Sequence Modelling for Sentence Classification in a Legal Summarisation System

Ben Hachey
School of Informatics, University of Edinburgh
2 Buccleuch Place
Edinburgh, Scotland
{bhachey}@inf.ed.ac.uk

Claire Grover
School of Informatics, University of Edinburgh
2 Buccleuch Place
Edinburgh, Scotland
{grover}@inf.ed.ac.uk

ABSTRACT

We describe a set of experiments using a wide range of machine learning techniques for the task of predicting the rhetorical status of sentences. The research is part of a text summarisation project for the legal domain for which we use a new corpus of judgments of the UK House of Lords. We present experimental results for classification according to a rhetorical scheme indicating a sentence's contribution to the overall argumentative structure of the legal judgments using four learning algorithms from the *Weka* package (C4.5, naïve Bayes, Winnow and SVMs). We also report results using maximum entropy models both in a standard classification framework and in a sequence labelling framework. The SVM classifier and the maximum entropy sequence tagger yield the most promising results.

Keywords

Automatic summarisation, Law, Discourse, Natural language, Artificial intelligence

1. INTRODUCTION

In this paper we report on a set of experiments to classify sentences for rhetorical status using a wide range of machine learning techniques. The task of classifying sentences forms part of a sentence extraction-based automatic summarisation system in the legal domain. The experiments described are part of an ongoing endeavor to determine the best classification techniques and the best feature sets for the task.

In the SUM project[1], we are exploring methods for generating flexible summaries of legal documents. Our approach to summarisation is described in detail in [6] and takes as a point of departure the work of Teufel and Moens [19, 18] (henceforth T&M). The T&M approach is an instance of what is known as the *text extraction* method of summarisation. In this approach a summary typically consists of sentences selected from the source text, with some smoothing (e.g reordering, anaphora resolution) to increase the coherence between them. Following T&M, we go beyond simple sentence selection and classify source sentences according to their rhetorical status (e.g. a description of background facts in the case, a reference to a point of law, etc.). With sentences classified in this manner, different kinds of summaries can be generated with prominence given to particular kinds of sentence. The rhetorical status classification task is the focus of this paper.

In Section 2 we describe the corpus of legal judgments that we have gathered and the manual annotation of rhetorical role classification that we have performed. Section 3 contains an overview of the feature sets that we use for our experiments. In Section 4 we report results from our experiments with four *Weka* classifiers and a maximum entropy classifier. In Section 5 we investigate treating the task as a sequence labelling problem and develop a maximum entropy tagger for this purpose. Finally, in Section 6, we draw conclusions and outline directions for future work.

2. THE HOLJ CORPUS

We have gathered a corpus of judgments of the House of Lords[2] (the HOLJ corpus). Each document contains a header providing structured information (e.g. respondent, appellant, date of hearing), followed by a sequence of (usually five) Law Lords' judgments consisting of free-running text, at least one of which is a substantial speech. Typically this will start with a statement of how the case came before the court, move on to a recapitulation of the facts, discuss one or more points of law, and then offer a ruling.

The corpus consists of 188 documents from the years 2001–2003. For a subset of these, manually created summaries are available and will be used for system evaluation.[3] The total number of words in the free text parts of the corpus is 2,887,037 and the total number of sentences is 98,645. The average sentence length is approximately 29 words. A document contains an average of 525 sentences while an individual Law Lord's judgment contains an average of 105 sentences.

The raw HTML documents are processed through a sequence of modules which convert to XML and add layers of linguistic annotation (see [6] for details); an individual Law Lord's judgment is encoded as a LORD element. All annotation is computed automatically except for manual annotation of sentences for their rhetorical status. The human annotation of rhetorical roles is performed

