Discourse-level argumentation in scientific articles: human and automatic annotation

Simone Teufel and Marc Moens
HCRC Language Technology Group
Division of Informatics
University of Edinburgh
S.Teufel@ed.ac.uk, M.Moens@ed.ac.uk

Abstract

In this paper we present a rhetorically defined annotation scheme which is part of our corpus-based method for the summarisation of scientific articles. The annotation scheme consists of seven non-hierarchical labels which model prototypical academic argumentation and expected intentional 'moves'. In a large-scale experiments with three expert coders, we found the scheme stable and reproducible. We have built a resource consisting of 80 papers annotated by the scheme, and we show that this kind of resource can be used to train a system to automate the annotation work.

1 Introduction

Work on summarisation has suffered from a lack of appropriately annotated corpora that can be used for building, training and evaluating summarisation systems. Typically, corpus work in this area has taken as its starting point texts target summaries: abstracts written by the researchers, supplied by the original authors or provided by professional abstractors. Training a summarisation system then involves learning the properties of sentences in those abstracts and using this knowledge to extract similar abstract-worthy sentences from unseen texts. In this scenario, system performance or development progress can be evaluated by taking texts in a test sample and comparing the sentences extracted from these texts with the sentences in the target abstract.

But this approach has a number of shortcomings. First, sentence extraction on its own is a very general methodology, which can produce extracts that are incoherent or under-informative especially when used for high-compression summarisation (i.e. reducing a document to a small percentage of its original size). It is difficult to overcome this problem, because once sentences have been extracted from the source text, the context that is needed for their interpretation is not available anymore and cannot be used to produce more coherent abstracts (Spärck Jones, 1998).

Our proposed solution to this problem is to extract sentences but also to *classify* them into one of a small number of possible argumentative roles, reflecting whether the sentence expresses a main goal of the source text, a shortcoming in someone else's work, etc. The summarisation system can then use this information to generate template-like abstracts: Main goal of the text:...; Builds on work by:...; Contrasts with:...; etc.

Second, the question of what constitutes a useful gold standard has not yet been solved satisfactorily. Researchers developing corpus resources for summarisation work have often defined their own gold standard, relying on their own intuitions (see, e.g. Luhn, 1958; Edmundson, 1969) or have used abstracts supplied by authors or by professional abstractors as their gold standard (e.g. Kupiec et al., 1995; Mani and Bloedorn, 1998). Neither approach is very satisfactory. Relying only on your own intuitions inevitably creates a biased resource; indeed, Rath et al. (1961) report low agreement between human judges carrying out this kind of task. On the other hand, using abstracts as targets is not necessarily a *good* gold standard for comparison of the systems' results, although abstracts are the only kind of gold standard that comes for free with the papers. Even if the abstracts are written by professional abstractors, there are considerable differences in length, structure, and information content. This is due to differences in the common abstract presentation style in different disciplines and to the projected use of the abstracts (cf. Liddy, 1991). In the case of our corpus, an additional problem was the fact that the abstracts are written by the authors themselves and thus susceptible to differences

in individual writing style.

For the task of summarisation and relevance decision between similar papers, however, it is essential that the information contained in the gold standard is *comparable* between papers. In our approach, the vehicle for comparability of information is similarity in argumentative roles of the associated sentences. We argue that it is more difficult to find the kind of information that preserves similarity of argumentative roles, and that it is not guaranteed that it will occur in the abstract.

A related problem concerns fair evaluation of the extraction methodology. The evaluation of extracted material necessarily consists of a comparison of *sentences*, whereas one would really want to compare the informational *content* of the extracted sentences and the target abstract. Thus it will often be the case that a system extracts a sentence which in that form does not appear in the supplied abstract (resulting in a low performance score) but which is nevertheless an abstract-worthy sentence. The mismatch often arises simply because a similar idea is expressed in the supplied abstract in a very different form. But comparison of content is difficult to perform: it would require sentences to be mapped into some underlying meaning representations and then comparing these to the representations of the sentences in the gold standard. As this is technically not feasible, system performance is typically performed against a fixed gold standard (e.g. the aforementioned abstracts), which is ultimately undesirable.

