Argumentative classification of extracted sentences as a first step towards flexible abstracting

Simone Teufel and Marc Moens
HCRC Language Technology Group
University of Edinburgh
2 Buccleuch Place
Edinburgh EH8 9LW, UK
S.Teufel@ed.ac.uk M.Moens@ed.ac.uk

Abstract

Knowledge about the rhetorical structure of a text is useful for automatic abstraction. We are interested in the automatic extraction of rhetorical units from the source text, units such as PROBLEM STATEMENT, CONCLUSIONS and RESULTS. We want to use such extracts to generate high-compression abstracts of scientific articles. In this paper, we present an extension of Kupiec, Pedersen and Chen's (1995) methodology for trainable statistical sentence extraction. Our extension additionally classifies the extracted sentences according to their rhetorical role.

1 Introduction

1.1 Flexible abstracting

Until recently, the world of research publications was heavily paper-oriented. Journals, dissertations and other publications were available only in paper form. To keep researchers informed of publications in their area of interest, secondary publishers produced journals with abstracts of research material. The main role of these abstracts was to act as a *decision* tool: on the basis of the abstract a researcher could decide whether the source text was worth a visit to the library or a letter to the author requesting a copy of the full article.

For reasons of consistency (and copyright) these abstracts often were not the abstracts produced by the original authors, but by professional abstractors, and written according to agreed guidelines and recommendations (Borko and Chatman, 1963). These guidelines suggest that such abstracts should be aimed at the "partially informed reader"—someone who knows enough about the field to understand the basic methodology and general goals of the paper but does not necessarily have enough of an overview of previous work to assess where a certain article is situated in the field or how articles are related to each other (Kircz, 1991). For a novice reader, such an abstract would be too terse; for experienced researchers the abstract would provide unnecessary detail. In addition, because the abstract is a pointer to an article not immediately available, the abstract has to be self-contained: the reader should be able to grasp the main goals and achievements of the full article without needing the source text for clarification.

Over the past few years this picture has changed dramatically. Research articles are now increasingly being made available on-line. Indeed, the goal of automated summarization presupposes that the full article is available in machine-readable form. As a result, abstracts will have different or additional functions from the ones they used to have.

A typical scenario might be one where a user receives a large quantity of machine-readable articles, for example in reply to a search query, from a database of scientific articles or from the Internet. In such a context, abstracts can still be used as a decision tool, to help the user decide which articles to look at first. But in this context abstracts could also be used as a *navigation* tool, helping users find their way through the retrieved document collection. When abstracts are generated as needed, rather than stored in a fixed form, they could show how certain articles are related to other articles in logical and chronological respect, e.g. they could summarize similarities between articles, indicating which of the retrieved articles share the same research questions or methodologies. This type of navigation within a set of papers can support users in making a more informed decision on how well a paper fits their information needs.

Abstracts also don't need to be self-contained anymore. They can contain pointers (e.g. in the form of hyperlinks) to certain passages in the full article. And they can be "embedded" in the source text, highlighting in context the most relevant sentences, as has been demonstrated with commercial products such as Microsoft's "AutoSummarize" feature in Word97.

Abstracts can thus play an important role for the non-linear reading of textual material—the process whereby readers efficiently take in the content of a text by jumping in seemingly arbitrary fashion from conclusion to table of contents, section headers, captions, etc. Nonlinear reading is typical for scientists (Pirelli et al., 1984; Bazerman, 1988); it serves to efficiently build a model of the text's structure as well as to ex-

tract the main concepts of the paper. However, O'Hara and Sellen (1997) have shown that nonlinear reading is something people only do well with paper: the physical properties of paper allow readers to quickly scan the document and jump back and forth without losing their place in the document. On-line display mechanisms do not as yet have such facilities. Embedded or otherwise contextualized abstracts can facilitate this process of nonlinear reading by revealing the text's logical and semantic organization.

The old type of abstract was a fixed, long-lived, stand-alone text, targeted at one particular type of user. The new type of abstract is more dynamic and user-responsive, generated automatically when needed and thus less long-lived. Even though such abstracts will be of a lower quality when compared to human-crafted abstracts, we predict that they will be of more use in many situations. It is the flexible automatic generation of such abstracts which we see as our long-term goal.

1.2 Our approach

We would like to develop a summarization system which is not tied to a particular scientific *domain*. The processing robustness needed for this, as well as the speed with which we would like to be able to deliver abstracts, suggests that a deep semantic analysis of the source text is not a viable option.

Many robust summarization systems have opted for statistical sentence extraction: systems have been designed which extract "important" sentences from a text, where the importance of the sentence is inferred from low-level properties which can be more or less objectively calculated. Over the years there have been many suggestions as to which low-level features can help determine the importance of a sentence in the context of a source text, such as stochastic measurements for the significance of key words in the sentence (Luhn, 1958), its location in the source text (Baxendale, 1958; Edmundson,1969), connections with other sentences (Skorochod'ko, 1972; Salton et al., 1994), and the presence of cue or indicator phrases (Paice, 1981) or of title words (Edmundson, 1969) . The result of this process is an *extract*, i.e. a collection of sentences selected verbatim from the text.

These extracts are then used as the abstract of the text. But this has a number of disadvantages. For one thing, they are just a collection of sentences, possibly difficult to interpret because of phenomena like unresolved anaphora and unexpected topic shifts. Postprocessing of the extracts can remove some of these shortcomings, e.g. by not using sentences in the extract which contain obviously anaphoric expressions or by including surrounding sentences into the extract which are likely to resolve the anaphora (Johnson et al., 1993). Of course, this may lead to extracts which are too long, or it might mean losing sentences which are crucial to the content of the source text, thereby reducing the value of the resulting extract.

But even if—after postprocessing—each individual sentence might be interpretable in isolation, that still does not mean that the extract as a whole will be easy to understand. Assuming that the text is coherent, people will try to fill in the semantics gaps between potentially unconnected sentences. In the act of doing so, they may introduce inappropriate semantics links and get the wrong idea about the content of the source text.

Another problem is that sentence extraction does not work very well for high compression summarization. Typical sentence extraction programs compress to about 10 or 15% of the original—for example, reducing a short newspaper article to a few sentences. Even if these sentences do not form a coherent text, that does not matter much: the extract is short enough to still make sense. But we are interested in summarizing longer texts, such as journal articles. Simple sentence extraction methods will reduce a 20-page article to a 2-page collection of unconnected sentences, a document surrogate which is not adequate as an abstract. Reducing the extract further to obtain a real abstract is difficult.

The reason for this difficulty is that once the abstract-worthy sentences have been extracted, the logical and rhetorical organization of the text is lost, and it becomes difficult to make sensible decisions on how to reduce the text further. To overcome this problem, we want to select abstract-worthy material from the source text, whilst at the same time keeping information about the overall rhetorical structure of the source text and of the role of each of the extract sentences in that rhetorical structure.

However, the full rhetorical structure of a paper (and the logical structure of the research it reports) is a very complex structure, and is difficult to model automatically. Although Marcu (1997) presents an approach for the automated rhetorical analysis of texts, these texts are considerably shorter than the ones we are interested in summarizing. Rather than attempting a full rhetorical analysis of the source text, we wanted to extract just enough rhetorical information so as to be able to determine the rhetorical contribution of all and only the abstract-worthy sentences, without modeling domain knowledge or performing domain-sensitive reasoning. We make use of meta-comments in the text, phrases like *"we have presented a method for"*, and *"however, to our knowledge there is no"* which signal rhetorical status.

The abstract we envisage is construed as an argumentative template, where the slots represent certain argumentative or rhetorical roles, such as GOAL, ACHIEVEMENT, BACKGROUND, METHOD, etc. Abstracting means analysing the argumentative structure of the source text and identifying textual extracts which constitute appropriate fillers for the template. For each slot in the template (i.e. each rhetorical role) the system identifies a number of plausible fillers (i.e. text excerpts), with different levels of confidence. We call this collection of meaningful sentences *together* with in-

formation about their rhetorical role in the full article a *rhetorically annotated extract*.

Our idea of an abstract is thus more related to the *structured abstracts* which have become prevalent in the medical domain in the past decade (Broer, 1971; Adhoc, 1987; Rennie and Glas, 1991). Hartley et al. (1996) and Hartley and Sydes (1997) show in user studies that these abstracts are easier to read and more efficient for information assessment than traditional summaries.

In a further step (the generation of the real abstract), some of this information can be added or suppressed, in order to allow abstracts of varying length to be generated. For example, the amount of BACKGROUND information supplied in the abstract can be varied depending on whether users have been identified as novices or experienced readers. Rhetorical roles for which only low-probability evidence was found in the source document can be pruned until an abstract of the required length is reached.