[1] http://www.ltg.ed.ac.uk/SUM/

[2] http://www.parliament.uk/judicial_work/judicial_work.cfm

[3] http://www.lawreports.co.uk/

Label	Freq.	Description
FACT	862 (8.5%)	The sentence recounts the events or circumstances which gave rise to legal proceedings. E.g. *On analysis the package was found to contain 152 milligrams of heroin at 100% purity.*
PROCEEDINGS	2434 (24%)	The sentence describes legal proceedings taken in the lower courts. E.g. *After hearing much evidence, Her Honour Judge Sander, sitting at Plymouth County Court, made findings of fact on 1 November 2000.*
BACKGROUND	2813 (27.5%)	The sentence is a direct quotation or citation of source of law material. E.g. *Article 5 provides in paragraph 1 that a group of producers may apply for registration . . .*
FRAMING	2309 (23%)	The sentence is part of the Law Lord's argumentation. E.g. *In my opinion, however, the present case cannot be brought within the principle applied by the majority in the Wells case.*
DISPOSAL	935 (9%)	A sentence which either credits or discredits a claim or previous ruling. E.g. *I would allow the appeal and restore the order of the Divisional Court.*
TEXTUAL	768 (7.5%)	A sentence which has to do with the structure of the document or with things unrelated to a case. E.g. *First, I should refer to the facts that have given rise to this litigation.*
OTHER	48 (0.5%)	A sentence which does not fit any of the above categories. E.g. *Here, as a matter of legal policy, the position seems to me straightforward*

Table 1: Rhetorical Annotation Scheme for Legal Judgments

on the documents after the tokenisation component has performed sentence boundary disambiguation. Manual annotation has been performed for 69 documents in the corpus and the experiments reported here were conducted using 40 of these. This subset of the corpus is similar in size to the corpus of 80 academic papers reported in Teufel and Moens [19]. Our corpus contains 290,793 words and 10,169 sentences while the T&M corpus contains 285,934 words and 12,188 sentences. Note that although our corpus contains marginally more words, the T&M corpus has a shorter average sentence length and thus contains more sentences.

The rhetorical roles that it is appropriate to assign to sentences vary from domain to domain and reflect the argumentative structure of the texts. Teufel and Moens [19] describe a set of non-hierarchically-structured labels which reflect regularities in the argumentative structure of research articles following from the author's communicative goals. For scientific articles the role labels reflect such things as the the goals of the paper, sentences describing generally accepted scientific background, etc. For the our legal domain, the author's primary communicative goal is to convince his peers that his position is legally sound, having considered the case with regard to all relevant points of law. We have analysed the structure of typical documents in our domain and derived from this seven rhetorical role categories, illustrated in Table 1. The second column shows the frequency of occurrence of each label in the manually annotated subset of the corpus. Apart from the OTHER category, the most infrequently assigned category is TEXTUAL while the most frequent is BACKGROUND. In general, the distribution across categories is more uniform than is found with the T&M labels: Teufel and Moens [19] report that their most frequent category (OWN) is assigned to 67% of sentences in their corpus while three other labels (BASIS, TEXTUAL and AIM) are each assigned to only 2% of sentences.

The 40 documents in our manually annotated subset were annotated by two annotators using guidelines which were developed by one of the authors, one of the annotators and a law professional. Eleven files were doubly annotated in order to measure inter-annotator agreement. We used the kappa coefficient of agreement as a measure of reliability. This showed that the human annotators distinguish the seven categories with a reproducibility of K=*.83* (N=1,955, k=2; where K is the kappa co-efficient, N is the number of sentences and k is the number of annotators). This is slightly higher than that reported by T&M and above the .80 mark which Krippendorf [8] suggests is the cut-off for good reliability.

3. FEATURE SETS

The feature set described in Teufel and Moens [19] includes many of the features which are typically used in sentence extraction approaches to automatic summarisation as well as certain other features developed specifically for rhetorical role classification. Briefly, the T&M feature set includes such features as: location of a sentence within the document and its subsections and paragraphs; sentence length; whether the sentence contains words from the title; whether it contains significant terms as determined by the information retrieval metric *tf*idf*; whether it contains a citation; linguistic features of the first finite verb; and cue phrases (described as meta-discourse features in Teufel and Moens [19]). The features that we have been experimenting with for the HOLJ corpus are broadly similar to those used by T&M and are described in the remainder of this section.

Location. For sentence extraction in the newswire domain, sentence location is an important feature and, though it is less dominant for T&M's scientific article domain, they did find it to be a useful indicator. T&M calculate the position of a sentence relative to segments of the document as well as sections and paragraphs. In our system, location is calculated relative to the containing paragraph and LORD element and is encoded in six integer-valued features: paragraph number after the beginning of the LORD element, paragraph number before the end of the LORD element, sentence number after the beginning of the LORD element, sentence number before the end of the LORD element, sentence number after the beginning of the paragraph, and sentence number before the end of the paragraph.