Our proposed solution to this problem is to build a corpus which details not only what the abstract-worthy sentences are but also what their argumentative role is. This corpus can then be used as a resource to build a system to similarly classify sentences in unseen texts, and to evaluate that system. This paper reports on the development of a set of such argumentative roles that we have been using in our work.

In particular, we employ human intuition to annotate argumentatively defined information. We ask our annotators to classify *every* sentence in the source text in terms of its argumentative role (e.g. that it expresses the main goal of the source text, or identifies open problems in earlier work, etc). Under this scenario, system evaluation is no longer a comparison of extracted sentences against a supplied abstract, or against a single sentence that was chosen as expressing (e.g.) the main goal of the source text. Instead, *every* sentence in the source text which expresses the main goal will have been identified, and the system's performance is evaluated against *that* classification.

Of course, having someone annotate text in this way may still lead to a biased or careless annotation. We therefore needed an annotation scheme which is simple enough to be usable in a stable and intuitive way for several annotators. This paper also reports on how we tested the stability of the annotation scheme we developed. A second design criterion for our annotation scheme was that we wanted the roles to be annotated automatically. This paper reports on preliminary results which show that the annotation process can indeed be automated.

To summarise, we have argued that discourse structure information will improve summarisation. Other researchers (Ono et al., 1994; Marcu, 1997) have argued similarly, although most previous work on discourse-based summarisation follows a different discourse model, namely Rhetorical Structure Theory (Mann and Thompson, 1987). In contrast to RST, we stress the importance of rhetorical moves which are *global* to the argumentation of the paper, as opposed to more local RST–type relations. Our categories are not hierarchical, and they are much less fine-grained than RST-relations. As mentioned above, we wanted them to a) provide context information for flexible summarisation, b) provide a higher degree of comparability between papers, and c) provide a fairer evaluation of superficially different sentences.

In the rest of this paper, we will first describe how we chose the categories (section 2). Second, we had to construct training and evaluation material such that we could be sure that the proposed categorisation yielded a reliable resource of annotated text to train a system against, a gold standard. The human annotation experiments are reported in section 3. Finally, in section 4, we describe some of the automated annotation work which we have started recently and which uses a corpus annotated according to our scheme as its training material.

2 The annotation scheme

The domain in which we work is that of scientific research articles, in particular computational linguistics articles. We settled on this domain for a number of reasons. One reason is that it is a domain we are familiar with, which helps for intermediate evaluation of the annotation work. The other reason is that computational linguistics is also a rather heterogeneous domain: the papers in our collection cover a wide range of subject matters, such as logic programming, statistical language modelling, theoretical semantics and computational psycholinguistics. This makes it a challenging test bed for our

BASIC SCHEME	BACKGROUND	Sentences describing some (generally accepted) background knowledge	**FULL SCHEME**
	OTHER	Sentences describing aspects of some specific other research in a neutral way (excluding contrastive or BASIS statements)	
	OWN	Sentences describing any aspect of the own work presented in this paper – except what is covered by AIM or TEXTUAL, e.g. details of solution (methodology), limitations, and further work.	
	AIM	Sentences best portraying the particular (main) research goal of the article	
	TEXTUAL	Explicit statements about the textual section structure of the paper	
	CONTRAST	Sentences contrasting own work to other work; sentences pointing out weaknesses in other research; sentences stating that the research task of the current paper has never been done before; direct comparisons	
	BASIS	Statements that the own work uses some other work as its basis or starting point, or gets support from this other work	

Figure 1: Overview of the annotation scheme

scheme which we hope to be applicable in a range of disciplines.

Despite its heterogeneity, our collection of papers does exhibit predictable rhetorical patterns of scientific argumentation. To analyse these patterns we used Swales' (1990) CARS (Creating a Research space) model as our starting point.

