution of the seven categories is very skewed, with 67% of all sentences being classified as OWN.

Figure 10 shows a confusion matrix between two annotators. Where does the remaining disagreement come from? 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 that category 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  $K=.71$ ,



**Figure 9**  
Distribution of Rhetorical Categories (Entire Document)

		ANNOTATOR B							Total
		AIM	CTR	TXT	OWN	BKG	BAS	OTH	
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 10**  
Confusion Matrix between Annotators B and C

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 existence of the guidelines is an important factor here.

The annotators were less consistent at determining BASIS ( $K=.49$ ) and CONTRAST ( $K=.59$ ). This same picture emerges if we look at precision and recall of single categories between two annotators (cf. Figure 11). 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 are usually found in typical locations (beginning or end of the introduction section) and are explicitly marked with meta-discourse, whereas CONTRAST, and even more so BASIS, 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 11**

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 in 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%, 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 take particularly great care in the abstracts to make clear who a certain statement is attributed to.

When we compared agreement with the full annotation scheme of all seven categories between abstract and entire document, we also found higher reproducibility in the abstracts ( $K=.79$ ), but the effect was much weaker.

So it might be the case that abstracts are easier to annotate than the rest of the paper, but this does not necessarily make them the best starting point for the definition of gold standards. As foreshadowed in section 2.4, abstracts do not contain many different 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 homogenous. When we inspected the rhetorical structures of abstracts in terms of sequences of rhetorical zones, we found a high amount 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 produced by professional abstractors (Liddy, 1991). 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 our gold standard.

### 4.3 Annotation of Relevance

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

In order to establish alignment between summary and document sentences, we used a semi-automatic method, using 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 similar semantics of the two sentences. The following sentence pair illustrates a *direct match*:

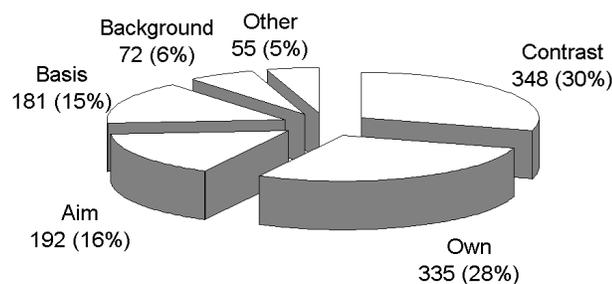
**Summary:** In understanding a reference, an agent determines his confidence in its adequacy as a means of identifying the referent.

**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 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 use abstract alignment as direct 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 (5 to 28 sentences per paper), resulting in 1183 sentences. She used the rhetorical zones and abstract similarity as aides in the relevance decision, but also skim-read the whole paper before making the decision.

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 summarization. If a selected sentence had the rhetorical class TEXTUAL, it was reclassified into one of the other six categories. Figure 12 shows the resulting category distribution amongst these 1183 sentences, which is far more evenly distributed than the one covering *all* sentences (cf. Figure 9). CONTRAST and OWN are the two most frequent categories.



**Figure 12**  
Distribution of Rhetorical Categories (Relevant Sentences)

We did not verify the relevance annotation with human experiments. We accept that there is no one best gold standard; instead, the set of sentences chosen by the human annotator make up *one* possible gold standard. More important, in our opinion, is that humans can agree as to what the rhetorical status of these sentences is. This assumption is in line with Liddy's observation about professional abstractors performing content 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 sentences selected by the first author, into the six rhetorical categories. The sentences are presented in order of occurrence in the document, but without any context in terms of surrounding sentences. We measure 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 4.2.2).

## 5 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 used for sentence extraction. Human annotation is used as a training material to learn the associations between these sentential features and the target (human) features, and also as gold standard for intrinsic system evaluation.

A simpler machine learning approach using only word frequency information and no other features, as typical in the task of text classification, could have been used (and indeed Nanba and Okumura (1999) do so for classifying citation contexts). To test if this was enough, we performed a text categorization experiment, using the Rainbow implementation of a Naive 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 already 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 (cf. the discussion in section 7.2). 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 features as possible. We predicted that this would increase precision and recall on the categories we are interested in. The text classification experiment, however, provides a non-trivial baseline for comparison with our intrinsic system evaluation (section 6).