	C4.5		**NB**		**Winnow**		**SVM**		**ME**	
	I	C	I	C	I	C	I	C	I	C
Cue Phrases	47.8	47.8	39.6	39.6	31.1	31.1	52.1	52.1	48.1	48.1
Location	**65.4**	54.9	34.9	47.5	34.2	40.2	35.9	55.0	42.5	51.9
Entities	35.5	54.4	32.6	48.8	26.0	40.2	33.1	56.5	35.8	53.7
Sent. Length	27.2	55.1	20.0	49.1	27.0	40.4	12.0	56.8	21.5	54.0
Quotations	28.4	59.5	29.7	**51.8**	23.3	41.1	27.8	60.2	25.7	57.3
Them. Words	30.4	59.7	21.2	51.7	25.7	**41.4**	12.0	**60.6**	27.7	**57.5**
Baseline	12.0									

Table 2: Micro-averaged F-score results for rhetorical classification.

Thematic Words. This feature is intended to capture the extent to which a sentence contains terms which are significant, or thematic, in the document. The thematic strength of a sentence is calculated as a function of the *tf*idf* measure on words (*tf*=‘term frequency’, *idf*=‘inverse document frequency’): words which occur frequently in the document but rarely in the corpus as a whole have a high *tf*idf* score. The thematic words feature in Teufel and Moens [19] records whether a sentence contains one or more of the 18 highest scoring words. In our system we summarize the thematic content of a sentence with a real-valued thematic sentence feature, whose value is the average *tf*idf* score of the sentence’s terms.

Sentence Length. In T&M, this feature describes sentences as short or long depending on whether they are less than or more than twelve words in length. We use an integer-valued sentence length feature which is a count of the number of tokens in the sentence.

Quotation. This feature, which does not have a direct counterpart in T&M, encodes the percentage of sentence tokens inside an in-line quote and whether or not the sentence is inside a block quote.

Entities. T&M do not incorporate full-scale Named Entity Recognition in their system, though they do have a feature reflecting the presence or absence of citations. We recognize a wide range of named entities and generate binary-valued entity type features which take the value 0 or 1 indicating the presence or absence of a particular entity type in the sentence.

Cue Phrases. The term ‘cue phrase’ covers the kinds of stock phrases which are frequently good indicators of rhetorical status (e.g. phrases such as *The aim of this study* in the scientific article domain and *It seems to me that* in the HOLJ domain). T&M invested a considerable amount of effort in building hand-crafted lexicons where the cue phrases are assigned to one of a number of fixed categories. A primary aim of the current research is to investigate whether this information can be encoded using automatically computable linguistic features. If they can, then this helps to relieve the burden involved in porting systems such as these to new domains. Our preliminary cue phrase feature set includes syntactic features of the main verb (voice, tense, aspect, modality, negation), which we have shown in previous work to be correlated with rhetorical status [5]. We also use sentence initial part-of-speech and sentence initial word features to roughly approximate formulaic expressions which are sentence-level adverbial or prepositional phrases. Subject features include the head lemma, entity type, and entity subtype. These features approximate the hand-coded agent features of T&M. A main verb lemma feature simulates T&M’s *type of action* and a feature encoding the part-of-speech after the main verb is meant to capture basic subcategorisation information.

4. CLASSIFICATION RESULTS

We ran per-feature and cumulative experiments for four classifiers in the *Weka* package: an implementation of Quinlan’s [16] decision tree algorithm (C4.5); an implementation of John and Langley’s [7] algorithm incorporating statistical methods for nonparametric density estimation of continuous variables in a naïve Bayes model (NB); an implementation of Littlestone’s [10] algorithm for mistake-driven learning of a linear separator (Winnow); and an implementation of Platt’s [15] sequential minimal optimization algorithm for training a support vector classifier using polynomial kernels (SVM). We also use a publicly available version of a maximum entropy (ME) estimation toolkit[4] which contains C++ implementations of the LMVM [11] and GIS [3] estimation algorithms.[5] We use continuous features for all algorithms except Winnow and maximum entropy. In order to evaluate these, we discretize continuous features using the *Weka* filter based on Fayyad and Irani’s [4] MDL method for discretization.