Two questions arise from this approach. The first question is how the building blocks of the abstract template, i.e. the rhetorical roles, should be defined. This is a particular problem for our approach because very little is known about what our new type of abstract should look like. Most of the information on good abstracts deals with the world of paper, not with the use of on-line research publications. That means that we cannot take existing guidelines on how to produce balanced, informative, concise abstracts at face value; we will need to fall back on a different set of intuitions as to what constitutes a good abstract. To answer this question, we take research on the argumentative structure of research articles and their abstracts as our starting point. This will be discussed in section 2.

The second question is how a system can be trained to find suitable fillers in a source text to complete such a template. In section 3 we report on our experiments to train a system to automatically detect meaningful sentences in the source text together with their rhetorical role.

2 The argumentative structure of research articles and their abstracts

2.1 Rhetorical divisions in research articles

Scholarly articles serve the process of communicating scientific information. The communicative function of a scientific research article is thus very well-defined: to present and refer to the results of specific research (Salager, 1992). In some scientific domains research follows predictable patterns of methodology and also of presentation. A rigid, highly structured building plan for research articles has evolved as a result, where rhetorical divisions are clearly marked in section headers (Kintsch and van Dijk, 1978). Prototypical rhetorical divisions include *Introduction, Purpose, Experimental Design, Results, Discussion,* and *Conclusions.* This is very efficient: researchers in psycholinguistics, for example, know with great accuracy where in any given article to find the information on the number of participants in an experiment.

The papers in our corpus do not show this pattern. This has undoubtedly to do with the fact that our corpus consists of articles in computational linguistics and cognitive science. The papers draw from many subdisciplines, and most papers in our collection cannot be uniquely classified by sub-discipline, because they report on truly interdisciplinary research coming from different sub-disciplines. As a rough estimate, about 45% of the articles in our collection are predominantly technical in style, describing implementations (i.e. engineering solutions); about 25% report on research in theoretical linguistics, with an argumentative tenet; the remaining 30% are empirical (psycholinguistic or psychological experiments or corpus studies). As a result, we found a heterogeneous mixture of methodologies and traditions of presentation, with fewer prototypical rhetorical divisions than expected. Even though most of our articles have an introduction and conclusions (sometimes occurring under headers with different names), and almost all of them cite previous work, the presentation of the problem and the methodology/solution are idiosyncratic to the domain and personal writing style. Figure 1 shows the headers with the highest frequency for 123 examined papers—surprisingly few of them correspond to prototypical rhetorical divisions; the rest contain content specific terminology.

vspace4m

Freq.	Header
104	Introduction
56	Conclusion
27	Conclusions
21	Acknowledgments
15	Discussion
14	Results
11	Experimental Results
8	Related Work
8	Implementation
8	Evaluation
7	Example
7	Background

Figure 1: Headers with highest frequency from our collection

Apart from not being easily identified in our corpus, distinctions as expressed in rhetorical divisions are also too coarse for our purposes, namely to analyze scientific articles with respect to document structure, in a way which is flexible enough to cover the variety found in our corpus. A rhetorical division like *Introduction* can contain a problem statement, a motivation, a description of previous relevant work, and other such units. These smaller units are the ones that we are interested

in, units which Swales (1981) calls *moves*, where a move is defined as "a semantic unit related to the writer's purpose".

2.2 Author intentions and argumentation in research articles

Swales (1990) claims that the main communicative goal of an author, far from the unbiased reporting of research, is to convince readers of the validity and importance of the work, in order to have the paper reviewed positively and thus published. Argumentation is used to show that the presented research was a contribution to science: that the solution proposed in the paper either solves a *new* problem, or, if a *known* problem is addressed, that the presented solution is better than that proposed by other researchers.

Swales analyzed several hundred introduction sections of scientific research papers from two data collections: research articles in the physical sciences and a mixture of research articles from several science and engineering fields. This analysis led to his CARS model ("Create a Research Space") which is schematically depicted in Figure 2; the right hand side of the figure shows examples from our corpus. This model describes prototypical rhetorical building plans of introductions, based on the rhetorical moves that authors typically employ to fulfill the communicative goal of writing a paper. One such rhetorical move is to motivate the need for the research presented (Move 2), which can be done in different ways, e.g. by pointing out a weakness of a previous approach (Move 2A/B) or by explicitly stating the research question (Move 2C). Note that context plays an important role for the classification of a sentence in Swales' system: the example sentence for Move 2D (which characterizes the work actually reported in the article) would constitute a different move if it had appeared towards the end of the article, or under the heading *Future Work*.

Inspection of introduction sections in our corpus showed that the steps defined by Swales' CARS model describe the argumentation phenomena at the right level of abstraction for our purposes; the author's typical intentions, expressed as predictable textual moves, seem to generalize well to the domain of computational linguistics and cognitive science.

We also observed a wide range of meta-comments in our corpus (the underlined phrases in the right hand side of Figure 2). The source of our collection being an unmoderated medium, writing style in the articles varies from formal to quite informal. About a third of the articles were not written (or subsequently edited) by native speakers of English. Also, meta-comments need not be unambiguous with respect to the rhetorical move they signal. Nevertheless, we claim that overall, they are still good enough indicators of rhetorical status to be extremely useful in a practical, shallow kind of discourse analysis.

2.3 Argumentative structure of abstracts

Although we argued that guidelines for abstracts cannot be taken at face value when designing a high-level framework for on-line abstracts, there is ample information in the literature which can be used to inform decisions about a desirable argumentative structure for abstracts.

As is the case with the communicative function of the whole paper, the communicative function of an abstract is one of a narrow range of things: it can be an indicative abstract, reporting the topic of the full article, or an informative abstract, reporting the topic of the source article as well as its main findings and conclusions (Cremmins, 1996; Rowley, 1982). As in the case of research articles, the communicative function of abstracts has led to common expectations of their rhetorical building blocks, such as *General Background*, *Specific Problem* tackled by full article, *Main Results*, *Recommendations*, etc. Buxton and Meadows (1978) provide a comparative survey of the contents of abstracts in the physics domain. They studied which rhetorical section in the source text (*Introduction–Method–Result–Discussion*) corresponds to the information in the abstracts and found, for example, that abstracts tend not to report material from the *Method* section. There is similar research on medical abstracts (Salager-Meyer, 1992) and sociological and humanities abstracts (Milas-Bracovic, 1987).

There is a consensus about the content units of informative abstracts for such articles in the experimental sciences—the majority of information in the descriptive and prescriptive abstracting literature seems to have concentrated on experimental sciences. Most authors agree that informative abstracts should mention the following four information units (ANSI, 1979; ISO, 1976; Day, 1995; Rowley, 1982; Cremmins, 1996):

1. the PURPOSE or PROBLEM of the full article,
2. the SCOPE or METHODOLOGY,
3. the RESULTS,
4. and CONCLUSIONS or RECOMMENDATIONS

In line with these recommendations, Manning (1990) argues that informative abstracts are not a miniature version of the full article in the sense of offering "a paraphrase of every rhetorical section" of the source article.

There is more disagreement about "peripheral" content units, such as BACKGROUND, INCIDENTAL FINDINGS, FUTURE WORK, RELATED WORK, and DATA. Of particular interest to us is the content unit BACKGROUND. According to Alley (1996) , BACKGROUND is a useful content unit in an abstract if it is restricted to being the first sentence of the abstract. Other authors (Rowley, 1982; Cremmins, 1996) recommend not to include any background information at all. We believe that background information is potentially important,

MOVE 1: ESTABLISHING A TERRITORY		
1.1	**Claiming centrality**	• *The last decade has seen a growing interest in the application of machine learning to different kinds of linguistic domains ...*
1.2	**Making topic generalizations** (background knowledge) OR (description of phenomena)	• *The traditional approach has been to plot isoglosses, delineating regions where the same word is used for the same concept.* • *In the Japanese language, the causative and the change of voice are realized by agglutinations of those auxiliary verbs at the tail of current verbs.*
1.3	**Reviewing previous research**	• *Brown et al. (1992) suggest a class-based n-gram model in which words with similar cooccurrence distributions are clustered in word classes.*
MOVE 2: ESTABLISHING A NICHE		
2A	**Counter-claiming**	• *However, we argue that such formalisms offer little help to computational linguists in practice.*
or 2B	**Indicating a gap**	• *...no formal framework has been proposed, to our knowledge, to regulate the interaction between regular and exceptional grammatical resources.*
or 2C	**Question-Raising**	• *Can the restrictive power of a single constraint be estimated in a reliable way to allow an effective scheduling procedure being devised?*
or 2D	**Continuing a tradition**	• *The remaining issue is to find a way of better accounting for unsymmetrical accommodation.*
MOVE 3: OCCUPYING A NICHE		
3.1A	**Outlining purpose**	• *The aim of this paper is to examine the role that training plays in the tagging process ...*
or 3.1B	**Announcing present research**	• *In this paper, we argue that instead of applying the arbitration process to the discourse level, it should be applied to...*
3.2	**Announcing principle findings**	• *In our corpus study, we found that three types of utterances (prompts, repetitions and summaries) were consistently used to signal control shifts....*
3.3	**Indicating article structure**	• *This paper is organized as follows: We begin in Section [CREF] by examining the distribution of possessive pronouns...*

Figure 2: Swales' (1990) CARS model with illustrative examples from our corpus

especially for self-contained abstracts and for abstracts for novice readers.