The annotation scheme we designed is summarised in Figure 1. The seven categories describe argumentative roles with respect to the overall communicative act of the paper. They are to be read as mutually exclusive labels, one of which is attributed to each sentence in a text. There are two kinds of categories in this scheme: *basic* categories and *non-basic* categories. Basic categories are defined by attribution of intellectual ownership; they distinguish between:

- statements which are presented as generally accepted (BACKGROUND);
- statements which are attributed to other, specific pieces of research outside the given paper, including the authors' own previous work (OTHER);
- statements which describe the authors' own *new* contributions (OWN).

The four additional (non-basic) categories are more directly based on Swales' theory. The most important of these is AIM, as this move on its own is already a good characterisation of the entire paper, and thus very useful for the generation of abstracts. The other categories are TEXTUAL, which provides information about section structure that might prove helpful for subsequent search steps. There are two moves having to do with the author's attitude towards previous research, namely BASIS and CONTRAST. We expect this kind of information to be useful for the creation of typed links for bibliometric search tools and for the automatic determination of rival approaches in the field and intellectual ancestry of methodologies (cf. Garfield's (1979) classification of the function of citation within researchers' papers).

The structure in Figure 2, for example, displays a common rhetorical pattern of scientific argumentation which we found in many introductions. A BACKGROUND segment, in which the history and the importance of the task is discussed, is followed by a longer sequence of OTHER sentences, in which specific prior work is described in a neutral way. This discussion usually terminates in a criticism of the prior work, thus giving a motivation for the own work presented in the paper. The next sentence typically states the specific goal or contribution of the paper, often in a formulaic way (Myers, 1992).

Such regularities, where the segments are contiguous, non-overlapping and non-hierarchical, can be

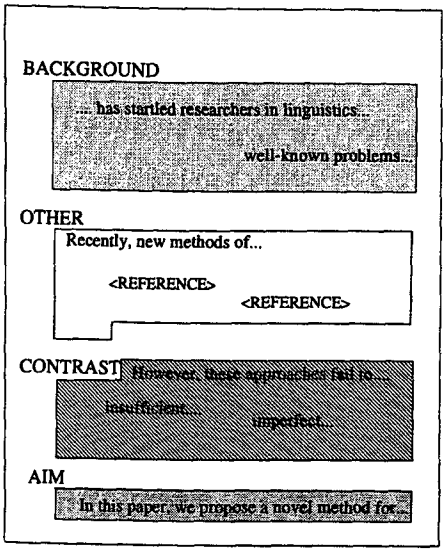

Figure 2: Typical rhetorical pattern in a research paper introduction

expressed well with our category labels. Whereas non-basic categories are typically short segments of one or two sentences, the basic categories form much larger segments of sentences with the same rhetorical role.

3 Human Annotation

3.1 Annotating full texts

To ensure that our coding scheme leads to less biased annotation than some of the other resources available for building summarisation systems, and to ensure that other researchers besides ourselves can use it to replicate our results on different types of texts, we wanted to examine two properties of our scheme: stability and reproducibility (Krippendorff, 1980). Stability is the extent to which an annotator will produce the same classifications at different times. Reproducibility is the extent to which different annotators will produce the same classification. We use the Kappa coefficient (Siegel and Castellan, 1988) to measure stability and reproducibility. The rationale for using Kappa is explained in (Carletta, 1996).

The studies used to evaluate stability and reproducibility we describe in more detail in (Teufel et al., To Appear). In brief, 48 papers were annotated by three extensively trained annotators. The training period was four weeks consisting of 5 hours of annotation per week. There were written instructions (guidelines) of 17 pages. Skim-reading and annotation of an average length (3800 word) paper typically took 20-30 minutes. The studies show that the training material is reliable. In particular, the basic annotation scheme is stable (K=.82, .81, .76; N=1220; k=2 for all three annotators) and reproducible (K=.71, N=4261, k=3), where k denotes the number of annotators, N the number of sentences annotated, and K gives the Kappa value. The full annotation scheme is stable (K=.83, .79, .81; N=1248; k=2 for all three annotators) and reproducible (K=.78, N=4031, k=3). Overall, reproducibility and stability for trained annotators does not quite reach the levels found for, for instance, the best dialogue act coding schemes, which typically reach Kappa values of around K=.80 (Carletta et al., 1997; Jurafsky et al., 1997). Our annotation requires more subjective judgements and is possibly more cognitively complex. Our reproducibility and stability results are in the range which Krippendorff (1980) describes as giving marginally significant results for reasonable size data sets when correlating two coded variables which would show a clear correlation if there were perfect agreement. As our requirements are less stringent than Krippendorff's, we find the level of agreement which we achieved acceptable.