### 5.1 Classifiers

We use a Naive Bayesian model as in Kupiec et al.'s (1995) experiment, cf. Figure 13.

*Sentential features* are collected for each sentence (Figure 14 gives an overview of the features we used). Learning is supervised: in the training phase, associations between these features and human-provided target-categories (relevant/non-relevant or the 7

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

$P(C F_0, \dots, F_{n-1})$ :	Probability that a sentence has target category $C$ , given its feature values $F_0, \dots, F_{n-1}$ ;
$P(C)$ :	(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 13**

Naive Bayesian Classifier

categories) are learned. 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 features of the sentence.

## 5.2 Features from Sentence Extraction

Some of the features in our feature pool are unique to our approach; they will be described in section 5.3. Others are borrowed from the text extraction literature (Paice, 1990) or related tasks and adapted to the problem of determining rhetorical status; we will describe these features here.

*Absolute location of a sentence:* In the news domain, sentence location is one of the unbeatable features 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 *own* method can be expected to be found towards the end of the article, whereas limitations of *other* people’s work often occur in the introduction. Rhetorical segmentation is not linear: we observed smaller rhetorical zones towards the beginning and the end of the article. We model this by assigning location values in a similar non-linear way, cf. Figure 15.

*Section structure:* Sections can have internal structuring; for instance, sentences towards the beginning of a section often have a summarizing 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.

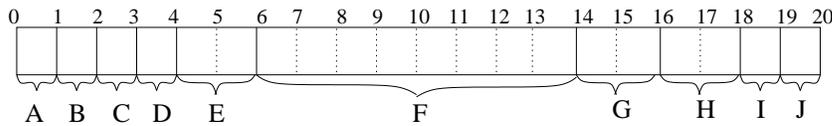
*Paragraph structure:* In many genres, paragraphs also have internal structure (Wiebe, 1994), with high-level or summarizing sentences occurring more often at the periphery of paragraphs. In this feature, sentences are distinguished into those leading or ending a paragraph, and all others.

*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 recognized by 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 matches, the sentence receives the value *Non-Prototypical*.

*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,

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 <i>Non-Prototypical</i>
Sentence length	Length	Is the sentence longer than a certain threshold, measured in words?	Yes or No
Content Features	Tf/idf	Does the sentence contain "significant terms" as determined by the <i>tf/idf</i> measure?	Yes or No
	Title	Does the sentence contain words also occurring in the title or headlines?	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 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?	Author Name, or None, or {Self Citation or not} X {Beginning, Middle, End}
Formulaic Expression	Formulaic	Type of formulaic expression occurring in sentence	18 Types of Formulaic Expressions + 9 Agent Types or None
Agentivity	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
History	History	Most probable previous category	7 Target Categories + "BEGIN"

**Figure 14**  
Overview of Feature Pool



**Figure 15**  
Values of Location Feature

by comparison to a fixed threshold (12 words).

*Title word contents:* Sentences containing many “content-bearing” words have been hypothesized 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 alternatively 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. The details of this analysis are given in section 5.4.

*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).

We automatically recognize 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 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 reports whether it is a self-citation, and 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 cannot be determined with certainty (it depends on the previous classification). 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 the application

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

**Figure 16**  
Formulaic Expression Lexicon (640 patterns, 18 classes)

in (Barzilay et al., 2000).

We will now turn to the last three features in our feature pool, the meta-discourse features.

### 5.3 Meta-Discourse Features

As our target categories are defined with respect to argumentation and authors' attitude towards other work, we hypothesize that another useful feature is explicit *meta-discourse* in scientific text: phrases like “*we argue that*” and “*in contrast to common belief, we*”. Meta-discourse, commonly defined as *discourse about discourse*, is a name for all those statements which fulfill other functions but to convey pure propositional contents in a text. It is ubiquitous in scientific writing: Hyland (1998) found a meta-discourse phrase on average after every 15 words in running text (mostly hedges).

A large proportion of scientific meta-discourse is conventionalized, particularly in 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 recognized 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*”. We model this phenomenon with a list of phrases modelled by regular expressions, similar to Paice's grammar (the feature `Formulaic`). Our list is divided into 18 semantic classes (cf. Figure 16). These classes model semantic indicators which we hypothesized to be important for rhetorical classification; the fact that phrases are clustered is a simple way of dealing with data sparseness. In fact, our experiments in section 6.1.2 will show the usefulness of the semantic clusters: the clustered list performs much better than the unclustered list (i.e., using the string itself as a value instead of its semantic class).

