Sentence extraction and rhetorical classification for flexible abstracts

Simone Teufel
Centre for Cognitive Science
and Language Technology Group
University of Edinburgh
S.Teufel@ed.ac.uk

Marc Moens
Language Technology Group
University of Edinburgh
M.Moens@ed.ac.uk

Abstract

Knowledge about the discourse-level structure of a scientific article is useful for flexible and sub-domain independent automatic abstraction. We are interested in the automatic identification of content units ("argumentative entities") in the source text, such as GOAL OR PROBLEM STATEMENT, CONCLUSIONS and RESULTS. In this paper, we present an extension of Kupiec et al.'s methodology for trainable statistical sentence extraction (1995). Our extension additionally classifies the extracted sentences according to their argumentative status; because only low-level properties of the sentence are taken into account and no external knowledge sources other than meta-level linguistic ones are used, it achieves robustness and partial domain-independence.

Introduction

Motivation

Until recently, the world of research publications was heavily paper-oriented. One of the roles of abstracts of research articles was to act as a decision tool: on the basis of the abstract a researcher could decide whether the paper was worth a visit to the library, whether it was worth a letter to the author requesting a copy of the full paper, whether it was worth postponing finishing one's own paper, etc.

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. These guidelines suggest that such an abstract should be aimed at the partially informed reader (Kircz 1991): someone who knows enough about the field to understand basic methodology and general goals 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. 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 was a pointer to an article not immediately available, the abstract had 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 summarisation presupposes that the full article is available in machine-readable form. A typical scenario will be one where a researcher receives a large quantity of machine-readable articles, for example in reply to a search query. As a result, abstracts will play a different role. They can still be used as a decision tool, to decide which articles to look at first. In addition, abstracts can be used as a *navigation* tool, helping users find their way through the retrieved document collection. Abstracts could indicate which of the retrieved articles share the same research questions or methodologies, or they could show how articles are related to other articles (in logical or chronological respect); such abstracts can be used to refine the original query.

Abstracts can also help with the nonlinear reading of textual material–the process whereby readers 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 serves to build a mental model of the text's structure as well as to extract the main concepts of the paper; it is a powerful and effective reading strategy which is typical for scientists (Bazerman 1988). However, (O'Hara & 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. As a result, users are forced to read electronic articles more linearly than they would paper articles.

Abstracts can facilitate this process of nonlinear

reading by revealing the text's logical and semantic organisation. Their third main function is to support navigation *within* a scientific article.

Note that in this scenario the abstracts needed to support any of these three functions can be very different from paper-based abstracts. When flexible and expandable abstracts serve as navigation (non-linear-reading) tools, the relevance decision phase is intertwined with a first text skimming phase. Thus, the traditional distinction between informative and indicative abstracts becomes less meaningful. Abstracts need no longer be self-contained since the full text can be made available at the same time, possibly alongside the abstract. Nor need the abstracts be targeted just at an idealised partially informed reader; instead, the abstract can be tailored more to the actual user, for example taking into account keywords in the user's original query.

By necessity, such abstracts will have to be generated automatically and on the fly. Even though they will be of lower quality when compared to human-crafted abstracts, we predict that they will be more useful in many situations. It is the flexible automatic generation of such abstracts that we see as our long-term goal.

General approach

We take sentence extraction as our starting point because this method is robust towards text type and author style. Sentence extracts are useful for relevance indication but not for navigation because they lack *rhetorical* information: information about the rhetorical justification for extracting a certain sentence. For example the sentence could have been rhetorically justified because it described the purpose of the research, or the conclusion. In our view, abstracting means analyzing the argumentative structure of the source text and identifying textual extracts which constitute representatives for the given argumentative step. The rhetorically annotated extracts could be subjected to further analysis before an abstract is deep-generated, or alternatively, it could be used without much further work (except anaphora resolution): the actual abstract could be construed as a template, where slots correspond to an argumentative step which are filled with rhetorically justified sentences (cf. the "structured abstracts" which have become prevalent in the medical domain in the past decade (Adhoc 1987, Hartley & Snydes 1997)). The argumentative template of the abstract will allow variations in length, and certain argumentative information can be added or suppressed: for example, the amount of BACKGROUND information can be varied depending on the experience of the user.

