What's yours and what's mine: Determining Intellectual Attribution in Scientific Text

Simone Teufel[†]
Computer Science Department
Columbia University
teufel@cs.columbia.edu

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

Abstract

We believe that identifying the structure of scientific argumentation in articles can help in tasks such as automatic summarization or the automated construction of citation indexes. One particularly important aspect of this structure is the question of who a given scientific statement is attributed to: other researchers, the field in general, or the authors themselves.
We present the algorithm and a systematic evaluation of a system which can recognize the most salient textual properties that contribute to the global argumentative structure of a text. In this paper we concentrate on two particular features, namely the occurrences of prototypical agents and their actions in scientific text.

1 Introduction

When writing an article, one does not normally go straight to presenting the innovative scientific claim. Instead, one establishes other, well-known scientific facts first, which are contributed by other researchers. Attribution of ownership often happens explicitly, by phrases such as *"Chomsky (1965) claims that"*. The question of intellectual attribution is important for researchers: not understanding the argumentative status of part of the text is a common problem for non-experts reading highly specific texts aimed at experts (Rowley, 1982). In particular, after reading an article, researchers need to know who holds the "knowledge claim" for a certain fact that interests them.

We propose that segmentation according to intellectual ownership can be done automatically, and that such a segmentation has advantages for various shallow text understanding tasks. At the heart of our classification scheme is the following trisection:

- BACKGROUND (generally known work)
- OWN, new work and
- specific OTHER work.

The advantages of a segmentation at a rhetorical level is that rhetorics is conveniently constant across different articles. Subject matter, on the contrary, is not constant, nor are writing style and other factors.

BACKGROUND:
Researchers in knowledge representation agree that one of the hard problems of understanding narrative is the representation of temporal information. Certain facts of natural language make it hard to capture temporal information [...]

OTHER WORK:
Recently, **Researcher-4** has **suggested** the following solution to this problem [...].

WEAKNESS/CONTRAST:
But **this solution cannot be used** to interpret the following Japanese examples: [...]

OWN CONTRIBUTION:
We propose a solution which **circumvents** this problem while retaining the explanatory power of Researcher-4's approach.

Figure 1: Fictional introduction section

We work with a corpus of scientific papers (80 computational linguistics conference articles (ACL, EACL, COLING or ANLP), deposited on the CMP_LG archive between 1994 and 1996). This is a difficult test bed due to the large variation with respect to different factors: subdomain (theoretical linguistics, statistical NLP, logic programming, computational psycholinguistics), types of research (implementation, review, evaluation, empirical vs. theoretical research), writing style (formal vs. informal) and presentational styles (fixed section structure of type Introduction–Method–Results–Conclusion vs. more idiosyncratic, problem-structured presentation).

One thing, however, is constant across all articles: the argumentative aim of every single article is to show that the given work is a contribution to science (Swales, 1990; Myers, 1992; Hyland, 1998). Theories of scientific argumentation in research articles stress that authors follow well-predictable stages of argumentation, as in the fictional introduction in figure 1.

[†]This work was done while the first author was at the HCRC Language Technology Group, Edinburgh.