However, we noticed a large number of meta-discourse statements in our corpus which are less formalized: statements about aspects of the problem-solving process or

- We employ Suzuki's algorithm to learn case frame patterns as dendroid distributions. (S-23, 9605013)
- Our method combines similarity-based estimates with Katz's back-off scheme, which is widely used for language modeling in speech recognition. (S-151, 9405001)
- Thus, we base our model on the work of Clark and Wilkes-Gibbs (1986), and Heeman and Hirst (1992) ... (S-15, 9405013)
- The starting point for this work was Scha and Polanyi's discourse grammar (Scha and Polanyi, 1988; Pruest et al., 1994). (S-4, 9502018)
- We use the framework for the allocation and transfer of control of Whittaker and Stenton (1988). (S-36, 9504007)
- Following Laur (1993), we consider simple prepositions (like "in") as well as prepositional phrases (like "in front of"). (S-48, 9503007)
- Our lexicon is based on a finite-state transducer lexicon (Karttunen et al., 1992). (S-2, 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)). (S-116, 9408014)

**Figure 17**  
Statements Expressing Research Continuation

the relation to other work. Figure 17, for instance, shows that there are many ways to express the fact that one piece of work is based on some previous other work. The resulting sentences do not look similar on the surface: in some sentences the syntactic subject is a method, in others it is the authors, and in others the originators of the method. Also, the verbs are very different ("*base, be related, use, follow*"). Many sentences seem metaphoric for the metaphors of change and creation. This wide range of linguistic expression presents a challenge for recognition and correct classification by standard IE patterns. We use a more complicated mechanism which recognizes the voice of a sentence on the basis of parts of speech (POS) and then independently recognizes and classifies subject-like entities ("agents") and predictates ("actions"), using a lexicon of semantic classes.

#### 5.4 Agent and Action Recognition

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 recognize many of them automatically. The types of roles that agents play in the argumentation are fixed and can be enumerated: as rivals, contributors of part of the solution ("*they*"), the entire research community in the field, or the authors of the paper themselves ("*we*"). Note the similarity of these agents to the three kinds of intellectual attribution that we described in section 2.1. 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*". Like in the Formulaic feature, similar agents and actions are generalized and clustered together to avoid data sparseness.

We use a manually created lexicon for patterns of agents, containing 167 patterns and 13 Types of Agents (Figure 18). The main three agent types we distinguish are US\_AGENT, THEM\_AGENT and GENERAL\_AGENT. A fourth type is US\_PREVIOUS\_AGENT (the authors, but in a *previous* paper).

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

**Figure 18**  
Agent Lexicon

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

**Figure 19**  
Action 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 with ambiguous reference (“*this system*”), namely REF\_AGENT, REF\_US\_AGENT, THEM-PRONOUN\_AGENT, and AIM\_REF\_AGENT. The total of 167 patterns in the lexicon expands to many more strings as we use a replace mechanism (e.g., the place holder **WORK\_NOUN** in the 6th row of Figure 18 can be replaced by a set of 37 nouns including “*theory, method, prototype, algorithm*”).

Agent classes were created based on intuition, but then 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 distribution of the target categories, as such features are not distinctive. We measured associations using the loglikelihood 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 where at least one category showed a high association ( $g_{score} > 5.0$ ), as that means that the distribution of this verb was significantly different from the overall distribution. The last column in Figure 18 shows that the classes THEM\_PRONOUN, GENERAL, SOLUTION, PROBLEM and REF were removed: this improved the performance of the Agent feature.

As far as intellectual attribution is concerned, we noticed that statements in a segment *without* any explicit attribution are often interpreted as belonging to the last explicit attribution (which often occurs at the beginning of an attribution zone, e.g., “*Other researchers claim that*”). We model this with a modified agent feature, which keeps track of previously recognized agents (feature SegAgent); unmarked sentences receive these previous agents as a value. Wiebe (1994) reports a similar segment-based feature to be useful in her experiments.

For verbs, we use a manually created action lexicon containing 365 verbs (summarized in Figure 19). 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*”). It is necessary to recognize negation, as the semantics of “*not solving*” is closer to “*being problematic*” than it is to “*solving*”.