Two questions then arise. A first question is how the building blocks of the abstract template, i.e. the argumentative roles, should be defined. This is a particular problem for our approach because very little is known about what our new type of abstracts should look like. Most of the information on good abstracts deal with the world of paper, not with the use of online 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, as discussed in the following section.

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

Argumentative structure of research articles

Scholarly articles serve the process of communicating scientific information by means of documents; their communicative function is to present, retell and refer to the results of specific research (Salager-Meyer 1992). Discourse linguistic theory suggests that texts that serve a common purpose among a community of users eventually take on a predictable structure of presentation. In scientific articles, prototypical rhetorical divisions have been established, typically Introduction, Purpose, Experimental Design, Results, Discussion, Conclusions, etc. This is especially the case for research texts in the exact sciences where rhetorical divisions tend to be very clearly marked in section headers. In non-experimental articles, the rhetorical divisions are still there, but implicitly so (Liddy 1991).

But the articles we are dealing with describe their research in a more idiosyncratic, less rigid and less well-structured way. Our corpus consists of articles in computational linguistics and cognitive science. Due to the fact that this research field is relatively young and interdisciplinary in approach, we found a heterogeneous mixture of methodologies and traditions of presentation. Our papers come from different sub-disciplines of cognitive science: some are from the field of theoretical linguistics, some are very technical, describing implementations, others are experimental. Even though most of them had an introduction and conclusions, and cited previous work, the presentation of the problem and the methodology/solution was completely idiosyncratic, and most headings are not prototypical

but subject matter specific.

So we were looking for a description of the argumentative structure at an abstraction level which is general enough to be sub-domain independent. Although we argued in the previous section that most 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 desirable argumentative structure in abstracts.

Most authors of prescriptive literature ("how should abstracts be constructed") agree with the ANSI recommendations (American National Standards Institute, Inc. 1979) that informative abstracts should mention the PURPOSE or PROBLEM of the full article, SCOPE or METHODOLOGY, RESULTS, and CONCLUSIONS or RECOMMENDATIONS (International Organisation for Standardisation 1976; Rowley 1982; Cremmins 1996), cf. the early, comprehensive survey by (Borko & Chatman 1963).

There is more disagreement about "peripheral" content units, such as RELATED WORK, BACKGROUND, INCIDENTAL FINDINGS, FUTURE WORK and DATA. Some authors (Rowley 1982; Cremmins 1996) recommend not to include any background information at all, but we believe that background information is potentially important, especially for self-contained abstracts and for abstracts for novice readers. Similarly, (Cremmins 1996) states that the content unit PREVIOUS WORK should not be included in an abstract unless the studies are replications or evaluations of the earlier work. However, depending on the information need, previous work might actually have been central to the query the user started off with, and we therefore want to preserve the possibility of including it in our modular abstracts.

The biggest problem we encountered when we tried to reuse existing argumentative taxonomies found in the literature was that many of the suggestions in the literature are by far too domain-specific for our purposes. Liddy's description of the components of abstracts (1991) is based on professional abstractors' intuitions and a corpus of abstracts, and it is very specific to the domain of empirical abstracts.[1] Kircz's taxonomy of argumentative entities (1991) includes dependencies between these entities, but it is also highly domain-specific for physics articles, and very fine-grained.

BACKGROUND	BACK
TOPIC/ABOUTNESS	TOPI
RELATED WORK	RWRK
PURPOSE/PROBLEM	PU/PR
SOLUTION/METHOD	SOLU
RESULT	RESU
CONCLUSION/CLAIM	CO/CL

Table 1: Top-level argumentative units in our taxonomy

A range of researchers have described the actual composition of human-written abstracts with respect to rhetorical units. (Buxton & Meadows 1978) provide a comparative survey about the contents of abstracts in the physics domain, (Salager-Meyer 1992) for medical abstracts and (Milas-Bracovic 1987) for sociological and humanities abstracts. Their taxonomies all resemble the ANSI one.