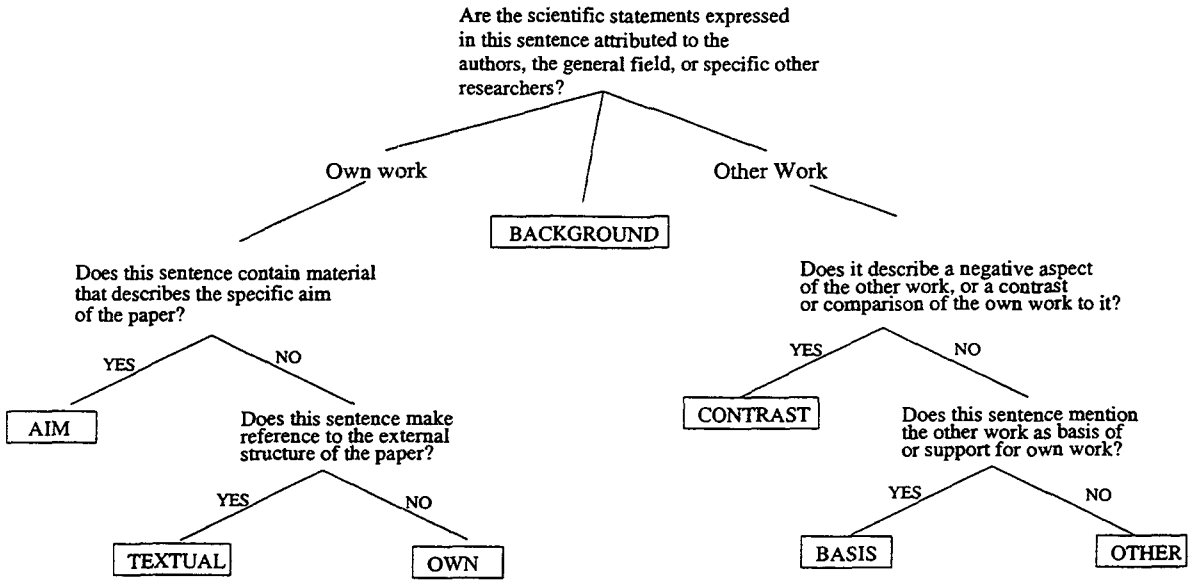

Figure 2: Annotation Scheme for Argumentative Zones

Our hypothesis is that a segmentation based on regularities of scientific argumentation and on attribution of intellectual ownership is one of the most stable and generalizable dimensions which contribute to the structure of scientific texts. In the next section we will describe an annotation scheme which we designed for capturing these effects. Its categories are based on Swales' (1990) CARS model.

1.1 The scheme

As our corpus contains many statements talking about *relations* between own and other work, we decided to add two classes ("zones") for expressing relations to the core set of OWN, OTHER and BACKGROUND, namely contrastive statements (CONTRAST; comparable to Swales' (1990) move 2A/B) and statements of intellectual ancestry (BASIS; Swales' move 2D). The label OTHER is thus reserved for neutral descriptions of other work. OWN segments are further subdivided to mark explicit aim statements (AIM; Swales' move 3.1A/B), and explicit section previews (TEXTUAL; Swales' move 3.3). All other statements about the own work are classified as OWN. Each of the seven category covers one sentence.

Our classification, which is a further development of the scheme in Teufel and Moens (1999), can be described procedurally as a decision tree (Figure 2), where five questions are asked about each sentence, concerning intellectual attribution, author stance and continuation vs. contrast. Figure 3 gives typical example sentences for each zone.

The intellectual-attribution distinction we make is comparable with Wiebe's (1994) distinction into subjective and objective statements. Subjectivity is a property which is related to the attribution of authorship as well as to author stance, but it is just one of the dimensions we consider.

1.2 Use of Argumentative Zones

Which practical use would segmenting a paper into argumentative zones have?

Firstly, rhetorical information as encoded in these zones should prove useful for summarization. Sentence extracts, still the main type of summarization around, are notoriously context-insensitive. Context in the form of argumentative relations of segments to the overall paper could provide a skeleton by which to tailor sentence extracts to user expertise (as certain users or certain tasks do not require certain types of information). A system which uses such rhetorical zones to produce task-tailored extracts for medical articles, albeit on the basis of manually-segmented texts, is given by Wellons and Purcell (1999).

Another hard task is sentence extraction from *long* texts, e.g. scientific journal articles of 20 pages of length, with a high compression. This task is hard because one has to make decisions about how the extracted sentences relate to each other and how they relate to the overall message of the text, before one can further compress them. Rhetorical context of the kind described above is very likely to make these decisions easier.

Secondly, it should also help improve citation indexes, e.g. automatically derived ones like Lawrence et al.'s (1999) and Nanba and Okumura's (1999). Citation indexes help organize scientific online literature by linking cited (outgoing) and citing (incoming) articles with a given text. But these indexes are mainly "quantitative", listing other works without further qualifying whether a reference to another work is there to extend the

AIM	*"We have proposed a method of clustering words based on large corpus data."*
TEXTUAL	*"Section 2 describes three unification-based parsers which are..."*
OWN	*"We also compare with the English language and draw some conclusions on the benefits of our approach."*
BACKGROUND	*"Part-of-speech tagging is the process of assigning grammatical categories to individual words in a corpus."*
CONTRAST	*"However, no method for extracting the relationships from superficial linguistic expressions was described in their paper."*
BASIS	*"Our disambiguation method is based on the similarity of context vectors, which was originated by Wilks et al. 1990."*
OTHER	*"Strzalkowski's Essential Arguments Approach (EAA) is a top-down approach to generation..."*

Figure 3: Examples for Argumentative Zones

earlier work, correct it, point out a weakness in it, or just provide it as general background. This "qualitative" information could be directly contributed by our argumentative zones.