Category	Percentage
OWN	69.4%
OTHER	15.8%
BACKGROUND	5.7%
CONTRAST	4.4%
AIM	2.4%
BASIS	1.4%
TEXTUAL	0.9%

Figure 3: Distribution of categories

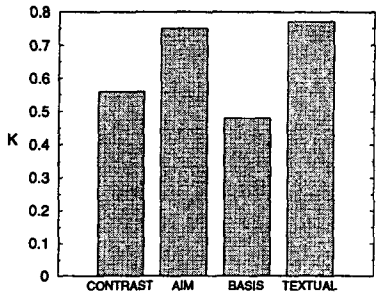

Figure 4: Reproducibility diagnostics: non-basic categories

Figure 3, which gives the overall distribution of categories, shows that OWN is by far the most frequent category. Figure 4 reports how well the four

non-basic categories could be distinguished from all other categories, measured by Krippendorff's diagnostics for category distinctions (i.e. collapsing all *other* distinctions). When compared to the overall reproducibility of .71, we notice that the annotators were good at distinguishing AIM and TEXTUAL, and less good at determining BASIS and CONTRAST. This might have to do with the location of those types of sentences in the paper: AIM and TEXTUAL are usually found at the beginning or end of the introduction section, whereas CONTRAST, and even more so BASIS, are usually interspersed within longer stretches of OWN. As a result, these categories are more exposed to lapses of attention during annotation.

The fact that the annotators are good at determining AIM sentences is an important result: as AIM sentences constitute the best characterisation of the research paper for the summarisation task at a very high compression to 1.8% of the original text length, we are particularly interested in having them annotated consistently in our training material. This result is clearly in contrast to studies which conclude that humans are not very reliable at this kind of task (Rath et al., 1961). We attribute this difference to a difference in our instructions. Whereas the subjects in Rath et al.'s experiment were asked to look for the most *relevant* sentences, our annotators had to look for specific argumentative roles which seems to have eased the task. In addition, our guidelines give very specific instructions for ambiguous cases.

These reproducibility values are important because they can act as a good evaluation measure as it factors random agreement out, unlike percentage agreement. It also provides a realistic upper bound on performance: if the machine is treated as another coder, and if reproducibity does not decrease then the machine has reached the theoretically best result, considering the cognitive difficulty of the task.

3.2 Annotating parts of texts

Annotating texts with our scheme is time-consuming, so we wanted to determine if there was a more efficient way of obtaining hand-coded training material, namely by annotating only parts of the source texts. For example, the abstract, introductions and conclusions of source texts are often like "condensed" versions of the contents of the entire paper and might be good areas to restrict annotation to. Alternatively, it might be a good idea to restrict annotation to the first 20% or the last 10% of any given text. Yet another possibility for restricting the range of sentences to be annotated is based on the 'alignment' idea introduced in (Kupiec et al., 1995): a simple surface measure determines sentences in the document that are maximally similar to sentences in the abstract.

Obviously, any of these strategies of area restriction would give us fewer gold standard sentences per paper, so we would have to make sure that we still had enough candidate sentences for all seven categories. On the other hand, because these areas could well be the most clearly written and informationally rich sections, it might be the case that the quality of the resulting gold standard is higher. In this case we would expect the reliability of the coding in these areas to be higher in comparison to the reliability achieved overall, which in turn would result in higher accuracy when this task is done automatically.