In trying to find the right level of granularity for the description, we found that the building-plan of a scientific article had to be seen as an instance of the generic problem-solving paradigm (similar to AI research on plan recognition). Argumentation follows the chronological course of the actual research only in exceptional cases; instead, it is explicitly or implicitly organized according to the logical steps of the problem-solving process (and its presentation).[2] Our taxonomy is an adaption of the ANSI suggestions; we have added 3 peripheral roles which we hope to be useful for further processing, namely BACKGROUND, TOPIC and RELATED WORK. Table 1 shows the 7 top-level argumentative units of our taxonomy.[3]

The content units in Table 1 are the high level argumentation steps of a problem-solving process; the real argumentative moves found in an article are instances of these moves. For example, mentioning a weakness of a previous approach is tantamount to stating the research goal of the given article, and there are more stylistic alternatives: as a simple purpose clause (*in*

[1] There are also technical reasons why we could not adopt her suggestions: Liddy defines an abstract's constituent components in a recursive fashion (i.e. they can be contained within other components), and most of them span parts of sentences rather than whole sentences. Neither of these options are available with our machine-learning technology, and so her suggestions could not be taken on board without significant change.

[2] In a scientific paper, the act of researching and the act of reporting on the research are intertwined. We distinguish between plans/argumentation steps that describe research phases (e.g. *I first classified the phonemes*), and textual steps (e.g. *in chapter 3, we will present our results*). We intend to filter such textual information in a separate step, in order to provide additional information for the information searching process. They are not included in this taxonomy.

[3] We also developed a feature-value-based subdivision of these roles, which we will not discuss here because these distinctions are too fine to be learnt automatically yet.

order to automatically...), as an explicit problem statement (*our goal is to...*) or implicitly as a mention of the absence of a solution *however, to our knowledge it has never been shown that...*). The underlying move is the same in all these realizations, namely *motivate that the research addresses a real problem; motivate that the problem is worth solving*; classified as PURPOSE/PROBLEM in our approach. Not all of these content units need to be realized in an article, but the ones that are must be in a logical order so that readers can grasp the argumentation and motivation behind the single research and report steps. It does not make sense to present a solution for which no corresponding problem has been posed; each problem mentioned should either find a solution within the article or receive the status "yet to be solved" in the FURTHER RESEARCH section; the sub-problem-solution space should be logically consistent, to name just a few of the obvious constraints.

Linguistic realization of argumentative structure: indicator phrases

We are particularly interested in the linguistic realisations of argumentative acts as low-level features which can be captured in sentence extraction. There has been research into which low-level linguistic factors correlate to different structures. Grosz and Sidner's intentional structure (Grosz & Sidner 1986) can be derived from lower-level linguistic markers, i.e. anaphora phenomena. Also, aspect and tense has been shown to correlate with discourse structures (Salager-Meyer 1992; Hwang & Schubert 1992). However, our main focus of interest here are explicit constructs, which (Paice 1981) called "indicator phrases".

(Cohen 1987) presents a theoretical framework for general argumentation, including pragmatic analysis and based on claim-evidence trees. Even though her aims are broader than ours and her framework is too general for our purposes, we find it interesting that she considers clue interpretation as "not only quite useful but feasible" and suggests the implementation of a separate clue module within her framework. However, there is an important difference between her clue words and our indicator phrases. Cohen defines clue words as all words and phrases used by the speaker to directly indicate the structure of the argument to the hearer (e.g. connectives). RST (Mann & Thompson 1987) uses a similar notion to mark local rhetorical relations between sentences and clauses. Our definition of indicator phrases is more restrictive: They are the phrases who directly indicate *global* argumentative moves between argumentative units, rather than between sentences or clauses. As such, they express relations to the rhetorical act constituted by the whole paper. The kind of discourse structure we envisage is thus much less detailed than RST or Cohen would suggest.