In this paper, we will describe the algorithm of an argumentative zoner. The main focus of the paper is the description of two features which are particularly useful for attribution determination: prototypical agents and actions.

2 Human Annotation of Argumentative Zones

We have previously evaluated the scheme empirically by extensive experiments with three subjects, over a range of 48 articles (Teufel et al., 1999). We measured *stability* (the degree to which the same annotator will produce an annotation after 6 weeks) and *reproducibility* (the degree to which two unrelated annotators will produce the same annotation), using the Kappa coefficient K (Siegel and Castellan, 1988; Carletta, 1996), which controls agreement $P(A)$ for chance agreement $P(E)$:

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

Kappa is 0 for if agreement is only as would be expected by chance annotation following the same distribution as the observed distribution, and 1 for perfect agreement. Values of Kappa surpassing .8 are typically accepted as showing a very high level of agreement (Krippendorff, 1980; Landis and Koch, 1977).

Our experiments show that humans can distinguish own, other specific and other general work with high stability (K=.83, .79, .81; N=1248; k=2, where K stands for the Kappa coefficient, N for the number of items (sentences) annotated and k for the number of annotators) and reproducibility (K=.78, N=4031, k=3), corresponding to 94%, 93%, 93% (stability) and 93% (reproducibility) agreement.

The full distinction into all seven categories of the annotation scheme is slightly less stable and reproducible (stability: K=.82, .81, .76; N=1220; k=2 (equiv. to 93%, 92%, 90% agreement); reproducibility: K=.71, N=4261, k=3 (equiv. to 87% agreement)), but still in the range of what is generally accepted as reliable annotation. We conclude from this that humans can distinguish attribution and full argumentative zones, if trained. Human annotation is used as training material in our statistical classifier.

3 Automatic Argumentative Zoning

As our task is not defined by topic coherence like the related tasks of Morris and Hirst (1991), Hearst (1997), Kan et al. (1998) and Reynar (1999), we predict that keyword-based techniques for automatic argumentative zoning will not work well (cf. the results using text categorization as described later). We decided to perform machine learning, based on sentential features like the ones used by sentence extraction. Argumentative zones have properties which help us determine them on the surface:

- Zones appear in typical positions in the article (Myers, 1992); we model this with a set of location features.
- Linguistic features like tense and voice correlate with zones (Biber (1995) and Riley (1991) show correlation for similar zones like "method" and "introduction"). We model this with syntactic features.
- Zones tend to follow particular other zones (Swales, 1990); we model this with an ngram model operating over sentences.
- Beginnings of attribution zones are linguistically marked by meta-discourse like *"Other researchers claim that"* (Swales, 1990; Hyland, 1998); we model this with a specialized agents and actions recognizer, and by recognizing formal citations.
- Statements without explicit attribution are interpreted as being of the same attribution as previous sentences in the same segment of attribution; we model this with a modified agent feature which keeps track of previously recognized agents.

3.1 Recognizing Agents and Actions

Paice (1981) introduces grammars for pattern matching of indicator phrases, e.g. *"the aim/purpose of this paper/article/study"* and *"we conclude/propose"*. Such phrases can be useful indicators of overall importance. However, for our task, more flexible meta-discourse expressions need to be determined. The description of a research tradition, or the statement that the work described in the paper is the continuation of some other work, cover a wide range of syntactic and lexical expressions and are too hard to find for a mechanism like simple pattern matching.