There is similar disagreement over the content unit RELATED WORK. Cremmins (1996) states that it should not be included in an abstract unless the studies are replications or evaluations of earlier work. However, depending on the information need, previous work might actually have been central to the original information need of the user. Therefore, we want to preserve the possibility of including it in our modular abstract.

For the experiments reported in this paper, we chose the four generally accepted categories, but we had to redefine each class slightly in order to achieve higher domain-independence.

For example, we use the label SOLUTION/METHOD instead of METHODOLOGY/SCOPE: unlike in purely experimental research, where methodologies are long-lived research tools that are agreed upon in the field and do not change often, the range of possible methodologies in computational linguistics is vast, and a new, short-lived methodology might be invented just for the given problem-solving task, in which case the label "solution" seems more appropriate.

We added the two controversial roles RELATED WORK and BACKGROUND. And we added the role TOPIC, as the name of the research area or of the most general problem in the field. Thus, we ended up with the seven argumentative units listed in Figure 3.

Note that the labels of our annotation scheme can be naturally defined by rhetorical moves, such as the

Rhetorical role	
Background	Back
Topic/Aboutness	Topi
Related Work	Rwrk
Purpose/Problem	Pu/Pr
Solution/Method	Solu
Result	Resu
Conclusion/Claim	Co/Cl

Figure 3: Rhetorical roles in our annotation scheme

ones in Swales' CARS model. For example, Move 1.1 ("claiming centrality") provides good fillers for the Topic slot, whereas Problem, i.e. the specific problem of the paper, is very likely to be found in Move 2A–D ("indicating a gap").

Our annotation scheme forms the basis of the manual and automatic classification which is reported in the next section.

3 Our experiment

3.1 Previous work

Kupiec et al. (1995) introduce the notion of corpus-based abstracting: they recast the problem of sentence extraction as statistical classification. More specifically, they use supervised learning to automatically adjust feature weights with a Naive Bayesian classifier, combining the features (heuristics) mentioned in the literature. They used a corpus of research articles and corresponding summaries. The new idea in Kupiec *et al.*'s work is how they defined their *gold standards*. Gold standards are the class of sentences that, by definition, constitute the correct set of answers, usually defined by an expert in the field. The gold standard has to be defined independently and before the experiment. In Kupiec *et al.*'s work, the gold standard sentences are defined as the set of sentences in the source text that "align" with a sentence in the summary—i.e. sentences that show sufficient semantic and syntactic similarity with a summary sentence. The underlying reason is that a sentence in the source text is abstract-worthy if professional abstractors used it or parts of it when producing their summary. In Kupiec *et al.*'s corpus of 188 engineering articles with summaries written by professional abstractors, 79% of sentences in the summary also occurred in the source text with at most minor modifications.

Kupiec *et al.* then try to determine the characteristic properties of abstract-worthy sentences according to a number of features, viz. presence of particular cue phrases, location in the text, sentence length, occurrence of thematic words, and occurrence of proper names. Each document sentence receives a score for each of the features, resulting in an estimate for the sentence's probability to also occur in the summary. This probability is calculated for each feature value as a combination of the probability of the feature-value pair occurring in a sentence which is in the summary (successful case) and the probability that the feature-value pair occurs unconditionally.

Evaluation of the training relies on cross-validation: the model is trained on a training set of documents, leaving all documents from one journal out at a time (the current test set). The model is then used to extract candidate sentences from all documents of the test set. Evaluation measures co-selection between the extracted sentences and the gold standard sentences in precision (number of sentences extracted correctly over total number of sentences selected) and recall (number of sentences extracted correctly over total number of gold standard sentences). Since from any given test text as many sentences are selected as there are gold standard sentences, numerical values for precision and recall are the same. The precision/recall values of the individual heuristics range between 20–33%; the highest cumulative result (44%) was achieved using paragraph, fixed phrases (indicators) and sentence length features.

3.2 Abstracting as stepwise classification

We decided to perform the automatic generation of rhetorically annotated extracts by a process of repeated classification, borrowing the classification methodology from Kupiec *et al.* The basic procedure for the sentence extraction and classification experiment is the following:

Step one: Extraction of abstract-worthy sentences. We try to separate sentences which carry *any* rhetorical roles (grey set of sentences in Figure 4) from irrelevant sentences, which are by far the larger part of the text (white set of sentences in Figure 4). The output of this step is called the *intermediate extract*. Errors in this task will lead to the inclusion of irrelevant material in the extracts (false positives), or the exclusion of relevant material from the extracts (false negatives).

Step two: Identification of the correct rhetorical role. Once good sentence candidates have been identified, we classify them according to one of the seven rhetorical roles (in Figure 4, this corresponds to the sub-classification of the grey sentences). The output of this step is called a *rhetorically annotated extract*.

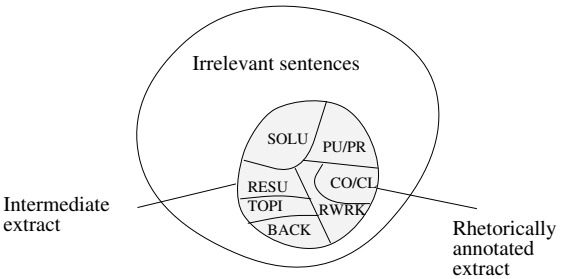

Figure 4: Abstracting as classification

We decided to split the task because we suspected that different heuristics would be more useful for the different tasks—a two-step process allows for the separation of these distinctions into two training processes.

Also, another motivation for the separation of the tasks stems from the fact that indicator phrases don't have to be unambiguous with respect to their argumentative status. For example, the phrase *"in this paper, we have"* is a very good overall relevance indicator, and it is quite likely that a sentence or paragraph starting with it will carry important global-level information. However, without an analysis of the following verb, we cannot be sure about the argumentative status of the extract. The sentence could continue with *"... used machine learning techniques for ..."*, in which case we have a solution instance; just as well, the sentence could be a conclusion (*"... argued that ..."*) or a problem statement (*"... attacked the hard problem of ..."*). Thus, the phrase *"in this paper we will"* is very useful for step one, but not useful for step two.

3.3 Corpus

Our corpus is a collection of 201 articles and their author-written summaries from different areas of computational linguistics and cognitive science, drawn from the computation and language archive (`http://xxx.lanl.gov/cmp-lg`). We assume that most of the articles had been accepted for publication in conference proceedings, although we have not verified this in each case. The documents were converted from LaTeX source into HTML in order to extract raw text and minimal structure automatically, then transformed into SGML format and manually corrected. We used all documents dated between 04/94 and 05/96 which we could semi-automatically retrieve with our conversion pipeline and which contained no less than 2,000 and no more than 10,000 words. The resulting corpus contains 568,000 word tokens; the average length of the documents is 187 sentences, the average length of the original summaries 4.7 sentences. In each text we marked up the following structural information: title, summary, headings, paragraph structure and sentences. We also removed tables, equations, figures, captions, references and cross references and replaced them by place holders (e.g. the symbol [REF] marks the place where a reference was cited in the text; [EQN] marks the place of equations).

We randomly divided our corpus into a training and test set of 123 documents which were further analyzed and annotated, and a remaining set of 78 documents which remain unseen. Only the first set was used for the experiments described here.

3.4 Annotation of gold standards

In line with Kupiec *et al.*'s method, we tried to use the summaries in our corpus for training and evaluation. However, the summaries of our articles were written by the authors themselves, and it is commonly assumed that author summaries are of a lower quality when compared to summaries by professional abstractors.

We first tested to which degree the authors' summaries reused sentences from the body of the document. In order to establish alignment between summary and document sentences, we used a semi-automatic method, assisted by a simple surface similarity measure which computed the longest common subsequence of non-stop-list words. Final alignment was decided by a human judge, where the criterion was similarity of semantic contents of the compared 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.

Unlike Kupiec *et al.*'s professional annotators, our authors had not reused document sentences to a large degree—we had a low 31% alignment rate as compared to Kupiec *et al.*'s 79%.

In addition to this, the authors had obviously not used a prototypical scheme to write their summaries, in contrast to professional abstractors surveyed by Liddy (1991). When we inspected the rhetorical contents of the sentences in the author summaries by applying our annotation scheme to them, we found that argumentative structure varied widely, even though most summaries are understandable and many are well-written. Some summaries are extremely short, and many of them are not self-contained, and would thus be difficult to understand for the partially informed reader. This again confirms the claim that author summaries are less systematically constructed than summaries by professional abstractors.