Of course, indicator phrases need not be unambiguous with respect to their argumentative status. For example, the phrase *in this paper, we have* is a very good 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...*). This motivates a separation of the tasks of a) extracting good sentence candidates and b) classifying them into their argumentative status, possibly with different indicator lists for the two tasks.

Authors use idiosyncratic style, especially in an unmoderated medium like the one where our articles come from. But we believe that automatically tracking the global-level argumentative structure of an article is possible, at least to a degree needed for the generation of more well-structured abstracts, by using shallow processing of the texts. Meta-linguistic information being the common denominator across articles, such a model should also be at least partially sub-domain independent.

Previous work in sentence selection

Sentence selection is an abstraction over different measurements of a sentence's importance (a high level property) from low-level, objectively determinable properties. Over the years there have been many suggestions as to which features contribute to making a sentence "meaningful" or abstract-worthy, in particular stochastic measurements for the significance of key words (Luhn 1958), its location in the source text (Baxendale 1958; Edmundson 1969), the presence of cue or indicator phrases (Paice 1981), or of title words (Edmundson 1969). The problem is that none of these features by themselves suffice, and weighted combinations need to be found.

(Kupiec, Pedersen, & Chen 1995) use supervised learning to automatically adjust feature weights, using a corpus of research articles and corresponding summaries. Kupiec et al.'s gold standard of abstract-worthy sentences is defined as the set of sentences in the source text that "align" with a sentence in the abstract—i.e. sentences that show sufficient semantic and syntactic similarity with a sentence in the abstract.

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 abstract. In Kupiec et al.'s corpus of 188 engineering articles with summaries written by professional abstractors, 79% of sentences in the abstract 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 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.

Evaluation of the training relies on cross-validation: the model is trained on a training set of documents, leaving one document out at a time (the current test document). The model is then used to extract candidate sentences from the test document, allowing evaluation of precision (sentences selected correctly over total number of sentences selected) and recall (sentences selected correctly over 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. Kupiec et al. report that precision of the individual heuristics ranges between 20–33%; the highest cumulative result (44%) was achieved using paragraph, fixed phrases and length cut-off features.

Our experiment

We borrow our methodology from Kupiec et al., but our orientation on meta-linguistic information makes our approach different from methods which extract sentences based on heuristics about the *contents* of sentences (e.g. using a tf/idf[4] model or lexical cohesion) and linking these together into an abstract.

We decided to split the task because we found that our heuristics were differently useful for the different tasks, so a two-step process avoids distortion of the training from these heuristics. The basic procedure for the sentence extraction and classification experiment is:

Step one: Extraction of abstract-worthy sentences. We try to identify sentences which carry *any* rhetorical roles (as defined by our annotation scheme) from irrelevant sentences (by far the larger part of the text). Failure to perform this task leads to the inclusion of irrelevant material in the abstracts (false positives), or the exclusion of relevant material from the abstract (false negatives). We call the result of this step an intermediate extract.

Step two: Identification of the correct rhetorical role. Once good abstract sentence candidates have been identified, we try to classify them according to their rhetorical role. We expect to do less well on this task, because the classification is inherently vague and ambiguous between certain classes (confusion classes). We call the collection of sentences with their argumentative status, an "argumentatively annotated extract".

Data and annotation of gold standards

Our corpus is a collection of 201 articles and their summaries from different areas of computational linguistics and cognitive science, most of them conference articles. The corpus contains 568,000 word tokens.[5] The following structural information is marked up: title, summary, headings, paragraph structure and sentences. Tables, equations, figures, captions, references and cross references were removed and replaced by place holders.

We assume that most of the articles had been accepted for publication, although this cannot be relied on as the archive is unmoderated. Although all the articles in this collection deal with computational linguistics, the corpus displays huge variation as to subdomain. The largest part (about 45%) are articles describing implementational work, but there are about 25% theoretical-linguistic articles, with an argumentative tenet, about 10% overview and general-opinion articles and 20% experimental articles (reporting corpus studies or psycholinguistic experiments). As a result of this, there is no explicit homogeneous discourse structure. Also, the writing style varies from extremely informal to formal. About a third of the articles were

[4] tf/idf or term-frequency times inverse document frequency is a method of document specific keyword weighing, which is commonly used in Information Retrieval (Salton & McGill 1983).