Agent Type	Example
US_AGENT	*we*
THEM_AGENT	*his approach*
GENERAL_AGENT	*traditional methods*
US_PREVIOUS_AGENT	*the approach given in* X (99)
OUR_AIM_AGENT	*the point of this study*
REF_US_AGENT	*this paper*
REF_AGENT	*the paper*
THEM_PRONOUN_AGENT	*they*
AIM_REF_AGENT	*its goal*
GAP_AGENT	*none of these papers*
PROBLEM_AGENT	*these drawbacks*
SOLUTION_AGENT	*a way out of this dilemma*
TEXTSTRUCTURE_AGENT	*the concluding chapter*

Figure 4: Agent Lexicon: 168 Patterns, 13 Classes

We suggest that the robust recognition of *prototypical* agents and actions is one way out of this dilemma. The agents we propose to recognize describe fixed role-players in the argumentation. In Figure 1, prototypical agents are given in boldface ("Researchers in knowledge representation, "Researcher-4" and "we"). We also propose prototypical actions frequently occurring in scientific discourse (shown underlined in Figure 1): the researchers "agree", Researcher-4 "suggested" something, the solution "cannot be used".

We will now describe an algorithm which recognizes and classifies agents and actions. We use a manually created lexicon for patterns for agents, and a manually clustered verb lexicon for the verbs. Figure 4 lists the agent types we distinguish. The main three types are US_AGENT, THEM_AGENT and GENERAL_AGENT. A fourth type is US_PREVIOUS_AGENT (the authors, but in a *previous* paper).

Additional agent types include non-personal agents like aims, problems, solutions, absence of solution, or textual segments. There are four equivalence classes of agents with ambiguous reference (*"this system"*), namely REF_US_AGENT, THEM_PRONOUN_AGENT, AIM_REF_AGENT, REF_AGENT. The total of 168 patterns in the lexicon expands to many more as we use a replace mechanism (@WORK_NOUN is expanded to *"paper, article, study, chapter"* etc).

For verbs, we use a manually created the action lexicon summarized in Figure 6. The verb classes are based on semantic concepts such as similarity, contrast, competition, presentation, argumentation and textual structure. For example, PRESENTATION_ACTIONS include communication verbs like *"present"*, *"report"*, *"state"* (Myers, 1992; Thompson and Yiyun, 1991), RESEARCH_ACTIONS include *"analyze"*, *"conduct"* and *"observe"*, and ARGUMENTATION_ACTIONS *"argue"*, *"disagree"*, *"object to"*. Domain-specific actions are contained in the classes indicating a problem (*"fail"*, *"degrade"*, *"overestimate"*), and solution-contributing actions (*"circumvent"*, *solve"*, *"mitigate"*).

The main reason for using a hand-crafted, genre-specific lexicon instead of a general resource such as WordNet or Levin's (1993) classes (as used in Klavans and Kan (1998)), was to avoid polysemy problems without having to perform word sense disambiguation. Verbs in our texts often have a specialized meaning in the domain of scientific argumentation, which our lexicon readily encodes. We did notice some ambiguity problems (e.g. *"follow"* can mean following another approach, or it can mean follow in a sense having nothing to do with presentation of research, e.g. following an arc in an algorithm). In a wider domain, however, ambiguity would be a much bigger problem.

Processing of the articles includes transformation from LaTeX into XML format, recognition of formal citations and author names in running text, tokenization, sentence separation and POS-tagging. The pipeline uses the TTT software provided by the HCRC Language Technology Group (Grover et al., 1999). The algorithm for determining agents in subject positions (or By-PPs in passive sentences) is based on a finite automaton which uses POS-input; cf. Figure 5.

In the case that more than one finite verb is found in a sentence, the first finite verb which has agents and/or actions in the sentences is used as a value for that sentence.

4 Evaluation

We carried out two evaluations. Evaluation A tests whether all patterns were recognized as intended by the algorithm, and whether patterns were found that should not have been recognized. Evaluation B tests how well agent and action recognition helps us perform argumentative zoning automatically.