Because of the low alignment and the heterogeneous rhetorical structure of the summaries, we decided not to use them directly for annotation and evaluation. Annotation of the training corpus had to proceed in the following three steps:

1. Alignment of summary and document sentences (semi-automatic);
2. Additional annotation of further relevant sentences (manual);
3. Annotation of the argumentative status of these sentences (manual).

A human judge annotated additional abstract-worthy sentences in the source text. We gave no restrictions as to how many additional sentences were to be selected. After this process, our texts had two gold standards of different origin: gold standard A, consisting of aligned sentences; and gold standard B, consisting of sentences selected by the human judge, 948 sentences in total.

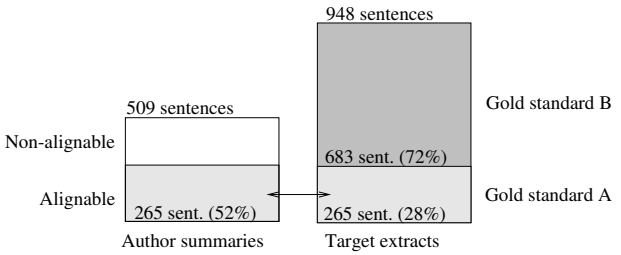

Figure 5: Composition of gold standards with respect to origin

Figure 5 shows the composition of gold standards: there are 2.5 times as many gold standard B sentences as there are gold standard A sentences. The alignment rate in our training and test set of 123 documents, which consists of the best-aligned documents, is 52% (the alignment rate of 31% refers to all 201 documents). With respect to compression (i.e. ratio of gold standard sentences to document sentences), our combined gold standards achieve 4.4% (as compared to Kupiec *et al.*'s 3.0% compression). Gold standard A had a compression of 1.2%, gold standard B 3.2%.

The second annotation step consisted of manually determining the argumentative roles for the abstract-worthy sentences (as defined in step one) for each article in the training set.

The following sentence with its rhetorical label illustrates this type of mark-up:

> Repeating the argument of Section 2, we conclude that a construction grammar that encodes the formal language [EQN] is at least an order of magnitude more compact than any lexicalized grammar that encodes this language. CONCLUSION/CLAIM

Difficulties encountered during annotation often concerned the status of a statement in the line of the argument, when the status was dependent on the context. For example, a weakness of the authors' solution might be classified as a limitation or as a local problem, depending on whether that problem will be solved later on in the given article. In cases of true ambiguity between two roles, we allowed for multiple annotation.

Another difficulty had to do with the fact that we annotated entire sentences: often, one sentence covers more than one role, as the following sentence illustrates:

> We also examined how utterance type related to topic shift and found that few interruptions introduced a new topic. PURPOSE/PROBLEM AND CONCLUSION/CLAIM

Figure 6 shows the composition of the gold standard sentences with respect to rhetorical roles. SOLUTION and PROBLEM are the most common rhetorical roles with about one third each of the judgements, the other roles sharing the last third. The least common role was RESULT.

There were 1172 instances of rhetorical roles in our 948 gold standard sentences. 232 sentences (24%) contained multiple mark-up (either ambiguous or concatenative). Figure 7 shows the distribution of *multiple* markup over the rhetorical roles, which is about proportional, except for a low involvement of BACKGROUND in multiple markup and a proportionally higher one for RELATED WORK and PROBLEM. We believe this is partly due to conceptual difficulties and partly due to concatenative markup: BACKGROUND sentences tend to contain nothing but background information, whereas the information units for PROBLEM statements and RELATED WORK tend to be smaller.

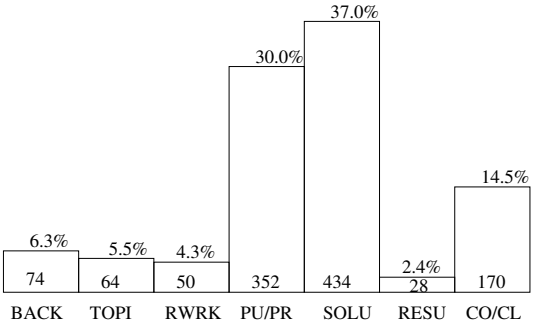

Figure 6: Composition of gold standard sentences with respect to rhetorical roles set

Rhetorical role	Multiple annotation	
BACKGROUND	16	(21%)
TOPIC/ABOUTNESS	25	(39%)
RELATED WORK	24	(48%)
PURPOSE/PROBLEM	168	(47%)
SOLUTION/METHOD	167	(38%)
RESULT	11	(39%)
CONCLUSION/CLAIM	64	(37%)

Figure 7: Percentages of judgements involving multiple annotation for the respective rhetorical roles

3.5 Heuristics Pool

We employed 7 heuristics in the two tasks: 4 of the heuristics used by Kupiec et al. (Indicator Quality Feature, Relative Location Feature, Sentence Length Feature and Thematic Word Feature), and 3 additional ones (Indicator Rhetorics Feature, Title Feature and Header Type Feature).

Indicator Quality Feature: The Indicator Quality Feature identifies meta-comments in a text, as opposed to subject matter. We use a list consisting of 1728 indicator phrases or formulaic expressions, such as communicative verbs and phrases related to argumentation and research activities. Our indicator phrase list was manually created by a cycle of inspection of extracted sentences and addition of indicator phrases to the list.

Figure 8 shows an extract from the indicator list; the first group of indicator phrases is centered around the concept *"argue"*, the second group uses the global indicator *"in this article"*, the third is centered around the concept *"attempt"*.

The largest part of these phrases is positive, but the last entry in Figure 8 illustrates a negative indicator phrase, typically occurring in the rhetorical division *Acknowledgements* (which is of no interest to content selection).

Indicator Phrase	**Quality Score**
we argued	2
we have argued	1
we have argued that	1
we will argue	1
what I have argued is	1
what we have argued is	1
This article	3
in this article	3
is an attempt to	1
I attempt to	2
I have attempted	2
I have attempted to	2
our work attempts	2
the present paper is an attempt	2
this paper is an attempt to	2
supported by grant	-1

Figure 8: An extract from the indicator list

Using the strings directly as values in a feature would result in a sparse distribution, and thus in an over-fitted feature, i.e. a feature that works well for the training data but not for different, but similar kinds of data. Thus, we classified the strings according to different criteria. For the Indicator Quality Feature, indicator phrases were manually classified into 5 quality classes according to their occurrence frequencies within the target extract sentences (cf. the column 'Quality Score' in Figure 8). The scores mirror the likelihood of a sentence containing the given indicator phrase to be included in the summary on a 5-valued scale from 'very likely to be included in a summary' to 'very unlikely'. For example, the likelihood of the phrase *"we argued"* to appear in the summary is higher than the likelihood of variations of this string in other tenses, a fact that is mirrored by its higher score of +2.

Indicator Rhetorics Feature: This feature tries to model the semantics (rhetorical contribution) of the phrases. Each indicator phrase was manually classified into one of 16 classes. Classes correspond to the 7 rhetorical roles (BACK, TOPI, RWRK, PU/PR, SOLU, RESU, CO/CL), and 8 confusion classes, viz. SOLU–PU/PR, SOLU–CO/CL, PU/PR–CO/CL, PU/PR–RWRK, PU/PR–BACK, CO/CL–RWRK, CO/CL–RESU, BACK–RWRK plus the value ZERO for phrases that do not predict a specific rhetorical role. The first group of phrases in Figure 8 (*"argue"*), for example, was classified as a most likely indicator of the rhetorical class CONCLUSION/CLAIM, and the third group (*"attempt"*) was classified as an indicator of PURPOSE/PROBLEM, whereas the second and fourth groups received the value ZERO.

Relative Location Feature: This feature distinguishes peripheral sentences in the document and within each paragraph, assuming a hierarchical organization of documents and paragraphs. The algorithm is sensitive to prototypical headings (e.g. *Introduction*); if such headings cannot be found, it uses a fixed range of paragraphs (first 7 and last 3 paragraphs). Document final and initial areas receive different values, but paragraph initial and final sentences are collapsed into one group.

Sentence Length Feature: All sentences under a certain length (current threshold: 15 tokens including punctuation) receive a 0 score, all sentences above the threshold a 1 score.

Thematic Word Feature: This feature is a variation of the "Term-frequency times inverse document frequency" (tf.idf) feature, a document specific keyword weighing method which is commonly used in Information Retrieval (Salton and McGill, 1993). It tries to identify key words that are characteristic for the contents of the document, viz. those of a medium range frequency relative to the overall collection. The 10 top-scoring words according to the tf.idf method are chosen as thematic words; sentence scores are then computed as a weighted count of thematic words in a sentence, meaned by sentence length. The 40 top-rated sentences obtain score 1, all others 0.