[5] The corpus was drawn from the computation and language archive (`http://xxx.lanl.gov/cmp-lg`), converted from LaTeX source into HTML in order to extract raw text and minimal structure automatically, then transformed into SGML format with a `perl` script, 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 were no less than 4 and no more than 10 pages length in Postscript originals, resulting in the above mentioned 201 documents. Data collection took place collaboratively with Byron Georgantopolous.

not written (or subsequently edited) by native speakers of English.

Summaries of the articles in our collection were provided by the authors themselves. In general, such summaries are of a lower quality (or at least less systematically constructed) than summaries by professional abstractors. After having had a close look at the author abstracts, our hopes were low that we would be able to use them directly for evaluation. Some abstracts are extremely short, and many of the abstracts are not self-contained, and would thus be difficult to understand for the partially informed reader. As an informal test to see if we could identify overall properties of the discourse level structure in the summaries, we applied our annotation scheme for rhetorical roles to the 123 summaries in our training and test corpus. We found that abstract structure varied widely. The authors, in contrast to professional abstractors surveyed by Liddy for example, had not used a prototypical scheme to write their abstracts. Even though most abstracts are still understandable and many are well-written, we hope that our method will create abstracts with a more systematic structure.

We divided our corpus into a training and test set of 123 documents which were further annotated for evaluation and training, and a remaining set of 78 documents which were unseen and were not used for the experiments described here. Annotation of the training and test set proceeded in two steps:

1. Determine which sentences are relevant/carry any high-level argumentative information;
2. determine the argumentative status of these sentences.

The creation of gold standards for the first step is described in (Teufel & Moens 1997). Similar to Kupiec et al.'s experiment, alignment between summary and document sentences was decided in a semi-automatic manner, and final alignment was decided by a human judge. The criterion was similarity of semantic contents of the compared sentences. As predicted, we found a substantially lower level of alignable sentences (31%[6]) in comparison to Kupiec et al.'s value of 79%. Because of the low alignment, we annotated source texts with additional abstract-worthy sentences, as selected by a human judge. We thus had two different gold standards: gold standard A is gained by alignment (265 sentences or 28%), gold standard B by human judgement (683 sentences or 72%).

[6] The alignment rate of 31% refers to all 201 documents; alignment rate in our training and test set of 123 documents, which consists of the best-aligned documents, is 52%.

With respect to compression (ratio of sentences in target extract to sentences in document), our combined gold standards achieve 4.4% as compared to Kupiec et al.'s 3.0% compression. Gold standard A meant 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.

What we were trying to annotate (and subsequently automatically learn) is a high-level property, viz. "which rhetorical role, if any, is expressed by the following sentence?" This can be a difficult question, but we found that humans find it easier to answer that question for a given sentence than to answer the related question "is this sentence a good candidate for inclusion in an abstract?" (irrespective of which role it has).

The task to decide on a certain role is nevertheless not easy. Often, the rhetorical role of a statement is dependent on the local context of the line of argument. For example, if the authors mention a weakness of their solution, it might be classified as SOLUTION[LIMIT] or as PURPOSE/PROBLEM[LOCAL], depending on whether that problem will be solved later on in the given article. Or, if there is a mention in the discussion that a certain problem does not occur with the presented solution, this might be viewed as a description of a tackled problem or as an advantage of the solution.

The following sentence with its judgements illustrates the 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 that any lexicalized grammar that encodes this language. CONCLUSION/CLAIM

We allowed for multiple annotation in ambiguous cases, but still faced problems, most of them having to do with the large unit of annotation (a whole sentence as opposed to a clause or even smaller unit) as enforced by our annotation and machine-learning technology. The following sentence shows a case where the sentence covers more than one role (AND denotes concatenations of roles within one sentence).