4.1 Evaluation A: Correctness

We first manually evaluated the error level of the POS-Tagging of finite verbs, as our algorithm crucially relies on finite verbs. In a random sample of 100 sentences from our corpus (containing a total of 184 finite verbs), the tagger showed a recall of

1. Start from the first finite verb in the sentence.
2. Check right context of the finite verb for verbal forms of interest which might make up more complex tenses. Remain within the assumed clause boundaries; do not cross commas or other finite verbs. Once the main verb of that construction (the "semantic" verb) has been found, a simple morphological analysis determines its lemma; the tense and voice of the construction follow from the succession of auxiliary verbs encountered.
3. Look up the lemma of semantic verb in Action Lexicon; return the associated Action Class if successful. Else return Action 0.
4. Determine if one of the 32 fixed negation words contained in the lexicon (e.g. *"not, don't, neither"*) is present within a fixed window of 6 to the right of the finite verb.
5. Search for the agent either as a by-PP to the right, or as a subject-NP to the left, depending on the voice of the construction as determined in step 2. Remain within assumed clause boundaries.
6. If one of the Agent Patterns matches within that area in the sentence, return the Agent Type. Else return Agent 0.
7. Repeat Steps 1–6 until there are no more finite verbs left.

Figure 5: Algorithm for Agent and Action Detection

Action Type	Example	Action Type	Example
AFFECT	*we hope to improve our results*	NEED	*this approach, however, lacks...*
ARGUMENTATION	*we argue against a model of*	PRESENTATION	*we present here a method for...*
AWARENESS	*we are not aware of attempts*	PROBLEM	*this approach fails...*
BETTER_SOLUTION	*our system outperforms ...*	RESEARCH	*we collected our data from...*
CHANGE	*we extend <CITE/>'s algorithm*	SIMILAR	*our approach resembles that of*
COMPARISON	*we tested our system against...*	SOLUTION	*we solve this problem by...*
CONTINUATION	*we follow <REF/> ...*	TEXTSTRUCTURE	*the paper is organized...*
CONTRAST	*our approach differs from ...*	USE	*we employ <REF/>'s method...*
FUTURE_INTEREST	*we intend to improve ...*	COPULA	*our goal is to...*
INTEREST	*we are concerned with ...*	POSSESSION	*we have three goals...*

Figure 6: Action Lexicon: 366 Verbs, 20 Classes

95% and a precision of 93%.

We found that for the 174 correctly determined finite verbs (out of the total 184), the heuristics for negation worked without any errors (100% accuracy). The correct semantic verb was determined in 96% percent of all cases; errors are mostly due to misrecognition of clause boundaries. Action Type lookup was fully correct, even in the case of phrasal verbs and longer idiomatic expressions (*"have to"* is a NEED_ACTION; *"be inspired by"* is a CONTINUE_ACTION). There were 7 voice errors, 2 of which were due to POS-tagging errors (past participle misrecognized). The remaining 5 voice errors correspond to a 98% accuracy. Figure 7 gives an example for a voice error (underlined) in the output of the action/agent determination.

Correctness of Agent Type determination was tested on a random sample of 100 sentences containing at least one agent, resulting in 111 agents. No agent pattern that should have been identified was missed (100% recall). Of the 111 agents, 105 cases were completely correct: the agent pattern covered the *complete* grammatical subject or by-PP intended (precision of 95%). There was one complete error, caused by a POS-tagging error. In 5 of the 111 agents, the pattern covered only *part*

At the point where John <ACTION TENSE=PRESENT VOICE=ACTIVE MODAL=NOMODAL NEGATION=0 ACTIONTYPE=0> **knows** </ACTION> **the truth has been** <FINITE TENSE=PRESENT_PERFECT VOICE=PASSIVE MODAL=NOMODAL NEGATION=0 ACTIONTYPE=0> **processed** </ACTION> **, a complete clause will have been** <ACTION TENSE=FUTURE_PERFECT VOICE=ACTIVE MODAL=NOMODAL NEGATION=0 ACTIONTYPE=0> **built** </ACTION> .

Figure 7: Sample Output of Action Detection

of a subject NP (typically the NP in a postmodifying PP), as in the phrase *"the problem with* **these approaches"** which was classified as REF_AGENT. These cases (counted as errors) indeed constitute no grave errors, as they still give an indication which type of agents the nominal phrase is associated with.

4.2 Evaluation B: Usefulness for Argumentative Zoning

We evaluated the usefulness of the Agent and Action features by measuring if they improve the classification results of our stochastic classifier for argumentative zones.

We use 14 features given in figure 8, some of which are adapted from sentence extraction techniques (Paice, 1990; Kupiec et al., 1995; Teufel and Moens, 1999).