The following classes were removed by the  $g_{score}$  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 19), AWARE only in positive context (“+” in last column). The following classes had low counts in 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 patterns over POS. A full parse would probably improve results, but time constraints forced us to develop a solution instead which can deal with corpora in the range of hundreds of thousands of sentences fast. 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. The processing assumes that commas and other finite verbs are 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. In active clauses, the agent pattern is matched in assumed subject position; in passive clauses, it is assumed to be found within a prepositional phrase headed by “*by*”.

As our classifier requires unique values of each feature per classified item, we have to choose one value in case a sentence contains 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 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 corpus which contain finite verbs at all (they happened to contain a total of 184 finite verbs), the tagger 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; errors are mostly 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 misrecognized). 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%). One error was caused by a POS-tagging error. In the remaining 5 (less severe) errors the pattern covered only *part* of a subject NP (typically the NP in a postmodifying PP), as in the phrase “*the problem with these approaches*” (where only “*these approaches*” were recognized, nevertheless leading to the correct agent classification). All in all, we consider the two features to be adequately robust to serve as sentential features in our system.

## 6 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 rhetorical status, namely OTHER, OWN and BACKGROUND, are more frequent but might also be extracted into the summary.

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

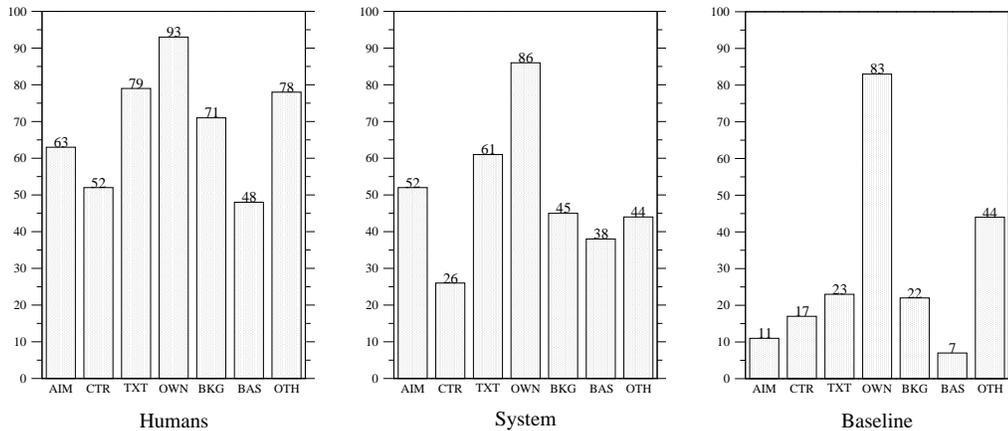
- We first report precision and recall values of all categories, in comparison to human performance and the text categorization baseline, as we are primarily interested in good performance on the categories AIM, CONTRAST, BASIS and BACKGROUND.
- We are also interested in good overall classification performance, 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.

### 6.1 Determination of Rhetorical Status

	AIM	CONTR.	TEXTUAL	OWN	BACKGR.	BASIS	OTHER
	p/r	p/r	p/r	p/r	p/r	p/r	p/r
System	44/65	34/20	57/66	84/88	40/50	37/40	52/39
Baseline	30/7	31/12	56/15	78/90	32/17	15/5	47/42
Humans	72/56	50/55	79/79	94/92	68/75	82/34	74/83

Figure 20

Performance per Category: Precision and Recall



**Figure 21**  
Performance per Category: F-measure

**6.1.1 Overall Results** 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 testing on seen data. Figure 20 and 21 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. The results are in the range of F-measures<sup>1</sup> from .52 (AIM), .61 (TEXTUAL) to .45 (BACKGROUND), .38 (BASIS) and .26 (CONTRAST). The recall for some categories is relatively low, which we do not find worrying, as our gold standard is designed to contain a lot of redundant information for the same category. However, low precision (e.g., 34% for CONTRAST, in contrast to human precision of 55%) could potentially present a problem for later steps.

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). Though no direct comparison with Kupiec et al.'s results is possible (their relevant sentences do not directly map into one of our categories), they are probably most comparable to our AIM sentences. In that case, our precision and recall of 44% and 65% compare favourably to theirs (42% and 42%).