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

Figure 1 shows the composition of 1172 instances of rhetorical roles for the 948 gold standard sentences in our training set. 232 sentences (24%) contained ambiguous mark-up.

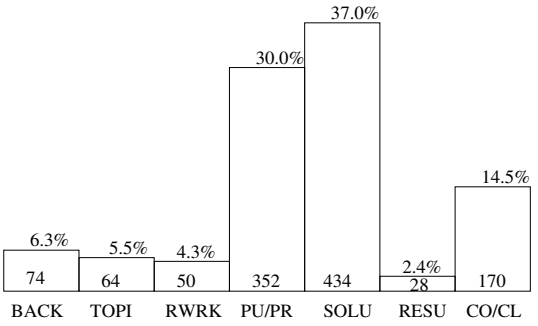

Figure 1: Composition of rhetorical roles for training set

Heuristics

We employed 6 different heuristics: 4 of the methods used by Kupiec et al. (indicator phrase method, location method, sentence length method and thematic word method), and 2 additional ones (title method and header method). We use different versions of methods for the two steps; typically, the methods used for the rhetorical classification are refinements of the methods used for extraction.

Indicator phrase quality method: The indicator phrase method uses linguistic text properties to identify meta-discourse (as opposed to subject matter) in a text. We use a list consisting of 1728 indicator phrases or formulaic expressions, like communicative verbs and research and argumentation related phrases. The largest part of these phrases is positive.

Our indicator phrase list was manually created by a cycle of inspection of extracted sentences and addition of indicator phrases to the list. Indicator phrases were manually classified into 5 quality classes according to their occurrence frequencies within the gold standard sentences. Thus, 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 unlikely' to 'very likely to be included in a summary'. For example, the phrase *we have given an account* received a high score of +3, whereas *supported by grant* receives a negative score. This method proved useful for the extraction step.

Indicator phrase identity method: For the classification step, however, some extra information is needed. Thus, we trained from the corpus by frequency counts *which* individual indicator phrases were associated with which roles how often, and how ambiguous this mapping was. Because some phrases occur very rarely and would thus arbitrarily bias the method too much to one or another role, we smoothed these frequency counts with groupings of indicator phrases. Groupings consist of indicator phrases which are syntactically and semantically similar, typically centred around one lexical item, e.g. *show*. They are manually created, and we currently have 194 groups. The probability of a certain role being associated with an indicator phrase is the means of the probability of the *phrase itself* having occurred with training examples of that role, and of the probability of the *group* being associated with the role (the means of the probabilities of that role for all members of the group).

Location method: This feature distinguishes peripheral sentences in the document and within each paragraph, assuming a hierarchical organisation of documents and paragraphs. The algorithm is sensitive to prototypical headings (*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. For the classification task, we work with a variant of the location method that separates the document in 10 proportional regions.

Sentence length method: 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 method: This method is a variation of the tf/idf method, which 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 word in sentence, meaned by sentence length. The 40 top-rated sentences obtain score 1, all others 0.

Title method: Words occurring in the title are good candidates for document specific concepts. The title method 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 method: For the rhetorical classification, the rhetorical section of a sentence can be a good indication. Therefore, we remembered for each sentence if its corresponding header was one of the prototypical ones, or if its header was an untypical one (all others, normally containing subject matter information). We clustered morphological variants of each other into one class, resulting in 15 classes.

Heuristics used in the first step are indicator quality,

sentence length, location, title and tf/idf; in the second step we use indicator identity, fine location and header method.

Classifiers

Kupiec et al.'s estimation for the probability that a given sentence is contained in the abstract 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, we similarly estimate the probability $P(s \in R_m|s \in E, F_1,\ldots,F_k)$, i.e. that sentence s in the source text carries role R_m, given its feature values (and given that it was included in the extract, which is the universe of the second classification).