Title Feature: Words occurring in the title are good candidates for document specific concepts. The Title Feature score of a sentence is the mean frequency of title word occurrences (excluding stop-list words). The 18 top-scoring sentences receive the value 1, all other sentences 0. We also experimented with taking words occurring in all headings into account (these words were scored according to the tf.idf method) but received better results for title words only.

Header Type Feature: The rhetorical division that a sentence appears in can be a good indication of its rhetorical status. The Header Type Feature uses a list of prototypical header key words like *discussion, introduction, concluding remarks, conclusions.* Each sentence is assigned one of 15 values, depending on the header it appears under. Headers are classified as one of 14 prototypical groups if they contain one or more of the header key words (or a morphological variant of it); otherwise (i.e. if they contain only domain-specific strings) they are classified as 'non-prototypical'.

3.6 Classifiers

As in Kupiec *et al.*'s (1995) experiment, each document sentence receives scores for each of the features, resulting in an estimate for the sentence's probability to also

occur in the summary. This probability is calculated for each feature value as a combination of the probability of the feature-value pair occurring in a sentence which is in the summary (successful case) and the probability that the feature-value pair occurs unconditionally.

Kupiec *et al.*'s estimation for the probability that a given sentence is contained in the summary is:

$$P(s \in E|F_1,\ldots,F_k) \approx \frac{P(s\in E)\prod_{j=1}^{k} P(F_j|s\in E)}{\prod_{j=1}^{k} P(F_j)}$$

where

$P(s \in E\|F_1,\ldots,F_k)$:	Probability that sentence s in the source text is included in the intermediate extract E, given its feature values;
$P(s \in E)$:	compression rate (constant);
$P(F_j\| s \in E)$:	probability of feature-value pair occurring in a sentence which is in the extract;
$P(F_j)$:	probability that the feature-value pair occurs unconditionally;
k:	number of feature-value pairs;
F_j:	j-th feature-value pair.

For the second step, the probability that a certain sentence from the new base set (the intermediate extract) is associated with a rhetorical role is calculated analogously as follows:

$$P(e \in R_i|F_1,\ldots,F_k) \approx \frac{P(e\in R_i)\prod_{j=1}^{k} P(F_j|e\in R_i)}{\prod_{j=1}^{k} P(F_j)}$$

where

$P(e \in R_i\|F_1,\ldots,F_k)$:	Probability that sentence e in the intermediate extract is assigned the rhetorical role R_i, given its feature values;
$P(e \in R_i)$:	probability of role R_i in extract (unconditional of feature values);
$P(F_j\| e \in R_i)$:	probability of feature-value pair occurring in an extract sentence which has rhetorical role R_i;
$P(F_j)$:	probability that the feature-value pair occurs unconditionally in the extract;
k:	number of feature-value pairs;
F_j:	j-th feature-value pair.

Assuming statistical independence of the features, $P(F_j)$ (for the two different base sets), $P(F_j|\ s \in E)$ and $P(F_j|e \in R_i)$ can be estimated from the corpus for each F_j and each R_i. The second step returns a vector of probabilities for each sentence in a document (cf. Figure 9), with each cell in the vector corresponding to a rhetorical role. For each sentence, the role with the highest probability is chosen (cf. grey boxes).

0.9e-9	0.1e-9	0.2e-9	0.6e-10	0.3e-9	0.7e-10	0.9e-10	0
BACK	TOPI	RWRK	PU/PR	SOLU	RESU	CO/CL	
0.3e-12	0.9e-14	0.3e-14	0.6e-13	0.1e-17	0.7e-14	0.9e-12	1
BACK	TOPI	RWRK	PU/PR	SOLU	RESU	CO/CL	
0.6e-14	0.4e-10	0.9e-11	0.5e-10	0.3e-7	0.7e-8	0.1e-10	2
BACK	TOPI	RWRK	PU/PR	SOLU	RESU	CO/CL	
. . .							
0.4e-8	0.9e-10	0.3e-9	0.5e-10	0.6e-8	0.7e-9	0.1e-10	235
BACK	TOPI	RWRK	PU/PR	SOLU	RESU	CO/CL	

Figure 9: Probability vectors for document sentences No. 0, 1, 2 and 235

3.7 Evaluation

The evaluation we report here is based on co-selection between the gold standard sentences (i.e. target extracts) and the automatic results. This kind of evaluation is useful in a corpus-based approach like ours to fine-tune the single heuristics, but in our opinion final evaluation should not be based on co-selection with target extracts. Co-selection measures might give a distorted picture of the quality of an extract, because there might be many good abstracts/extracts, but a comparison with a target can only ever measure how well it approximates *one* of these. Real evaluation should be task-based, i.e. measure how well a certain document surrogate supports a human in fulfilling a certain task.

In our experiments, co-selection measures were used as follows: for extraction, co-selection reports how many of the extracted sentences had independently been identified as relevant sentences by the human annotator. For classification, co-selection reports how often the rhetorical roles identified by the algorithm were indeed the roles the human annotator had assigned. The numerical results reported for classification refer to the intermediate extract as a base set (i.e. those sentences that have been correctly identified in the first step). Cross-validation is used: the model is trained on a training set of documents, leaving a single document out at a time (the current test document). We did not have an indication as to subject matter like Kupiec *et al.* did (by journal name), so we chose to use all other documents but the single test document for training. After training, the model is used to extract candidate sentences from the test document, and co-selection values are measured.

Numerical values in the tables always give precision and recall rates as percentages. Due to the setup of the experiment (there are always as many sentences chosen as there are gold standards), precision and recall values are identical for extraction and for the *overall* results of classification. However, it is possible that precision and recall values for the classification of a *specific* rhetorical role differ. This is because it is possible that the algorithm overestimates the frequency of one role X at the

expense of another role Y, in which case the recall of X would increase, but the precision of X would decrease. For multiply-annotated gold standard sentences, a correct classification was scored when the algorithm identified *one* of the ambiguous roles correctly.

As a baseline for the first task we chose sentences from the beginning of the source text, which constituted a recall and precision of 28.0%. This "from-top" baseline is a more conservative baseline than random order: it is more difficult to beat, as prototypical document structure places a high percentage of relevant information in the beginning.

The baseline for the second task (classification) is computed by classifying each sentence as the most frequent role (SOLUTION); it stands at an amazing 40.1% which means that this task is statistically much easier than extraction.

3.8 Results

3.8.1 Extraction

Extraction	Indiv.	Cumul.
Indicator Quality Feature	54.4	54.4
Relative Location Feature	41.0	63.9
Sentence Length Feature	28.9	65.6
Title Feature	21.6	65.6
Header Type Feature	39.6	65.3
Thematic Word Feature	16.2	66.0
Indicator Rhetorics Feature	44.0	65.6
Baseline	28.0	

Figure 10: Impact of individual heuristics on extraction

Figure 10 summarizes the contribution of the features, individually and cumulatively. Precision and recall values for the features vary between 16.2% (Thematic Word Feature) and 54.4% (Indicator Quality Feature). The most successful combination of the 7 available heuristics at 66.0% actually excludes the Indicator Rhetorics Feature—including it would decrease the results slightly (by 0.4%). The fact that a subset of all heuristics achieves a better result than all heuristics taken together means that the combination of heuristics in our implementation is non-monotonic. Non-monotonicity would be an unfortunate property in a real world setting where there are no gold standards available, and where we have to rely on the fact that *each* heuristic in the pool contributes positively to the results. However, in the supervised experiments described here co-selection measures are used to fine-tune the heuristics, in order to identify weaknesses of features (or features that should be removed from the pool completely).

Also note that even such weak features as the Title Feature and Thematic Word Feature with precision and recall lower than the baseline can still contribute positively to the results, whereas the relatively strong Indicator Rhetorics Feature does not. This does not mean that the Indicator Rhetorics Feature is not a good feature, but only that it is not completely independent from the more successful features, contrary to assumption (in this case, it is probably very similar to the Indicator Quality Feature). Thus, how helpful a heuristic will be in combination with others cannot be judged from its individual performance alone, but also from its similarity to the other heuristics.

Overall, these results reconfirm the usefulness of Kupiec *et al.*'s method of heuristic combination. The method increases precision for the best feature by around 20%.