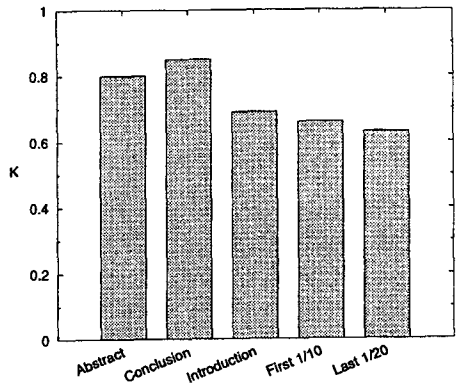

Figure 5: Reproducibility by annotated area

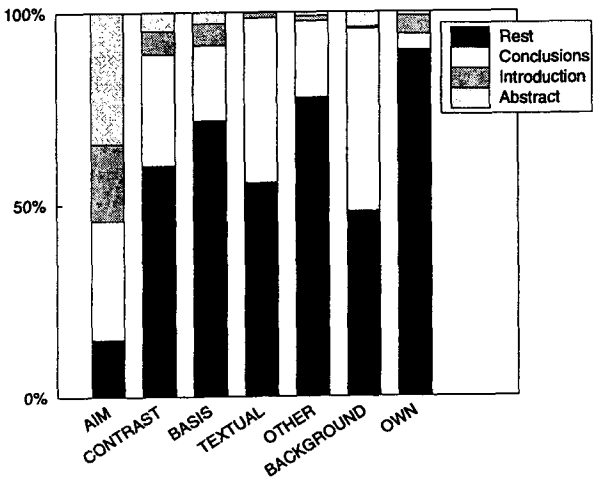

Figure 6: Label distribution by annotated area

We did extensive experiments on this. Figure 5 shows reliability values for each of the annotated portions of text, and Figure 6 shows the composi-

tion in terms of our labels for each of the annotated portions of text. The implications for corpus preparation for abstract generation experiments can be summarised as follows. If one wants to avoid manually annotating entire papers but still make all argumentative distinctions, one can restrict the annotation to sentences appearing in the introduction section, even though annotators will find them slightly harder to classify (K=.69), or to all alignable abstract sentences, even if there are not many alignable abstract sentences detectable overall (around 50% of the sentences in the abstract), or to conclusion sentences, even if the coverage of argumentative categories is very restricted in the conclusions (mostly AIM and OWN sentences).

We also examined a fall-back option of just annotating the first 10% or last 5% of a paper (as not all papers in our collection have an explicitly marked introduction and conclusion section), but the reliability results of this were far less good (K=.66 and K=.63, respectively).

4 Automatic annotation

All the annotation work is obviously in aid of development work, in particular for the training of a system. We will provide a brief description of training results so as to show the practical viability of the proposed corpus preparation method.

4.1 Data

Our training material is a collection of 80 conference papers and their summaries, taken from the Computation and Language E-Print Archive (`http://xxx.lanl.gov/cmp-lg/`). The training material contains 330,000 word tokens.

The data is automatically preprocessed into xml format, and the following structural information is marked up: title, summary, headings, paragraph structure and sentences, citations in running text, and reference list at the end of the paper. If one of the paper's authors also appears on the author list of a cited paper, then that citation is marked as self citation. Tables, equations, figures, captions, cross references are removed and replaced by place holders. Sentence boundaries are automatically detected, and the text is POS-tagged according to the UPenn tagset.

Annotation of rhetorical roles for all 80 papers (around 12,000 sentences) was provided by one of our human judges during the annotation study mentioned above.

4.2 The method

(Kupiec et al., 1995) use supervised learning to automatically adjust feature weights. 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.

We extend Kupiec et al.'s estimation of the probability that a sentence is contained in the abstract, to the probability that it has rhetorical role R (cf. Figure 7).

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

where

$P(s \in R\|F_1,\ldots,F_k)$:	Probability that sentence s in the source text has rhetorical role R, given its feature values;
$P(s \in R)$:	relative frequency of role R (constant);
$P(F_j\| s \in R)$:	probability of feature-value pair occurring in a sentence which is in rhetorical class R;
$P(F_j)$:	probability that the feature-value pair occurs unconditionally;
k:	number of feature-value pairs;
F_j:	j-th feature-value pair.