Results

In the following, we present success rates for the two kinds of tasks. Numerical values in the tables always give precision and recall rates. For the first task (extraction), evaluation is based on cross-validation like in Kupiec et al.'s experiment.[7]

For the second task (classification), we report numerical values for those sentences that have been correctly identified by the first step. Precision reports for each role the proportion of correctly identified instances against those sentences identified as the given role; recall reports for each role the proportion of correctly identified instances against those sentences identified in our gold standards as carrying that role. In the case of ambiguity between several roles in the gold standard, the evaluation is harsh on our algorithm: our algorithm, which can only identify one role, will receive a share of a point, if the role it identified is included in the roles in the gold standard, depending on how many ambiguous roles there were. In the case of ambiguity, it is therefore theoretically not possible to score 100% for an algorithm like ours which only identifies one role per sentence.

Extraction Table 2 summarises the contribution of the individual methods, individually and cumulatively. Using the indicator phrase method (method 1) is clearly the strongest single heuristic. Overall, these results reconfirm the usefulness of Kupiec et al.'s method of heuristic combination. The method increases precision for the best method by around 20%.[8]

	Indiv.	**Cumul.**
Method 1 (indicator)	55.2	55.2
Method 2 (location)	32.1	65.3
Method 3 (length)	28.9	66.3
Method 4 (tf/idf)	17.1	66.5
Method 5 (title)	21.7	68.4
Baseline	28.0	

Table 2: Impact of individual heuristics

Classification Tables 3 and 4 show the confusion matrix and precision and recall values for each rhetorical role for 2 combinations of heuristics: header, location and indicator identity or indicator identity alone. The columns refer to the roles assigned by our algorithm; the lines denote roles assigned in the gold standard ("reality"). Successful classifications are shown in boxes. In this task, combination of heuristics also results in a *decrease* of precision and recall. The indicator phrase method proves to be highly predictive; again, it is the strongest method, as expected. The overall precision and recall, solely based on the disambiguation power of the single cue phrases, is quite remarkable at 68%. When less predictive, location-based methods are mixed in, precision and recall improve slightly.

Taken together, the extraction and classification success rate of the method (precision and recall) lies by 46.9%, at a compression rate of around 4%. It is worth pointing out that such high-compression abstracting is a more useful and more difficult task than lower-compression abstracting as usually reported in the literature (with abstracts of only 25% compression). That means, that out of a long document (average: 203 sentences), the algorithm picks 4% sentences and manages to assign about half of them the correct rhetorical status, as defined in our annotation scheme.

[7] As a baseline we chose sentences from the beginning of the source text, which obtained 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.

[8] For a more comprehensive discussion of the results, cf. (Teufel & Moens 1997).

		CLASSIFIED AS							
		BACK	TOPI	RWRK	SOLU	PUPR	RESU	CLCO	total
R	BACK	6.50			22.50	5.00			34.00
E	TOPI		18.33		6.33	22.00			46.67
A	RWRK			3.00	11.00	5.50			19.50
L	SOLU	0.50	1.00		147.00	56.50	3.33	5.50	213.83
I	PUPR		2.00	1.00	47.00	135.00		18.00	203.00
T	RESU				1.00	1.00	9.50	2.50	14.00
Y	CLCO				14.00	18.50		80.50	113.00
	total	7.00	21.33	4.00	248.83	243.50	12.83	106.50	644.00
	precision	92.86	85.94	75.00	59.08	55.44	74.05	75.58	62.09
	recall	19.21	42.65	15.38	68.74	66.59	67.86	71.24	62.09

Table 3: Confusion matrix: classification results using only indicators

		CLASSIFIED AS							
		BACK	TOPI	RWRK	SOLU	PUPR	RESU	CLCO	total
R	BACK	7.00			27.00				34.00
E	TOPI		21.50	0.50	7.50	16.83		0.33	46.67
A	RWRK			8.50	7.50	3.00		0.50	19.50
L	SOLU				154.00	48.50	0.50	10.83	213.83
I	PUPR		1.50		31.50	160.00		10.00	203.00
T	RESU				2.50	1.00	6.50	4.00	14.00
Y	CLCO				13.00	15.67		84.33	113.00
	total	7.00	23.00	9.00	243.00	245.00	7.00	110.00	644.00
	precision	100.00	93.48	94.44	63.37	65.31	92.86	76.67	68.61
	recall	20.59	46.07	43.59	72.02	78.82	46.43	74.63	68.61