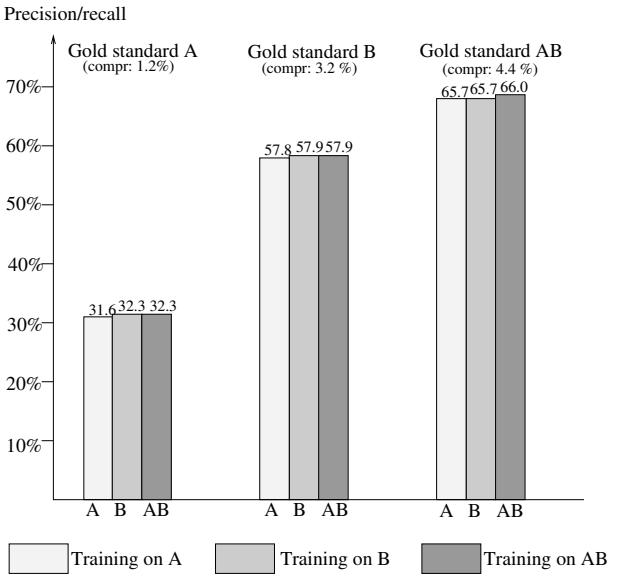

Figure 11: Influence of training material/gold standards

In order to see how the different origins of our gold standards contribute to the results, we trained three models (cf. Figure 11): one by training only on gold standard A sentences (light grey), one by training only on gold standards B (medium grey), and the third by training on both kinds of gold standards (dark grey). We then used the 3 models for 3 different tasks—first trying to identify A gold standards, then B gold standards and then both. Due to the higher compression of the task, extraction in the first task is statistically more difficult, which accounts for the much lower precision and recall values when compared to the other tasks. If we compare the values *within* extraction tasks, where the only difference is in *training*, the results show a surprising consistency: the distribution of heuristics values was almost identical between gold standards, no matter which gold standards we had trained our model on. The practical conclusion from this experiment is that we can get intermediate extracts of a similar quality (if we were to be content with these as end results) by training only on the relatively cheaply attainable gold standard A (alignment), rather than using the labor-intensive gold standard B (human judgement).

3.8.2 Classification

Classification	Indiv.	Cumul.
Indicator Rhetorics Feature	56.3	56.3
Relative Location Feature	46.5	63.8
Title Feature	40.0	64.2
Indicator Quality Feature	45.9	63.8
Sentence Length Feature	39.7	61.6
Thematic Word Feature	16.2	61.5
Header Type Feature	39.6	57.2
Baseline	40.1	

Figure 12: Impact of individual heuristics on classification

Figure 12 summarizes the contribution of the individual features for classification, taken individually and cumulatively. Precision and recall values for the features vary between 16.2% (Thematic Word Feature) and 56.3% (Indicator Rhetorics Feature). The most successful combination consisted of Indicator Rhetorics Feature, Relative Location Feature and Title Feature (with a combined precision/recall value of 64.2%). The combination is non-monotonic to a higher degree than in the extraction task: addition of the other 4 heuristics steadily decreased precision and recall to 57.2%.

Where does the system make errors? The confusion matrix in Figure 13 shows the distribution of machine and human classifications for the different roles (best heuristic combination), where the columns in the table refer to the roles assigned by our algorithm ("Machine") and the rows denote roles assigned in the gold standard sentences ("Human"). For example, out of the 227 SOLUTION gold standard sentences that the human judge identified, the system found 170 correctly; it misclassified 41 as PROBLEM and the remaining 16 as CONCLUSION. The grey boxes along the diagonal show the absolute numbers of successful machine classifications per role; also, precision and recall values of the automatic classification are given for each rhetorical role.

It is obvious that the system significantly underestimates low-frequency roles—there are only very few RELATED WORK and RESULT roles assigned by the system, and none at all for TOPIC. In comparison, the estimation of the frequency of the higher frequency roles is quite adequate.

The confusion matrix illustrates that our system often misclassifies PROBLEMS as SOLUTIONS (38 times) and SOLUTIONS as PROBLEMS (41 times). But these roles are often co-classified by the human judge, as Figure 14 shows: 113 out of the 434 SOLUTION instances and the 352 PROBLEM instances were co-classifications "PROBLEM and/or SOLUTION". Apart from ambiguities between PROBLEM and SOLUTION, there were also many misclassifications including these roles and CONCLUSION (cf. the hatched boxes in Figure 14). These were exactly the ones where our algorithm had a high percentage of misclassifications (cf. the hatched boxes in Figure 13), which implies that the low performance

MACHINE

HUMAN	Background	Topic	Related Work	Problem	Solution	Result	Conclusion	Total	Recall
Background	48			2	3			53	0.91
Topic	8	0	1	28	5		2	44	0.00
Related Work	5		0	9	4			18	0.00
Problem	2		2	137	38		10	189	0.72
Solution				41	170		16	227	0.75
Result				2	5	3	4	14	0.21
Conclusion	1			13	32	1	62	109	0.57
Total	65	0	3	232	257	4	94	654	0.64
Precision	0.75	0.00	0.00	0.59	0.66	0.75	0.65	0.64	

Figure 13: Confusion matrix for argumentative classification by roles (machine)

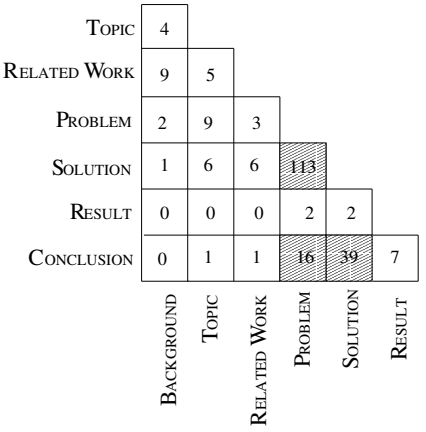

Figure 14: Number of sentences involved in multiple markup (gold standards)

of the system must be partly attributed to the inherent difficulty of the task. The distinction between these roles is conceptually difficult: conclusions are often statements *about* properties of the solution or *about* phenomena in the world (which are annotated as problems); problems and solutions co-occur often in the same sentence, and sometimes it is difficult to distinguish between a research goal and its solution, i.e. to find out if the sentence describes a goal in itself or a research step towards the main goal. This decision is particularly hard where the status of the sentence is not linguistically marked. In that case, only inference on the argumentation in the article as a whole might help a human judge disambiguate, a possibility obviously not open to our system.

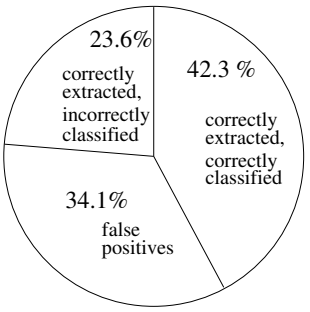

Figure 15: Overall results

Overall results for both tasks of the experiment (extraction and classification) are shown in Figure 15. At our high compression of 4.4%, 42.3% of all gold-standard sentences have been both correctly extracted and classified. This number includes the cases where one of several ambiguous roles has been identified correctly. A further 23.6% of the presented sentences can be counted as almost correct; they have been correctly extracted but have been assigned the wrong rhetorical role. 34.1% of all sentences are false positives, i.e. they should not have been extracted at all because the human annotator had not marked them.

Figure 16 shows a typical example of a rhetorically annotated extract. It is the output of our system after processing the first article in our collection, cmp_lg-9404003. Examples for correctly extracted and classified sentences are sentences 0 and 4, and also sentences 235, 236 and 238 (where one role was correctly identified). Correctly extracted, but incorrectly classified, are sentences 2 and 7. In our example, the only false positive is sentence 8.

The example also shows just how difficult rhetorical classification is. Consider sentence 7—a point could be made for the system's classification as well as for the human classification. Is "redefinition of synchronous TAG derivation" the Topic of the paper, or is it the Solution? Or is the Problem "How can synchronous TAG derivation be redefined?" One of these possibilities had to be chosen by objective criteria, which are documented in the coding manual for the annotation task.