1.	Absolute location of sentence in document
2.	Relative location of sentence in section
3.	Location of a sentence in paragraph
4.	Presence of citations
5.	Location of citations
6.	Type of citations (self citation or not)
7.	Type of headline
8.	Presence of tf/idf key words
9.	Presence of title words
10.	Sentence length
11.	Presence of modal auxiliaries
12.	Tense of the finite verb
13.	Voice of the finite verb
14.	Presence of Formulaic Expressions

Figure 8: Other features used

All features except `Citation Location` and `Citation Type` proved helpful for classification. Two different statistical models were used: a Naive Bayesian model as in Kupiec et al.'s (1995) experiment, cf. Figure 9, and an ngram model over sentences, cf. Figure 10. Learning is supervised and training examples are provided by our previous human annotation. Classification preceeds sentence by sentence. The ngram model combines evidence from the context (C_{m-1}, C_{m-2}) and from l sentential features $(F_{m,0}...F_{m,l-1})$, assuming that those two factors are independent of each other. It uses the same likelihood estimation as the Naive Bayes, but maximises a context-sensitive prior using the Viterbi algorithm. We received best results for n=2, i.e. a bigram model.

The results of stochastic classification (presented in figure 11) were compiled with a 10-fold cross-validation on our 80-paper corpus, containing a total of 12422 sentences (classified items).

As the first baseline, we use a standard text categorization method for classification (where each sentence is considered as a document*) Baseline 1 has an accuracy of 69%, which is low considering that the most frequent category (OWN) also covers 69% of all sentences. Worse still, the classifier classifies almost all sentences as OWN and OTHER segments (the most frequent categories). Recall on the rare categories but important categories AIM, TEXTUAL, CONTRAST and BASIS is zero or very low. Text classification is therefore not a solution.

*We used the Rainbow implementation of a Naive Bayes tf/idf method, 10-fold cross-validation.

Baseline 2, the most frequent category (OWN), is a particularly bad baseline: its recall on *all* categories except OWN is zero. We cannot see this bad performance in the percentage accuracy values, but only in the Kappa values (measured against one human annotator, i.e. k=2). As Kappa takes performance on rare categories into account more, it is a more intuitive measure for our task.

In figure 11, NB refers to the Naive Bayes model, and NB+ to the Naive Bayes model augmented with the ngram model. We can see that the stochastic models obtain substantial improvement over the baselines, particularly with respect to precision and recall of the rare categories, raising recall considerably in all cases, while keeping precision at the same level as Baseline 1 or improving it (exception: precision for BASIS drops; precision for AIM is insignificantly lower).

If we look at the contribution of single features (reported for the Naive Bayes system in figure 12), we see that `Agent` and `Action` features improve the overall performance of the system by .02 and .04 Kappa points respectively (.36 to .38/.40). This is a good performance for single features. `Agent` is a strong feature beating both baselines. Taken by itself, its performance at K=.08 is still weaker than some other features in the pool, e.g. the `Headline` feature (K=.19), the `Citation` feature (K=.18) and the `Absolute Location` Feature (K=.17). (Figure 12 reports classification results only for the stronger features, i.e. those who are better than Baseline 2). The `Action` feature, if considered on its own, is rather weak: it shows a slightly better Kappa value than Baseline 2, but does not even reach the level of random agreement (K=0). Nevertheless, if taken together with the other features, it still improves results.

Building on the idea that intellectual attribution is a segment-based phenomena, we improved the `Agent` feature by including history (feature `SAgent`). The assumption is that in unmarked sentences the agent of the previous attribution is still active. Wiebe (1994) also reports segment-based agenthood as one of the most successful features. `SAgent` alone achieved a classification success of K=.21, which makes `SAgent` the best single features available in the entire feature pool. Inclusion of `SAgent` to the final model improved results to K=.43 (bigram model).

Figure 12 also shows that different features are better at disambiguating certain categories. The `Formulaic` feature, which is not very strong on its own, is the most diverse, as it contributes to the disambiguation of six categories directly. Both `Agent` and `Action` features disambiguate categories which many of the other 12 features cannot disambiguate (e.g. CONTRAST), and `SAgent` additionally contributes towards the determination of BACKGROUND zones (along with the `Formulaic` and the `Absolute Location` feature).