<sup>1</sup> F-measure, a convenient way of reporting precision and recall in one value, has been defined by van Rijsbergen (1979) as  $2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$

Figure 22 shows a confusion matrix between one annotator and the system.

		MACHINE							Total
		AIM	CTR	TXT	OWN	BKG	BAS	OTH	
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 22**  
Confusion Matrix: Human vs. Automatic Annotation

With respect to overall agreement, Figure 23 shows the results in terms of three overall measures: Kappa, percentage accuracy, and Macro-F. We define Macro-F as the mean of the F-measures of all seven categories, as did Lewis (1991). The reason why we use Macro-F and Kappa is that our distribution is very skewed, and that we want to measure success particularly on the rare categories which are needed for our final task: AIM, BASIS and CONTRAST. Therefore, we need to make sure that the performance on the frequent, but less relevant category OWN is not overestimated, as is the case in micro-averaging techniques like traditional accuracy. This situation has parallels to information retrieval, where precision and recall are used because accuracy overestimates the performance on irrelevant items.

In the case of Macro-F, the rare categories are treated as one unit, just as the frequent ones are, so the classification success of the individual items in rare categories is given more importance than classification success of frequent category items, which is exactly the effect we're looking for. 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, such classifications which incorrectly favour frequent categories are punished due to a high random agreement. This effect can be shown best using the baselines. The most ambitious baseline we consider is the output of a text categorization system, as described in section 5. Other possible baselines, which are all easier to beat, include the most-frequent word classification. This is a "baseline" which turns out to be trivial, as it does not extract *any* of the rare rhetorical categories we are particularly interested in, and which 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 a hard-to-beat 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	Text Class.	Random	Random (Distr.)	Most Freq.	3 Humans
	Comparison to One Human Annotator					
Kappa	.45	.30	-.10	0	-.13	.71
Accuracy	.73	.72	.14	.48	.67	.87
Macro-F	.50	.30	.09	.14	.11	.69

**Figure 23**

Overall Classification Results

**6.1.2 Feature Impact** The previous results were compiled by 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 if 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 classify each sentence as `OWN` (and therefore achieve  $K=-.12$ ): `Relative Location`, `Paragraphs`, `tf/idf`, `Title`, `Sentence Length`, `Tense`, `Voice`, `Modality`. `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 very likely that another `BACKGROUND` sentence will follow after a `BACKGROUND` sentence.

However, each of these features still contributes to the final result: if taken out of the feature pool, classification performance decreases. How can this be, given that the 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 Naive 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.

It is also interesting to see that certain categories are disambiguated particularly well by certain features (cf. Figure 24). 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 rhetori-

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 24**

Precision and Recall of Rhetorical Classification, Individual Features

cal categories have typical cue phrases associated with them (whereas not all categories might have a preferred location in the document). Both `Agent` and `Action` features disambiguate categories which many of the other 12 features alone cannot disambiguate (e.g., `CONTRAST`), and `SegAgent` additionally contributes towards the determination of `BACKGROUND` zones (along with the `Formulaic` and the `Absolute Location` feature). Not surprisingly, `Location` and `History` are the features particularly useful for detecting `BACKGROUND` sentences.

## 6.2 Relevance Determination

The main workhorse in our implementation is the classifier for rhetorical status which we evaluate in the previous section. The next step is the determination of relevant sentences in the text. One solution for relevance decision is 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.

However, we already know that comparison with a fixed gold standard is a harsh way of comparison, as many other extracts might have been possible. How do the numbers compare to the quality of the end product? Figure 25 shows the output of the system for the example paper (all `AIM`, `CONTRAST` and `BASIS` sentences). The second column gives the human rhetorical role judgement (a tick for correct guesses, or a different category for incorrect ones). The direct comparison shows that 10 out of the 15 extracted sentences have been classified correctly, but we also see that rhetorical status is not always a straightforward distinction. 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 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, one can see why the agent and action features influenced the system to decide (plausibly) that the first part of sentence 21 had a comparative aspect. All in all, we think that the extracted material conveys the rhetorical status adequately, and that the end result provides considerable added value when compared to sentence extracts.

In terms of relevance, the asterisk in Figure 25 marks sentences which the human judge found particularly relevant in the overall context (cf. the full set in Figure 2). 6 out of all 15 sentences, and 6 out of the 10 sentences which received the correct rhetorical status, were judged relevant by this measure.