Figure 7: Naive Bayesian classifier

Evaluation of the method relies on cross-validation: the model is trained on a training set of documents, leaving one document out at a time (the test document). The model is then used to assign each sentence a probability for each category R, and the category with the highest probability is chosen as answer for the sentence.

4.3 Features

The features we use in training (see Figure 8) are different from Kupiec et al.'s because we do not estimate overall importance in one step, but instead guess argumentative status first and determine importance later.

Many of our features can be read off directly from the way the corpus is encoded: our preprocessors determine sentence-boundaries and parse the reference list at the end. This gives us a good handle on structural and locational features, as well as on features related to citations.

Type of feature	Name	Feature description	Feature values
Explicit structure	Struct–1	Type of Headline of current section	8 prototypical headlines or 'non-prototypical'
	Struct–2	Relative position of sentence within paragraph	initial, medial, final
	Struct–3	Relative position of sentence within section	first, second or last third
Relative location	Loc	Paper is segmented into 10 equally-sized segments	1–10
Citations	Cit–1	Does the sentence contain a citation or the name of an author contained in the reference list?	Full Citation, Author Name or None
	Cit–2	Does the sentence contain a *self* citation?	Yes or No
Syntactic features	Syn–1	Tense (associated with first finite verb in sentence)	Present, Past, Present Perfect, Past Perfect, Future or Nothing
	Syn–2	Modal Auxiliaries	Present or Not
	Syn–3	Voice	Active or Passive
	Syn–4	Negation	Present or Not
Semantic features	Sem–1	Action type of first verb in sentence	20 different Action Types (cf. Figure 9) or Nothing
	Sem–2	Type of Agent	Authors or Others or Nothing
	Sem–3	Type of formulaic expression occurring in sentence	18 different types of Formulaic Expressions (cf. Figure 9) or Nothing
Content Features	Cont–1	Does the sentence contain keywords as determined by the tf/idf measure?	Yes or No
	Cont–2	Does the sentence contain words also occurring in the title or headlines?	Yes or No

Figure 8: Features for supervised learning

The syntactic features rely on determining the first finite verb in the sentence, which is done symbolically using POS-information. Heuristics are used to determine the tense and possible negation.

The semantic features rely on template matching. In the feature Sem–1, a hand-crafted lexicon is used to classify the verb into one of 20 Action Classes (cf. Figure 9, left half), if it is one of the 388 verbs contained in the lexicon. The feature Sem–2 encodes whether the agent of the action is most likely to refer to the authors, or to other agents, e.g. other researchers (177 templates). Heuristic rules determine that the agent is the subject in an active sentence, or the head of the by-phrase (if present) in a passive sentence. Sem–3 encodes various other formulaic expressions (indicator phrases (Paice, 1981), meta-comments (Zukerman, 1991)) in order to exploit explicit rhetoric phrases the authors might have used, cf. Figure 9, right half (414 templates).

The content features use the tf/idf method and title and header information for finding contentful words or phrases. In contrast to all other features they do not attempt to model the form or meta-discourse contained in the sentences but instead model their domain (object-level) contents.

4.4 Results

When the Naive Bayesian Model is added to the pool of coders, the reproducibility drops from K=.71 to K=.55. This reproducibility value is equivalent to the value achieved by 6 human annotators with no prior training, as found in an earlier experiment (Teufel et al., To Appear). Compared to one of the annotators, Kappa is K=.37, which corresponds to percentage accuracy of 71.2%. This number cannot be directly compared to experiments like Kupiec et al.'s because in their experiment a compression of around 3% was achieved whereas we classify each sentence into one of the categories.