Micro-averaged[6] F-scores for each classifier are presented in Table 2.[7] The I columns contain individual scores for each feature type and the C columns contain scores which incorporate features incrementally. C4.5 performs very well (65.4) with location features only, but is not able to successfully incorporate other features for improved performance. SVMs perform second best (60.6) with all features. The maximum entropy model achieves an F-score of 57.5 with all features. NB is next (51.8) with all but thematic word features. Winnow has the poorest performance with all features giving a micro-averaged F-score of 41.4.

For the most part, these scores are considerably lower than the micro-averaged F-score of 72.0 achieved by T&M. However, the picture is slightly different when we consider the systems in the context of their respective baselines. Teufel and Moens [19] report a macro-averaged F-score of 11 for always assigning the most frequent rhetorical class, similar to the simple baseline they use in earlier work. This score is 54 when micro-averaged because of the skewed distribution of rhetorical categories (67% of sentences fall into the most frequent category).

[4] Written by Zhang Le: `http://www.nlplab.cn/zhangle/maxent_toolkit.html`

[5] We used LMVM for early experiments, but all final results presented in sections 4 and 5 use GIS.

[6] Micro-averaging weights categories by their frequency in the corpus. By contrast, macro-averaging puts equal weight on each class regardless of how sparsely populated it might be.

[7] Note that while the *Weka* experiments use 10-fold cross-validation, the maximum entropy experiments use per-Lord cross-validation in anticipation of the sequencing experiments where individual Lord’s speeches should remain intact.

With the more uniform distribution of rhetorical categories in the HOLJ corpus, we get baseline numbers of 6.2 (macro-averaged) and 12.0 (micro-averaged). Thus, the actual per-sentence (micro-averaged) F-score improvement is relatively high, with our system achieving an improvement of between 29.4 and 53.4 points (to 41.4 and 65.4 respectively for the Winnow and C4.5 feature sets) where the T&M system achieves an improvement of 18 points. Like T&M, our cue phrase features are the most successful feature subset (excepting C4.5 decision trees). We find these results encouraging given that we have not invested any time in developing cue phrase features but have attempted to simulate these through fully automatic, largely domain-independent linguistic information.

Although ME approaches have proved very successful for natural language tasks, they are not in common use in the text summarisation community. Teufel and Moens [19] state simply that they experimented with maximum entropy but it did not show significant improvement over naïve Bayes. We hypothesize that this is due to the very carefully constructed feature set optimized for naïve Bayes. Results from Osborne [14], where maximum entropy was shown to perform much better than naïve Bayes when features are highly dependent, support this hypothesis. Our results also support this hypothesis. The feature subset containing the most interdependencies in our system is that which uses automatically generated linguistic features to represent cue phrase information. On this feature set, the ME classifier performs nearly 10 points better than naïve Bayes. Maximum entropy outperforms the other classifiers as well for most feature types, falling short only of the C4.5 decision tree on location features and the SVM on cue phrase and quotation features, though the cumulative numbers indicate that it is not integrating diverse information as well as the SVM does. This may be overcome using explicitly conjoined features. Furthermore, ME allows the integration of diverse information and has proved highly effective in natural language tasks with large, noisy feature sets such as text categorization, part-of-speech tagging, and named entity recognition. We focus on maximum entropy modelling for the sequencing experiments in the next section.

5. SEQUENCE MODELLING

Order is a general characteristic of natural languages that distinguishes many problems from classification tasks in other domains.[8] For example, when predicting a word's part-of-speech, a classifier should consider the surrounding labels to approximate syntactic constraints. Likewise, it is important in named entity recognition to consider the context of boundary and entity type predictions. Order is also implicit in sentence-level tasks where label contexts capture discourse constraints. Our rhetorical status classification task falls in this category since sentences of the same rhetorical class tend to cluster together in blocks.