4 Discussion

We find the results encouraging: with shallow processing, in a high-compression task, our algorithm finds 66% of all marked-up gold standard sentences in our training text and subsequently associates the right role for 64% of the correctly extracted sentences. Even though these results are only measurements of co-selection, they support our hypothesis that argumentative document structure can be approximated by low-level properties of the sentence. We see our prototype as a shallow document structure analyzer, specially designed for scientific text and geared towards the kinds of meta-

	MACHINE	HUMAN
0 The formalism of synchronous tree-adjoining grammars [REF], a variant of standard tree-adjoining grammars (TAG), was intended to allow the use of TAGs for language transduction in addition to language specification.	BACK	BACK
2 This paper concerns the formal definitions underlying synchronous tree-adjoining grammars.	SOLU	TOPI
4 This sort of rewriting definition of derivation is problematic for several reasons.	PROB	PROB
7 In this paper, we describe how synchronous TAG derivation can be redefined so as to eliminate these problems.	PROB	SOLU TOPIC
8 The redefinition relies on an independent redefinition of the notion of tree-adjoining derivation [REF] that was motivated completely independently of worries about the generative capacity of synchronous TAGs, but which happens to solve this problem in an elegant manner.	PROB	—
235 We have introduced a simple, natural definition of synchronous tree-adjoining derivation, based on isomorphisms between standard tree-adjoining derivations, that avoids the expressivity and implementability problems of the original rewriting definition.	SOLU	SOLU PROB
236 The decrease in expressivity, which would otherwise make the method unusable, is offset by the incorporation of an alternative definition of standard tree-adjoining derivation, previously proposed for completely separate reasons, that allows for multiple adjunctions at a single node in an elementary tree.	PROB	SOLU PROB
238 Nonetheless, some remaining problematic cases call for yet more flexibility in the definition; the isomorphism requirement may have to be relaxed.	SOLU	SOLU RWRK

Figure 16: Example of a rhetorically annotated extract, with gold standard judgement ("Human")

linguistic, argumentative constructs typically found in this text type.¡

However, our approach crucially depends on the quality of the indicator list. As our indicator list is handcrafted, (i.e. gained during the reading and annotation of the 123 papers in our training corpus), as opposed to automatically acquired, one might be suspicious of its performance—it might be over-fitted to the data, i.e. too dependent on phrases that occur only rarely rather than relying on generic phrases. As a result, it might not generalize well to other documents from the same source. The first question is thus how robust the indicator list is to different data of the same source.

In order to test the robustness of the list, we need *unseen* data, i.e. documents which were not taken into account when building the system or its knowledge sources, but for which gold standard judgements exist. As the process of annotation and indicator phrase addition happened simultaneously in our experiment, we do not have gold standards for the unseen part of our corpus. But we can simulate 'unseen' data as follows. We compare versions of our indicator phrase list before and after the annotation process for the last third of our training set (42 documents). Before the annotation process for that third, the indicator phrase list already contained 1501 indicator phrases; the annotation process for the last third only contributed another 262 phrases. When using the indicator list before the annotation process, the last third of the training data is practically treated as unseen: only indicator phrases are used that already occurred in the first two thirds of our training corpus. We report results only for the Indicator Features, because the performance of the other heuristics would not change by the analysis of more data. The results (Figure 17) show that there is only a minor decrease in performance if the first list is used (left column). This means that the indicator list, even though hand-crafted, is robust and general enough for our purposes; it generalizes reasonably well to texts of a similar kind, viz. research articles in computational linguistics of around 6 to 20 pages in length.

Another question is how well the list of phrases we collected might scale up to other domains. We make no claims about other *text types*, e.g. newspaper articles on scientific topics, or articles in *Scientific American*; our method depends on the explicitness of meta-linguistic information of scientific research articles which is not necessarily present in other text types.

We are interested in different domains, however, because we believe that the definition of rhetorical roles in our annotation scheme are generic rhetorical steps in scientific research papers. We are now planning to move to articles in the medical domain, in order to validate this hypothesis. With our corpus already consisting of articles from different sub-domains of computational linguistics, we are confident that performance should be similar in different domains as long as we have the right indicator phrases available. In the light of these considerations, the main challenge is to make the indicator

Extraction		
	Last 42 files treated as	
Heuristics used	seen	unseen
Indicator Quality Feature	57.62	54.32
Indicator Rhetorics Feature	47.76	44.48
Indicator Quality, Title, Sentence Length, and Header Type Features	68.36	64.78
Baseline	25.67	

Classification		
	Last 42 files treated as	
Heuristics used	seen	unseen
Indicator Quality Feature	50.21	49.36
Indicator Rhetorics Feature	56.47	55.79
Indicator Rhetorics, Relative Location and Title Features	61.37	60.26
Baseline	45.26	

Figure 17: Difference between seen and unseen data

features more adaptive to new text. What is needed is a method for the automatic and reliable acquisition of indicator phrases from corpus data, so that indicators get recognized even if the linguistic expression found is not identical, but only similar to one of the examples in the list.

We have run some preliminary experiments in indicator list acquisition. We used a simple method: using the gold standard sentences as a base, we compiled frequency lists of strings of different length occurring under each rhetorical role. Because subject matter specific strings get automatically cancelled out during this procedure, we ended up with a proto-list of around 500 very frequently occurring indicator phrases. In the extraction/classification experiment, this list performed about 30% below our hand-crafted list, a drop in performance which we believe to be mostly due to the fact that the new list is very short compared to the manually created list. On the positive side, the automatically created list is very unlikely to be over-fitted to our data. Further research could aim at improving this baseline by taking more sophisticated criteria like statistical interaction between the words in phrases into account, and by using different similarity measures to cluster similar phrases together.

In our approach, the rhetorically annotated extracts are collections of *sentences*. Although sentences are a natural choice of information unit when the collection of sentences is itself the abstract, there are several reasons why sentences are *not* the ideal information units for the approach we take. One problem is that as sentences are rhetorically connected to previous ones, they might not mean the same thing in isolation. They certainly don't look the same: Salager (1992), who analyzed summaries in the medical domain for the use of hedging and their rhetorical structure, found that in

summaries claims are stated boldly without explanations or comments, whereas in the full article a sentence conveying the same information tends to be formulated much more tentatively and with a higher level of reserve. Thus, it is unlikely that we will be able to use sentences extracted from the body of the text without change. The main problem is that sentences are too large a unit for rhetorical annotation and extraction, as became apparent during the human annotation phase: ideally, one would like to annotate and extract a unit that corresponds to a proposition, i.e. a clause. However, due to problems of ambiguity between sentential and phrasal coordination (and subordination), it is difficult to find clauses automatically with low-level tools like tokenizers. For now, we have to content ourselves with sentences as our selection unit for purely practical reasons. The sentence-based approach put forward here achieves good results, which might be improved later by a more sophisticated unit identification.

One of the main motivations behind our definition of rhetorical roles found in scientific articles is that this classification is intuitive to humans. This could be relevant for the procedure of how gold standards for training are gained. Typically, when human annotation is used to define gold standards for sentence extraction (Zechner, 1995; Marcu, 1997), the instructions to the annotators are vague and phrased in terms of importance ("annotate important sentences"). Due to the subjectivity and task dependence of the term 'important', such instructions usually result in individually varying annotations. If our claim that our annotation scheme defines relevance criteria in a more objective way is true, a definition of importance in terms of these rhetorical roles should make the task of annotating gold standards easier.

An experiment is currently underway to substantiate this claim. We have written a coding manual, i.e. an operational description of how the rhetorical roles are to be annotated, based on Swales' rhetorical moves, indicator phrases, and context. In the experiment, we compare the inter-annotator reliability of annotators who have read the annotation guidelines to that of two control groups: a second group who has been instructed to mark instances of the seven rhetorical roles without any further instructions, and a third group which had only been instructed to annotate important sentences. If our definitions of the rhetorical roles can be conveyed to other humans operationally, group 1 will have the highest inter-annotator reliability. If they are intuitive, group 2 will annotate similarly to group 1. Inter-annotator reliability should be higher in either group 1 or 2 than in group 3.

The usability of gold standards gained in an annotation based on our rhetorical roles will have to be established in an independent, task-based evaluation.

5 Related Work

Paice (1981) was probably the first attempt at implementing an extraction mechanism for physics articles that relied on pattern-matching operations, based on indicator phrases. Indicator phrases have been frequently used since then (Johnson et al., 1993). Paice and Jones (1993) made the method more flexible by supplying a finite state grammar for indicator phrases specific to the agriculture domain. However, we are the first to explicitly use the rhetorical status of indicator phrase for extraction and rhetorical classification.

There is a similar notion of *cue* phrases, typically used in discourse analysis, which is closely related to our notion of indicator phrases. Cohen (1987) defines cue words as all words and phrases used by the speaker to directly indicate the structure of the argument to the hearer. Cue phrases are typically short and come from a closed-class vocabulary (e.g. adverbials or sentence connectives (Litman, 1996)). As a result, the linguistic realization of the cue phrases between different authors tends to be invariant. Our indicator phrases, on the other hand, are longer and more variable; because they depend on the individual writing style, they are more difficult to identify automatically.

Rhetorical Structure Theory (RST) defines local rhetorical relations between sentences and clauses (Mann and Thompson, 1987), in order to build up a fixed rhetorically annotated tree structure through a complete rhetorical analysis of the text. There are automatic procedures for recognizing RST relations, either heuristically (Miike et al., 1994; Sumita et al., 1992) or by full rhetorical parse (Marcu, 1997) .

There are some analogies between these approaches and the analysis proposed in this paper, even though they are not obvious. We, too, believe that the main discourse structure of a paper is a hierarchical, rhetorically annotated tree structure. The branches are annotated differently, but one could argue that our rhetorical roles are text-type specific realizations of RST relations.