Further analysis of our results shows the system performs well on the frequent category OWN, cf. the confusion matrix in Fig. reftab:confusion. Indeed, as Figure 3 shows, OWN is so frequent that choosing OWN all the time gives us a seemingly hard-to-beat baseline with a high percentage agreement of 69% (Baseline 1). However, the Kappa statistic, which controls for expected random agreement, reveals just how bad that baseline really is: Kappa is K=–.12 (machine vs. one annotator). Random choice of categories according to the distribution of categories (Baseline 2) is a better baseline; Kappa

Action Types		Formulaic Expression Types	
AFFECT	*we hope to improve these results*	GENERAL_AGENT	*linguists*
ARGUMENTATION	*we argue against an application of*	SPECIFIC_AGENT	*according to <REF>*
AWARENESS	*we know of no other attempts...*	GAP_INTRODUCTION	*to our knowledge*
BETTER_SOLUTION	*our system outperforms that of ...*	AIM	*main contribution of this*
CHANGE	*we extend <CITE/>'s algorithm*	TEXTSTRUCTURE	*in section <CREF/>*
COMPARISON	*we tested our system against...*	DEIXIS	*in this paper*
CONTINUATION	*we follow X in postulating that*	CONTINUATION	*following the argument in*
CONTRAST	*our approach differs from X's ...*	SIMILARITY	*bears similarity to*
FUTURE_INTEREST	*we intend to improve our results...*	COMPARISON	*when compared to our*
INTEREST	*we are concerned with ...*	CONTRAST	*however*
NEED	*this approach, however, lacks...*	METHOD	*a novel method for XX-ing*
PRESENTATION	*we present here a method for...*	PREVIOUS_CONTEXT	*elsewhere, we have*
PROBLEM	*this raises the problem of how to...*	FUTURE	*avenue for improvement*
RESEARCH	*we collected our data from...*	AFFECT	*hopefully*
SIMILAR	*our approach resembles that of X...*	PROBLEM	*drawback*
SOLUTION	*we solve this problem by...*	SOLUTION	*insight*
TEXTSTRUCTURE	*the paper is organized as follows...*	POSITIVE_ADJECTIVE	*appealing*
USE	*we employ X's method...*	NEGATIVE_ADJECTIVE	*unsatisfactory*
COPULA	*our goal is to...*		
POSSESSION	*our approach has three advantages...*		

Figure 9: Types of actions and formulaic expressions

		MACHINE							
		AIM	CONTRAST	TEXTUAL	OWN	BACKGROUND	BASIS	OTHER	Total
HUMAN	AIM	115	4	10	46	15	13	4	207
	CONTRAST	11	79	5	280	92	40	89	596
	TEXTUAL	13	4	115	71	5	3	12	223
	OWN	75	61	61	7666	168	125	279	8435
	BACKGROUND	11	20	3	286	295	21	84	720
	BASIS	10	10	5	40	4	102	55	226
	OTHER	7	35	10	1120	203	173	466	2014
	Total	242	213	209	9509	782	477	989	12421

Figure 10: Confusion matrix: human vs. automatic annotation

for this baseline is K=0.

AIM categories can be determined with a precision of 48% and a recall of 56% (cf. Figure 11). These values are more directly comparable to Kupiec et al.'s results of 44% co-selection of extracted sentences with alignable summary sentences. We assume that most of the sentences extracted by their method would have fallen into the AIM category. The other easily determinable category for the automatic method is TEXTUAL (p=55%; r=52%), whereas the results for the other non-basic categories are relatively lower – mirroring the results for humans.

As far as the individual features are concerned, we found the strongest heuristics to be location, type of header, citations, and the semantic classes (indicator phrases, agents and actions); syntactic and content-based heuristics are the weakest. The first column in Figure 12 gives the predictiveness of the feature

Category	Precision	Recall
AIM	48%	56%
CONTRAST	37%	13%
TEXTUAL	55%	52%
OWN	81%	91%
BACKGROUND	38%	41%
BASIS	21%	45%
OTHER	47%	23%

Figure 11: Precision and recall per category

on its own, in terms of kappa between machine and one annotator. Some of the weaker features are not predictive enough on their own to break the dominance of the prior; in that case, they behave just like Baseline 1 (K=–.12).