There are a number of approaches to sequence modelling in the natural language processing literature. Hidden Markov models have been the standard for speech applications for some time and have been been applied to word-level tasks such as named entity recognition and shallow parsing, e.g. [13]. Maximum entropy Markov models (MEMMs) and conditional random fields (CRFs) have also been proposed for sequence modelling. In this work, we implement the approach used by Ratnaparkhi [17] for part-of-speech tagging and also used by Curran and Clark [1, 2] for supertagging and named entity recognition. Here, the conditional probability of a tag sequence $y_1..y_n$ given a Lord's speech $s_1..s_n$ is approximated as:

$$p(y_1..y_n|s_1..s_n) \approx \prod_{i=1}^{n} p(y_i|x_i) \qquad (1)$$

where $p(y_i|x_i)$ is the normalized probability at sentence i of a tag y_i given the context x_i. The conditional probability $p(y_i|x_i)$ has the following log-linear form:

$$p(y_i|x_i) = \frac{1}{Z(x_i)} exp(\sum_j \lambda_j f_j(x_i,y_i)) \qquad (2)$$

where the f_j include the features described in section 3 and features defined in terms of the previous two tags. This framework is very similar to that of MEMMs, a graphical framework that separates transition functions for different source states [12]. However, Ratnaparkhi's [17] model allows arbitrary state-transition structures, and because it combines all of the different source states into a single exponential model, it is likely to cope better with sparse data.

	ME		**PL**		**SEQ**	
	I	C	I	C	I	C
Cue Phrases	48.1	48.1	51.6	51.6	52.6	52.6
Location	42.5	51.9	38.0	54.0	39.5	56.2
Entities	35.8	53.7	32.0	55.2	35.5	56.5
Sent. Length	21.5	54.0	28.6	56.4	27.9	58.1
Quotations	25.7	57.3	28.5	57.7	30.5	**61.2**
Them. Words	27.7	**57.5**	26.7	**58.1**	31.7	60.8
Baseline	12.0					

Table 3: Maximum entropy F-score results for rhetorical classification.

Table 3 gives the results for sequencing (SEQ) as well as results for a model incorporating previous labels but no search (PL) and results on the original feature set (ME). Sequence modelling provides significant improvements over the classifier scores, the optimal configuration achieving an F-score gain of 3.7 points over the optimal ME classification configuration.

The following table contains results on a per category basis and shows precision (*P*), recall (*R*) and F-scores (*F*) for each rhetorical category using the optimal sequencing model. The final two columns show the distributions of the categories in the source documents and in the summaries respectively. (The latter was calculated by propagating the annotations from aligned sentences of the full document for 47 document-summary pairs.) Note that source documents and summaries exhibit different relative frequencies for the categories with e.g. DISPOSAL sentences accounting for a larger proportion of a summary than of the source document.

Rhet Role	***P***	***R***	***F***	**DocDist**	**SumDist**
FACT	57.0	49.9	53.2	8.5	10.3
PROCEEDINGS	59.7	58.1	58.9	24.0	18.4
BACKGROUND	57.9	62.1	60.0	27.5	10.2
FRAMING	56.7	66.4	61.2	23.0	30.0
DISPOSAL	71.5	47.7	57.2	9.0	31.1
TEXTUAL	89.7	81.5	85.4	7.5	0.2
OTHER	0.0	0.0	0.0	0.5	0.0
Micro Average	61.4	60.9	61.2	–	–

The system performs very well on TEXTUAL sentences because sentences having to do with document structure are easy to iden-

[8] The biomedical domain is a notable exception. Order is also implicit in gene sequencing tasks, for instance.

tify as they rarely contain a verb at all. Also, the average sentence length for TEXTUAL sentences (~ 8.3) is a reliable indicator, falling far below the overall average of ~ 29.6 words. Conversely, for FACT sentences, the performance suffers because of the heterogeneity of the lexical cue phrase features (e.g. main verb and subject) for this category, where subjects and actions range greatly from horses jumping fences to businesses starting up to councils hiring and firing employees.

A confusion matrix shows that errors for all rhetorical categories are distributed roughly proportionally to their gold standard distribution. Notable exceptions are between PROCEEDINGS and BACKGROUND and between BACKGROUND and FRAMING where errors are roughly double their gold standard distributions. These four substitutions alone account for 47.9% of the errors. Also, though they account for a much smaller number of overall errors, FACT tends to be misclassified as both PROCEEDINGS and BACKGROUND (9.3% of errors) and DISPOSAL tends to be misclassified as FRAMING (9.3% of errors).