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

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

Figure 9: Naive Bayesian Classifier

$$P(C_m|F_{m,0},\ldots,F_{m,l-1},C_0,\ldots,C_{m-1}) \approx P(C_m|C_{m-1},C_{m-2}) \frac{\prod_{j=0}^{l-1} P(F_{m,j}|C_m)}{\prod_{j=0}^{l-1} P(F_{m,j})}$$

m:	index of sentence (mth sentence in text)
l:	number of features considered
C_m:	target category associated with sentence at index m
$P(C_m\|F_{m,0},\ldots,F_{m,l-1},C_0,\ldots,C_{m-1})$:	Probability that sentence m has target category C_m, given its feature values $F_{m,0}$, ..., $F_{m,l-1}$ and given its context $C_0,...C_{m-1}$;
$P(C_m\|C_{m-1},C_{m-2})$:	Probability that sentence m has target category C, given the categories of the two previous sentences;
$P(F_{m,j}\|C_m)$:	Probability of feature-value pair F_j occurring within target category C at position m;
$P(F_{m,j})$:	Probability of feature value $F_{m,j}$;

Figure 10: Bigram Model

5 Discussion

The result for automatic classification is in agreement with our previous experimental results for human classification: humans, too, recognize the categories AIM and TEXTUAL most robustly (cf. Figure 11). AIM and TEXTUAL sentences, stating knowledge claims and organizing the text respectively, are conventionalized to a high degree. The system's results for AIM sentences, for instance, compares favourably to similar sentence extraction experiments (cf. Kupiec et al.'s (1995) results of 42%/42% recall and precision for extracting "relevant" sentences from scientific articles). BASIS and CONTRAST sentences have a less prototypical syntactic realization, and they also occur at less predictable places in the document. Therefore, it is far more difficult for both machine and human to recognize such sentences.

While the system does well for AIM and TEXTUAL sentences, and provides substantial improvement over both baselines, the difference to human performance is still quite large (cf. figure 11). We attribute most of this difference to the modest size of our training corpus: 80 papers are not much for machine learning of such high-level features. It is possible that a more sophisticated model, in combination with more training material, would improve results significantly. However, when we ran them on our data as it is now, different other statistical models, e.g. Ripper (Cohen, 1996) and a Maximum Entropy model, all showed similar numerical results.

Another factor which decreases results are inconsistencies in the training data: we discovered that 4% of the sentences with the same features were classified differently by the human annotation. This points to the fact that our set of features could be made more distinctive. In most of these cases, there were linguistic expressions present, such as subtle signs of criticism, which humans correctly identified, but for which the features are too coarse. Therefore, the addition of "deeper" features to the pool, which model the semantics of the meta-discourse shallowly, seemed a promising avenue. We consider the automatic and robust recognition of agents and actions, as presented here, to be the first incarnations of such features.

6 Conclusions

Argumentative zoning is the task of breaking a text containing a scientific argument into linear zones of the same argumentative status, or zones of the same intellectual attribution. We plan to use argumentative zoning as a first step for IR and shallow document understanding tasks like summarization. In contrast to hierarchical segmentation (e.g. Marcu's (1997) work, which is based on RST (Mann and Thompson, 1987)), this type of segmentation aims at capturing the argumentative status of a piece of text in respect to the overall argumentative act of the paper. It does not deter-

Method	Acc. (%)	K	Precision/recall per category (in %) AIM	CONTR.	TXT.	OWN	BACKG.	BASIS	OTHER
Human Performance	87	.71	72/56	50/55	79/79	94/92	68/75	82/34	74/83
NB+ (best results)	71	.43	40/53	33/20	62/57	85/85	30/58	28/31	50/38
NB (best results)	72	.41	42/60	34/22	61/60	82/90	40/43	27/41	53/29
Basel. 1: Text categ.	69	.13	44/9	32/42	58/14	77/90	20/5	47/12	31/16
Basel. 2: Most freq. cat.	69	-.12	0/0	0/0	0/0	69/100	0/0	0/0	0/0

Figure 11: Accuracy, Kappa, Precision and Recall of Human and Automatic Processing, in comparison to baselines