Table 4: Confusion matrix: classification results using indicators, location and headers

Discussion

It is questionable if our argumentatively annotated extracts are stable enough to be presented to users without further work, especially considering that some roles do not get chosen often enough to fill an abstract slot with them. However, subsequent information extraction (and possibly reasoning about argumentation steps) could correct suboptimal performance of both steps. The algorithm gives us an set of abstract-worthy sentences with a relatively high reliability, along with an indication as to what their most probable argumetative status is. These candidate sentences are the ones that are worth further, deeper, more resource intensive analysis, especially if the rhetorical role of that sentence is needed. In the case of ambiguities, other modules in the abstracting process could decide how indispensable a given role is, and if the ambiguity needs to be resolved. Obviously, how much a given role is needed depends on the structure of the abstract frame and alternative information resources, e.g. from textual cues (like *in chapter 4, we will define our goal in more detail*) or other extraction results.

We find the results encouraging: with shallow processing only, our algorithm determines 68% of all marked-up gold standards sentences in our training text and subsequently associates the right role for again 68% of the correctly extracted sentences (i.e. 46% in toto). However, we have to keep in mind that we are dealing with seen data and that our indicator list has been manually created. Our ambition now is to make the indicator method more adaptive to new text of a similar kind; this will have to be done in a learning phase. We are experimenting with maximum entropy methods for determining indicator phrases and possible groupings between them automatically. Our hope is that this work will reconfirm our hypothesis that there is enough overlap in the linguistic realizations of rhetorical roles to keep the classification stable across sub-domains.

The described experiments are a first step in automatic rhetorical classification, and even though the results are far from perfect, they support our hypothesis that argumentative document structure can be approximated by low-level properties of the sentence. With respect to our general approach, we have consciously chosen to use rhetoric structure of paper as op-

posed to attempting to model any information about the subject-domain. This is quite different from AI-based knowledge-intensive approaches (semantic analysis) or other, more shallow methods for modelling subject matter (e.g. lexical chains or the tf/idf method). We argue that the domain of argumentation is small enough to be modelled and expressive enough to provide a classification which is useful for subsequent steps.

Conclusion

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 proportion to users' expertise and specific information needs. Our goal is to recognize abstract-worthy sentences with respect to rhetorical structure, and to perform a classification of these sentences into a set of predefined, generic rhetorical roles. We have presented a robust method which uses statistical classification to deduce these rhetorical roles from lower-level properties of sentences.

The results are encouraging; our algorithms determines two out of three marked-up gold standards sentences in our training text and additionally associates the right role for about 46% of all sentences it extracts. Even though this level of precision in the classification is not reliable enough to use the extracts without further processing, our results seem to point to the general feasibility of a shallow processing of discourse structure.

References

American National Standards Institute, Inc. 1979. *American National Standard for Writing Abstracts.* New York: American National Standards Institute, Inc. ANSI Z39.14.1979.

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.

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.

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

Grosz, B. J., and Sidner, C. L. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics* 12(3):175–204.

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

Hwang, C. H., and Schubert, L. K. 1992. Tense trees as the 'fine structure' of discourse. In *Proceedings of ACL-92*, 232–240.

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

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.

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.

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.*

Paice, C. D. 1981. The automatic generation of literary abstracts: an approach based on the identification of self-indicating phrases. In Norman, O. R.; Robertson, S. E.; van Rijsbergen, C. J.; and Williams, P. W., eds., *Information Retrieval Research.* London: Butterworth. 112.

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.

Teufel, S., and Moens, M. 1997. Sentence extraction as a classification task. In Mani, I., and Maybury, M. T., eds., *Proceedings of the workshop on Intelligent Scalable Text Summarization, in association with ACL/EACL-97.*