Figure 26 reports the result of comparing the system’s output of correctly classified rhetorical categories to human judgement. In all cases, the results are far above the

System	Human		
AIM	(OTH)	8	<i>In Hindle's proposal, words are similar if we have strong statistical evidence that they tend to participate in the same events.</i>
	✓	* 10	<i>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.</i>
	✓	11	<i>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.</i>
	(OWN)	12	<i>More specifically, we model senses as probabilistic concepts or clusters <math>c</math> with corresponding cluster membership probabilities <math>EQN</math> for each word <math>w</math>.</i>
	✓	* 22	<i>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.</i>
	(CTR)	41	<i>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.</i>
	(OWN)	150	<i>We also evaluated asymmetric cluster models on a verb decision task closer to possible applications to disambiguation in language analysis.</i>
	✓	* 162	<i>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.</i>
BAS	✓	19	<i>The corpus used in our first experiment was derived from newswire text automatically parsed by Hindle's parser Fidditch (Hindle, 1993).</i>
	✓	20	<i>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).</i>
	✓	* 113	<i>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 <math>EQN</math> following an annealing schedule.</i>
CTR	✓	* 9	<i>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.</i>
	✓	* 14	<i>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.</i>
	(OWN)	21	<i>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".</i>
	✓	43	<i>This is a useful advantage of our method compared with agglomerative clustering techniques that need to compare individual objects being considered for grouping.</i>

Figure 25

System Output for Example Paper

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,

	AIM		CONTR.		BASIS		BACKGROUND			
	p	r	p	r	p	r	without classifier		with classifier	
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 26**

Relevance by Human Selection: Precision and Recall

simply taking all 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 26) raises precision for BACKGROUND sentences to 38%, while keeping recall high at 88%.

## 7 Discussion

### 7.1 Contribution

We have presented a new method for robust 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 by rhetorical aspects (“Contrast”) rather than by domain-specific aspects (“Perpetrator”), which 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 summarization 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 summarizer 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 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 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, beating 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%/70% for goal statements, 70%/24% for contrast, 71%/39% for continuation). We see these results as an indication that shallow discourse processing with a well-designed set of surface-based indicators is possible.

The extracted material and their rhetorical status can be used for task- and user-tailored, multi- and single document summaries. We have also argued for an incorporation of citation information into summaries, as the relatedness of scientific work to other work can provide important information of a different kind to the envisaged users of our summaries.

## 7.2 Related work

We propose in this paper a method for *content selection* from scientific articles which adds context (in the form of *rhetorical* context) to extracted material. Apart from the obvious parallels with work on sentence extraction and knowledge-based summarization and to automatic citation induction (section 3.2), our analysis of rhetorical status bears similarity to other strands of work in discourse linguistics.

Our type of rhetorical analysis is much less fine-grained than the rhetorical relations in Rhetorical Structure Theory (Mann and Thompson, 1987), and our processing is more shallow than Marcu's (1999) automatic RST analysis. In contrast to RST, 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. Also, our analysis is linear rather than hierarchical. 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.

Linear segmentation, similar to ours, is also used in the task of topic segmentation (Morris and Hirst, 1991; Hearst, 1997; Kan, Klavans, and McKeown, 1998). In these approaches, lexical phenomena like word and concept repetition are used in order to segment text into zones which are coherent with respect to topic. In our approach the zone breaks are given by shifts in *rhetorical organization* rather than by topic shifts. We see topic and rhetorical structure as two complementary discourse organizations, which create a cross-classification of text: one topic can span several rhetorical zones, and one rhetorical zone can contain many topics. We believe that both types of text structure contribute entirely different, yet useful, information for information compression and structuring tasks.

Our work is similar to other forms of empirical discourse analysis, e.g. dialogue act coding (Carletta et al., 1997; Alexandersson, Maier, and Reithinger, 1995; Stolcke et al., 2000). In both types of work, a text segment (in the case of dialogue act coding, an utterance) is assigned a flat rhetorical label indicating the function of this utterance in the dialogue context (e.g., "Backchannel" or "Answer"). However, in dialogue act coding applications, there are no utterances which are more or less important — all utterances should be correctly classified. This difference influences, for example, the metrics used to determine success of automatic methods: Kappa, accuracy and Micro-F are metrics for complete classification, whereas precision and recall measure retrieval with respect to a single class of relevant items.