We believe that the upper parts of the tree are more important for abstracting than the lower level parts. Unlike RST, we are not concerned with rhetorical relations between each sentence or clause, but we concentrate on the higher levels of the tree, what we call *global* rhetorical relations: relations of content units with respect to the content of the whole article. We use indicator phrases which mark global rhetorical moves, rather than those that mark rhetorical relations between sentences or clauses.

As a result, we can perform a robust rhetorical analysis without the need for a *full* analysis. Our two-step approach ensures that we find global fillers for this flat tree structure with a reasonably high confidence level, at the cost of some detail in the lower areas of the tree. Indeed, the annotation scheme described in this chapter only allows us to build rhetorical trees which are one level deep.

Of course the representation of the text's structure

as a flat tree is a simplification. In principle, our annotation scheme could be extended to include these lower-level relations (e.g. the subproblem relationship between two problems), with intersegment relations holding at each level (e.g. the problem-solution relationship between a given problem and a solution when more than one problem is mentioned). This more detailed analysis may prove useful for the construction of longer and even more modular abstracts. But we believe that many of the local rhetorical relations between sentences and clauses are not immediately important for robust high-compression abstracting.

Our use of meta-linguistic information makes our approach different from methods which aim at representing the *contents* of the text. Lexical cohesion methods (Barzilay and Elahad, 1997), like statistical, keyword based methods, model main document concepts shallowly by using presumably content-specific lexical items observed in the text. Our method, in contrast, employs structural heuristics alone and uses *everything else* in the text but the content-specific lexical items.

Having said this, we can very well envisage our method cooperating with a complementary module that is based on an analysis of content rather than form. In a larger summarizing system, information from both types of module could flow together, in order to fulfill the tasks needed for generating abstracts from rhetorically annotated extracts: finding duplicate fillers, deciding on the best candidates for a filler, and resolving conflicts between fillers.

6 Conclusion

Robust, high-compression abstracting can be improved greatly if the discourse structure of the text is taken into account. We have argued that rhetorical classification of extracted material is a useful subtask for the production of a new kind of abstract that can be tailored in length and focus to users' expertise and specific information needs.

Our goal is to recognize abstract-worthy sentences with respect to global rhetorical structure, and to perform a subsequent classification of these sentences into a set of predefined rhetorical roles. We have presented a robust method which uses supervised learning techniques to deduce rhetorical roles from lower-level properties of sentences. This is technically feasible, because restrictions with respect to the task of the reader on the one hand, and knowledge about the typical argumentation of the writers on the other hand, can be exploited.

The results are encouraging; our algorithm determines 66% of all marked-up gold standards sentences in our training text and subsequently associates the right role for 64% of the correctly extracted sentences.

7 Acknowledgements

Data collection of our corpus took place collaboratively with Byron Georgantopolous. The first author is supported by an EPSRC studentship.

References

Ad Hoc Working Group For Critical Appraisal Of The Medical Literature. 1987. A proposal for more informative abstracts of clinical articles. *Annals of Internal Medicine.*

Alley, M. 1996. *The craft of scientific writing.* Englewood Cliffs, N.J.: Prentice-Hall.

American National Standards Institute, Inc. 1979. American national standard for writing abstracts. Technical report, American National Standards Institute, Inc., New York. ANSI Z39.14.1979.

Barzilay, R., and Elahad, M. 1997. Using lexical chains for text summarization. In Mani, I., and Maybury, M. T., eds., *Proceedings of the ACL/EACL-97 workshop on Intelligent Scalable Text Summarization.* Association for Computational Linguistics.

Baxendale, P. B. 1958. Man-made index for technical literature – an experiment. *IBM journal on research and development* 2(4):354–361.

Bazerman, C. 1988. *Shaping writing knowledge.* Madison: University of Wisconsin Press.

Borko, H., and Chatman, S. 1963. Criteria for acceptable abstracts: a survey of abstractors' instructions. *American Documentation* 14(2):149–160.

Broer, J. W. 1971. Abstracts in block diagram form. *IEEE Transactions on Engineering Writing and Speech* 14(2):64–67. ISA, 72-1626.

Buxton, A. B., and Meadows, A. J. 1978. Categorization of the information in experimental papers and their author abstracts. *Journal of Research in Communication Studies* 1:161–182.

Cohen, R. 1987. Analyzing the structure of argumentative discourse. *Computational Linguistics* 13:11–24.

Cremmins, E. T. 1996. *The art of abstracting.* Information Resources Press.

Day, R. A. 1995. *How to write and publish a scientific paper.* Cambridge: Cambridge University Press.

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

Hartley, J., and Sydes, M. 1997. Are structured abstracts easier to read than traditional ones? *Journal of Research in Reading* 20(2):122–136.

Hartley, J.; Sydes, M.; and Blurton, A. 1996. Obtaining information accurately and quickly: are structured abstracts more efficient? *Journal of Information Science* 22(5):349–356.

International Organisation for Standardisation. 1976. Documentation – Abstracts for publication and documentation. Technical report, International Organisation for Standardisation. ISO 214-1976.

Johnson, F. C.; Paice, C. D.; Black, W. J.; and Neal, A. P. 1993. The application of linguistic processing to automatic abstract generation. *Journal of Document and Text Management* 1(3):215–42.

Kintsch, W., and van Dijk, T. A. 1978. Toward a model of text comprehension and production. *Psychological Review* 85(5):363–394.

Kircz, J. G. 1991. The rhetorical structure of scientific articles: the case for argumentational analysis in information retrieval. *Journal of Documentation* 47(4):354–372.

Kupiec, J.; Pedersen, J. O.; and Chen, F. 1995. A trainable document summarizer. In *Proceedings of the 18th ACM-SIGIR Conference*, 68–73.

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

Litman, D. J. 1996. Cue phrase classification using machine learning. *Journal of Artificial Intelligence Research* 5:53–94.

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

Mann, W. C., and Thompson, S. A. 1987. Rhetorical structure theory: A theory of text organisation. Technical report, Information Sciences Institute, U of South California. ISI/RS-87-190.

Manning, A. 1990. Abstracts in relation to larger and smaller discourse structures. *Journal of Technical Writing and Communication* 20(4):369–390.

Marcu, D. 1997. *The rhetorical parsing, summarization, and generation of natural language texts.* Ph.D. Dissertation, University of Toronto.

Miike, S.; Itoh, E.; Ono, K.; and Sumita, K. 1994. A full-text retrieval system with a dynamic abstract generation function. In *Proceedings of the 17th ACM-SIGIR Conference, Association for Computing Machinery, Special Interest Group Information Retrieval*, 152–163.

Milas-Bracovic, M. 1987. The structure of scientific papers and their author abstracts. *Informatologia Yugoslavica* 19(1-2):51–67.

O'Hara, K., and Sellen, A. 1997. A comparison of reading paper and on-line documents. In *Proceedings of CHI-97, Special Interest Group on Computer & Human Interaction.*

Paice, C. D., and Jones, A. P. 1993. The identification of important concepts in highly structured technical papers. In *Proceedings of the Sixteenth Annual International ACM-SIGIR conference on research and development in IR, Association for Computing Machinery, Special Interest Group Information Retrieval.*

Paice, C. D. 1981. The automatic generation of literary abstracts: an approach based on the identification of self-indicating phrases. In Oddy, R. N.; Robertson, S. E.; van Rijsbergen, C. J.; and Williams, P. W., eds., *Information Retrieval Research.* London: Butterworth. 172–191.

Pinelli, T. E.; Cordle, V. M.; and Vondran, R. F. 1984. The function of report components in the screening and reading of technical reports. *Journal of Technical Writing and Communication* 14(2):87–94.

Rennie, D., and Glass, R. M. 1991. Structuring abstracts to make them more informative. *Journal of the American Medical Association* 266(1).

Rowley, J. 1982. *Abstracting and indexing.* London: Bingley.

Salager-Meyer, F. 1992. A text-type and move analysis study of verb tense and modality distributions in medical English abstracts. *English for Specific Purposes* 11:93–113.

Salton, G., and McGill, M. J. 1983. *Introduction to Modern Information Retrieval.* Tokyo: McGraw-Hill.

Salton, G.; Allan, J.; Buckley, C.; and Singhal, A. 1994. Automatic analysis, theme generation, and summarisation of machine readable texts. *Science* 264:1421–1426.

Skorochod'ko, E. F. 1972. Adaptive method of automatic abstracting and indexing. In *Information Processing 71*, volume 2. North Holland Publishing company. 1179–1182.

Sumita, K.; Ono, K.; Chino, T.; Ukita, T.; and Amaro, S. 1992. A discourse structure analyzer for japanese text. In *Proceedings of the International Conference on Fifth Generation Computer Systems.*

Swales, J. 1981. Aspects of article introductions. Aston ESP Research Project No. 1. Technical report, The University of Aston, Birmingham, U.K.

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

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