Features used (Naive Bayes System)	Acc. (%)	K	Precision/recall per category (in %) AIM	CONTR.	TXT.	OWN	BACKG.	BASIS	OTHER
`Action` alone	68	-.11	0/0	43/1	0/0	68/99	0/0	0/0	0/0
`Agent` alone	67	.08	0/0	0/0	0/0	71/93	0/0	0/0	36/23
`SAgent` alone	70	.21	0/0	17/0	0/0	74/94	53/16	0/0	46/33
`Abs. Location` alone	70	.17	0/0	0/0	0/0	74/97	40/36	0/0	28/9
`Headlines` alone	69	.19	0/0	0/0	0/0	75/95	0/0	0/0	29/25
`Citation` alone	70	.18	0/0	0/0	0/0	73/96	0/0	0/0	43/30
`Citation Type` alone	70	.13	0/0	0/0	0/0	72/98	0/0	0/0	43/24
`Citation Locat.` alone	70	.13	0/0	0/0	0/0	72/97	0/0	0/0	43/24
`Formulaic` alone	70	.07	40/2	45/2	75/39	71/98	0/0	40/1	47/13
12 other features	71	.36	37/53	32/17	54/47	81/91	39/41	22/32	45/22
12 fea.+`Action`	71	.38	38/57	34/22	58/59	81/91	39/40	25/38	48/22
12 fea.+`Agent`	72	.40	40/57	35/18	59/51	82/91	39/43	25/34	52/29
12 fea.+`SAgent`	73	.40	39/57	33/19	61/51	81/91	42/43	25/33	52/29
12 fea.+`Action`+`Agent`	71	.43	40/53	33/20	62/57	85/85	30/58	28/31	50/38
12 fea.+`Action`+`SAgent`	73	.41	41/59	34/22	62/61	82/91	41/42	27/39	51/29

Figure 12: Accuracy, Kappa, Precision and Recall of Automatic Processing (Naive Bayes system), per individual features

mine the rhetorical structure within zones. Subzone structure is most likely related to domain-specific rhetorical relations which are not directly relevant to the discourse-level relations we wish to recognize.

We have presented a fully implemented prototype for argumentative zoning. Its main innovation are two new features: prototypical agents and actions — semi-shallow representations of the overall scientific argumentation of the article. For agent and action recognition, we use syntactic heuristics and two extensive libraries of patterns. Processing is robust and very low in error. We evaluated the system without and with the agent and action features and found that the features improve results for automatic argumentative zoning considerably. History-aware agents are the best single feature in a large, extensively tested feature pool.

References

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

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

Cohen, William W. 1996. Learning trees and rules with set-valued features. In *Proceedings of AAAI-96*.

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

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

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

Kan, Min-Yen, Judith L. Klavans, and Kathleen R. McKeown. 1998. Linear Segmentation and Segment Significance. In *Proceedings of the Sixth Workshop on Very Large Corpora (COLIN G/ACL-98)*, 197–205.

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

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

Kupiec, Julian, Jan O. Pedersen, and Francine Chen.

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

Landis, J.R., and G.G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33: 159–174.

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

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

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

Marcu, Daniel. 1997. From Discourse Structures to Text Summaries. In Inderjeet Mani and Mark T. Maybury, eds., *Proceedings of the ACL/EACL-97 Workshop on Intelligent Scalable Text Summarization*, 82–88.

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

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

Nanba, Hidetsugu, and Manabu Okumura. 1999. Towards multi-paper summarization using reference information. In *Proceedings of IJCAI-99*, 926–931. `http://galaga.jaist.ac.jp:8000/~nanba/study/papers.html`.

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

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

Reynar, Jeffrey C. 1999. Statistical models for topic segmentation. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*, 357–364.

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

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

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

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

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

Teufel, Simone, and Marc Moens. 1999. Argumentative classification of extracted sentences as a first step towards flexible abstracting. In Inderjeet Mani and Mark T. Maybury, eds., *Advances in Automatic Text Summarization*, 155–171. Cambridge, MA: MIT Press.

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

Wellons, M. E., and G. P. Purcell. 1999. Task-specific extracts for using the medical literature. In *Proceedings of the American Medical Informatics Symposium*, 1004–1008.

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