The intellectual-attribution distinction which is part of our analysis is similar to Wiebe's (1994) notion of *evidentiality* or point of view in narrative (subjective vs. objective statements). Subjectivity is a property which might prove relevant for scientific papers also, as it is related to the attribution of authorship as well as to author stance.

However, there are obvious genre differences: in news reporting and narrative, segments presenting opinions and verbal reactions are openly subjective, whereas in scientific text, authors take great care to hide the subjective elements (Ziman, 1969).

### 7.3 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. We are planning to apply this or a similar method to our data and to compare the performance of automatically gained resources with manual ones.

One of our continuing research efforts concerns the semantic verb clusters described in section 5.4. 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: verbs in our texts often have a specialized meaning in the domain of scientific argumentation, which our lexicon readily encodes. Klavans and Kan's classes are also manually created, but they are based on Levin's classes as a starting point. 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 recognize the fact that it *is* indeed a named approach, and that it is associated with the authors (information available elsewhere in the text). It is very likely that such a treatment would improve results, as named approaches are frequent in our domain. Information about named approaches in themselves would also be an important aspect to include in summaries or citation maps (section 3.2).

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 (this feature had been excluded from the `agent` feature; we include it now in its 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  $K=.14$  (Naive Bayesian model), for `SegAgent` from  $K=.19$  to  $K=.22$ . This is a clear indication of the added value of anaphora

resolution for our task.

As far as the statistical classification results are concerned, 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 an Maximum Entropy model, RIPPER and decision trees; preliminary results do not show significant improvement over the Naive 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 before 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 — at the price of having to identify clauses, which are a linguistically less well-defined entity than sentences are.

Two further projects that we are currently pursuing are an extrinsic evaluation of the annotated lists achieved as output of our system, and the porting of the annotation scheme and feature determination to other genres, namely medical and legal texts.

## References

- Alexandersson, Jan, Elisabeth Maier, and Norbert Reithinger. 1995. A robust and efficient three-layered dialogue component for a speech-to-speech translation system. In *Proceedings of the Seventh Meeting of the European Chapter of the Association for Computational Linguistics*, pages 188–193.
- 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.
- Bazerman, Charles. 1985. Physicists reading physics, schema-laden purposes and purpose-laden schema. *Written Communication*, 2(1):3–23.
- 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.
- Carletta, Jean, Amy Isard, Stephen Isard, Jacqueline C. Kowtko, Gwyneth Doherty-Sneddon, and Anne H. Anderson. 1997. The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23(1):13–31.
- 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.
- Fenichel, Carol Hansen. 1981. Online searching: Measures that discriminate between users with different types of experience. *Journal of the American Society for Information Science*, 32(1):23–32.
- 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.
- Hylland, 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, M. P. 1984. *Rhetoric of Everyday English Texts*. George Allen and Unwin, London, UK.
- Kan, Min-Yen, Judith L. Klavans, and Kathleen R. McKeown. 1998. Linear segmentation and segment significance. In *Proceedings of the Sixth Workshop on Very Large Corpora (COLIN G/ACL-98)*, pages 197–205.
- Kircz, Joost G. 1991. The rhetorical structure of scientific articles: The case for argumentational analysis in information retrieval. *Journal of Documentation*, 47(4):354–372.
- 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.R. and G.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.
- MacRoberts, Michael H. and Barbara R. MacRoberts. 1984. The negational

- reference: Or the art of dissembling. *Social Studies of Science*, 14:91–94.
- 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.
- Moravcsik, Michael J. and Poovanalingan Murugesan. 1975. Some results on the function and quality of citations. *Social Studies of Science*, 5:88–91.
- 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.
- 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.
- Shum, Simon Buckingham. 1998. Evolving the web for scientific knowledge: First steps towards an “HCI knowledge web”. *Interfaces, British HCI Group Magazine*, 39:16–21.
- 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.
- Stolcke, Andreas, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Dan Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26:339–373.
- 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, 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 Inderjeet Mani and Mark T. Maybury, editors, *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).
- 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.
- Weinstock, Melvin. 1971. Citation indexes. In *Encyclopedia of Library and Information Science*, volume 5. Dekker, New York, NY, pages 16–40.
- 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.