The second column gives kappa for experiments using all features except the given feature, i.e. the results if this feature is left out of the pool of fea-

Feature Code	Alone	Left out
Struct–1	–.12	.37
Struct–2	–.12	.36
Struct–3	.16	.36
Struct–1–3	.18	.34
Loc	.17	.34
Cit–1	.18	.37
Cit–2	.13	.37
Cit–1–2	.18	.36
Syn–1	–.12	.37
Syn–2	–.12	.37
Syn–3	–.12	.37
Syn–4	–.12	.37
Syn–1–4	–.12	.37
Sem–1	–.12	.36
Sem–2	.07	.35
Sem–3	–.03	.36
Sem–1–3	.13	.31
Cont–1	–.12	.37
Cont–2	–.12	.37
Cont–1–2	–.12	.37
Baseline 1 (all OWN): K=–.12		
Baseline 2 (random by distr.): K=0		

Figure 12: Disambiguation potential of individual heuristics

tures. These numbers show that some of the weaker features contribute some predictive power in combination with others.

While not entirely satisfactory, these results might be taken as an indication that we have indeed managed to identify the right kinds of features for argumentative sentence classification. Taking the context into account should further increase results, as preliminary experiments with n-gram modelling have shown. In these experiments, we replaced the prior $P(s \in R)$ in Figure 7 with a n-gram based probability of that role occurring in the given context.

5 Conclusions

In this paper we have presented an annotation scheme for corpus based summarisation. In tests, we have found this annotation scheme to be stable and reproducible. On the basis of this scheme, we have created a new kind of resource for training summarisation systems: a corpus annotated with labels which indicate the argumentative role of each sentence in the text. Results of our training work show that the annotation work can be automated.

References

Jean Carletta, Amy Isard, Stephen Isard, Jacqueline C. Kowtko, Gwyneth Doherty-Sneddon, and Anne H. Anderson. 1997. The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23(1):13–31.

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

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

E. Garfield. 1979. *Citation indexing: its theory and application in science, thechnology and humanities.* Wiley, New York.

Daniel Jurafsky, Elizabeth Shriberg, and Debra Biasca, 1997. *Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual.* University of Colorado, Institute of Cognitive Science. TR-97-02.

Klaus Krippendorff. 1980. *Content analysis: an introduction to its methodology.* Sage Commtext series; 5. Sage, Beverly Hills London.

Julian Kupiec, Jan O. Pedersen, and Francine Chen. 1995. A trainable document summarizer. In *Proceedings of the 18th ACM-SIGIR Conference*, pages 68–73.

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

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

Inderjeet Mani and Eric Bloedorn. 1998. Machine learning of generic and user-focused summarization. In *Proceedings of the Fifteenth National Conference on AI (AAAI-98)*, pages 821–826.

William C. Mann and Sandra A. Thompson. 1987. Rhetorical structure theory: description and construction of text structures. In G. Kempen, editor, *Natural Language Generation: New Results in Artificial Intelligence, Psychology and Linguistics*, pages 85–95, Dordrecht. Nijhoff.

Daniel Marcu. 1997. From discourse structures to text summaries. In *Proceedings of the ACL/EACL workshop on Intelligent Scalable Text Summarization.*

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

Kenji Ono, Kazuo Sumita, and Seijii Miike. 1994. Abstract generation based on rhetorical structure extraction. In *Proceedings of the 15th International conference on Computational Linguistics (COLING-94).*

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

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

Sidney Siegel and N.J. Jr. Castellan. 1988. *Nonparametric statistics for the Behavioral Sciences*. McGraw-Hill, second edition.

Karen Spärck Jones. 1998. Automatic summarising: factors and directions. In *AAAI Spring Symposium on Intelligent Text Summarization*.

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

Simone Teufel, Jean Carletta, and Marc Moens. To Appear. An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics (EACL-99)*.

Ingrid Zukerman. 1991. Using meta-comments to generate fluent text in a technical domain. *Computational Intelligence: Special Issue on Natural Language Generation*, 7(4):276.