6. CONCLUSIONS AND FUTURE WORK

We have presented classifier experiments in the context of summarisation of legal texts, for which we are developing a new corpus of UK House of Lords judgments with detailed linguistic markup in addition to rhetorical status annotation. We have compared a number of machine learning algorithms that have previously shown good performance on natural language tasks. Among these, support vector machines and maximum entropy models prove to be the best suited to our task. We introduced a robust and generic method for capturing cue phrase information based on widely available linguistic analysis tools. And we presented a sequence modelling approach to a sentence-level natural language task which improved performance significantly over the basic classifier.

While generic linguistic analysis tools (e.g. part-of-speech tagging, chunking) are easy to come by in many languages, detailed named entity recognition may not be available for a given new domain. We have invested a considerable amount of time in writing NER rules by hand for the HOLJ domain. However, current research is addressing bootstrapping NER systems from small amounts of seed data. Effective bootstrapping methods for NER will make our linguistic features fully domain-independent for domains and languages where the tools for shallow linguistic analysis are available. In current research, we are exploring active learning to minimize the time necessary to create state-of-the-art systems for named entity recognition in novel domains such as astronomy and law.

Finally, on the system level, we are currently developing the sentence extraction component. As with the rhetorical parser, the core of this component will be a classifier that predicts, in this case, whether or not a sentence is a good summary sentence. Once this is finished, we will have the building blocks for our summaries. Content will initially be structured using rhetorical templates. We will then be ready to carry out user studies to assess the quality of our system's output. and compare our summary text structuring with other methods such as Lapata's [9] probabilistic approach.

7. ACKNOWLEDGMENTS

This work is supported by EPSRC grant GR/N35311. We are grateful to Steve Clark for discussions on sequencing and to Vasilis Karaiskos for the annotation.

8. REFERENCES

[1] J. R. Curran and S. Clark. Investigating GIS and smoothing for maximum entropy taggers. In *EACL'03*, 2003.

[2] J. R. Curran and S. Clark. Language independent NER using a maximum entropy tagger. In *CoNLL-03*, 2003.

[3] J. N. Darroch and D. Ratcliff. Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics*, 43(5):1470–1480, 1972.

[4] U. Fayyad and K. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In *IJCAI'93*, 1993.

[5] C. Grover, B. Hachey, I. Hughson, and C. Korycinski. Automatic summarisation of legal documents. In *ICAIL 2003*, Edinburgh, Scotland, 2003.

[6] B. Hachey and C. Grover. A rhetorical status classifier for legal text summarisation. In *ACL-2004 Text Summarization Branches Out Workshop*, 2004.

[7] G. H. John and P. Langley. Esitmating continuous distributions in bayesian classifiers. In *UAI'95*, 1995.

[8] K. Krippendorff. *Content analysis: An Introduction to its Methodology*. Sage Publications, Beverly Hills, CA, 1980.

[9] M. Lapata. Probabilistic text structuring: Experiments with sentence ordering. In *ACL-2003*, Sapporo, 2003.

[10] N. Littlestone. Learning quickly when irrelevant attributes are abound: A new linear threshold algorithm. *Machine Learning*, 2:285–318, 1988.

[11] R. Malouf. A comparison of algorithms for maximum entropy parameter estimation. In *CoNLL-2002*, 2002.

[12] A. McCallum, D. Freitag, and F. Pereira. Maximum entropy markov models for information extraction and segmentation. In *ICML-2000*, 2000.

[13] A. Molina and F. Pla. Shallow parsing using specialized HMMs. *The Journal of Machine Learning Research*, 2:595–613, 2002.

[14] M. Osborne. Using maximum entropy for sentence extraction. In *ACL-2002 Automatic Summarization Workshop*, Philadelphia, USA, July 2002.

[15] J. C. Platt. Fast training of support vector machines using sequential minimal optimization. In B. Schölkopf, C. J. Burges, and A. J. Smola, editors, *Advances in Kernel Methods: Support Vector Learning*. MIT Press, 1998.

[16] R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA, 1993.

[17] A. Ratnaparkhi. A maximum entropy part-of-speech tagger. In *EMNLP-1996*, 1996.

[18] S. Teufel and M. Moens. Discourse-level argumentation in scientific articles: human and automatic annotation. In *ACL-1999 Towards Standards and Tools for Discourse Tagging Workshop*, 1999.

[19] S. Teufel and M. Moens. Summarising scientific articles-experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445, 2002.