A Rhetorical Status Classifier for Legal Text Summarisation

Ben Hachey and Claire Grover
School of Informatics
University of Edinburgh
{bhachey,grover}@inf.ed.ac.uk

Abstract

We describe a classifier which determines the rhetorical status of sentences in texts from a corpus of judgments of the UK House of Lords. Our summarisation system is based on the work of Teufel and Moens where sentences are classified for rhetorical status to aid sentence selection. We experiment with a variety of linguistic features with results comparable to Teufel and Moens, thereby demonstrating the feasibility of porting this kind of system to a new domain.

1 Introduction

Law reports form an interesting domain for automatic summarisation. They are texts which record the proceedings of a court and, due to the role that precedents play in English law, easy access to them is essential for a wide range of people. For this reason, they are frequently manually summarised by legal experts, with summaries varying according to target audience (e.g. students, solicitors).

In the SUM project, we are exploring methods for generating flexible summaries of legal documents, taking as our point of departure the Teufel and Moens (2002; 1999a; 1999b) approach to automatic summarisation (henceforth T&M). We have chosen to work with law reports for three main reasons: (a) the existence of manual summaries means that we have evaluation material for the final summarisation system; (b) the existence of differing target audiences allows us to explore the issue of tailored summaries; and (c) the texts have much in common with the academic papers that T&M worked with, while remaining challengingly different in many respects. Our general aims are comparable with those of the SALOMON project (Moens et al., 1997), which also deals with summarisation of legal texts, but our choice of methodology is designed to test the portability of the T&M approach to a new domain.

The T&M approach is an instance of what Spärck Jones (1999) terms *text extraction* where a summary typically consists of sentences selected from the source text, with some smoothing to increase the coherence between the sentences. Since the academic texts they use are rather long and the aim is to produce flexible summaries of varying length and for various audiences, T&M go beyond simple sentence selection and classify source sentences according to their rhetorical status (e.g. a description of the main result, a criticism of someone else's work, etc.). With sentences classified in this manner, different kinds of summaries can be generated. Sentences can be reordered, since they have rhetorical roles associated with them, or they can be suppressed if a user is not interested in certain types of rhetorical roles.

In the second stage of our project we will explore techniques for sentence selection. Following the T&M methodology, we will annotate sentences in the corpus for 'relevance'. For our corpus we hope to be able to compute relevance by using automatic techniques to pair up sentences from manually created abstracts with sentences in the source text. The addition of this layer of annotation will provide the training and testing material for sentence extraction, with the rhetorical role labels helping to constrain the type of summary generated.

In this paper we focus on our rhetorical status classifier. This is a key part of the summarisation process and our work can be thought of as a test of portability of the T&M approach to a new domain. At the same time, our methods differ in important respects from those of T&M and in reporting our work we will attempt to draw comparisons wherever possible.

In Section 2 we describe the House of Lords corpus we have gathered and annotated. We explain the rhetorical role annotation scheme that we have developed and contrast it with the T&M scheme for academic articles. We provide inter-annotation agreement results for the annotation scheme. In Section 2.3 we give an overview of the tools and techniques we have used in the automatic linguistic processing of the judgments. Section 3 describes our sentence classifier. In Section 3.1 we review the kinds of features that can be used by a classifier and describe the set of features used in our experiments. In Section 3.2 we present the results of experiments with four classifiers and discuss the relative effectiveness of the methods and the feature sets. Finally, in Section 4 we draw some conclusions and outline future work.

2 The HOLJ Corpus

2.1 Corpus Overview

The texts in our corpus are judgments of the House of Lords[1], which we refer to as HOLJ. These texts contain a header providing structured information, followed by a sequence of Law Lord's judgments consisting of free-running text. The structured part of the document contains information such as the respondent, appellant and the date of the hearing. The decision is given in the opinions of the Law Lords, at least one of which is a substantial speech. This often starts with a statement of how the case came before the court. Sometimes it will move to a recapitulation of the facts, moving on to discuss one or more points of law, and then offer a ruling.

We have gathered a corpus of 188 judgments from the years 2001–2003 from the House of Lords website. (For 153 of these, manually created summaries are available[2] and will be used for system evaluation). The raw HTML documents are processed through a sequence of modules which automatically add layers of annotation. The first stage converts the HTML to an XML format which we refer to as HOLXML. In HOLXML, a House of Lords Judgment is defined as a J element whose BODY element is composed of a number of LORD elements (usually five). Each LORD element contains the judgment of one individual lord and is composed of a sequence of paragraphs (P elements) inherited from the original HTML. The total number of words in the BODY elements in the corpus is 2,887,037 and the total number of sentences is 98,645. The average sentence length is approx. 29 words. A judgment contains an average of 525 sentences while an individual LORD speech contains an average of 105 sentences.

All annotation is computed automatically except for manual annotation of sentences for their rhetorical status. The automatic processing is divided into two stages, tokenisation, which also includes part-of-speech (POS) tagging and sentence boundary disambiguation, followed by linguistic annotation (described in detail in Section 2.3 below). The human annotation of rhetorical roles is performed on the documents after tokenisation has identified the sentences. This annotation is work in progress and so far we have 40 manually annotated documents. The classifiers described in this paper have been trained and evaluated on this manually annotated subset of the corpus.

Our working subset of the corpus is similar in size to the corpus reported in (Teufel and Moens, 2002): the T&M corpus consists of 80 conference articles while ours consists of 40 HOLJ documents. The T&M corpus contains 12,188 sentences and 285,934 words while ours contains 10,169 sentences and 290,793 words. The experimental results reported in this paper were obtained using 10-fold cross validation over the 40 documents.

2.2 Rhetorical Status Annotation

The rhetorical roles that it would be appropriate to assign to sentences[3] vary from domain to domain and reflect the argumentative structure of the texts. Teufel and Moens (2002) describe a set of labels which reflect regularities in the argumentative structure of research articles following from the authors' communicative goals. The scientific article rhetorical roles include labels such as AIM, which is assigned to sentences indicating the goals of the paper, and BACKGROUND, which is assigned to sentences describing generally accepted scientific background.

For the legal domain, the communicative goal is slightly different; 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, as 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. The distribution across categories is more uniform than that of the T&M labels: Teufel and Moens (2002) report that their most frequent category (OWN) is assigned to 67% of sentences while three other labels (BASIS, TEXTUAL and AIM) are each assigned to only 2% of sentences.

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

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

[3] We take the sentence as the level of processing for rhetorical role annotation. While clause-level annotation might allow more detailed discourse information, there are considerably more clauses in the HOLJ documents than sentences and annotating at the clause level would be significantly more expensive. Moreover, clause boundary identification is less reliable than sentence boundary identification.

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

The 40 judgments 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.[4] 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 (1980) suggests is the cut-off for good reliability.

In striving to achieve high quality summarisation, it is tempting to consider using an annotation system which reflects a more sophisticated analysis of rhetorical roles. However, our kappa co-efficient is currently on the bottom end of the range suggested as indicating 'good' reliability. Therefore, we suspect that the methods we are using may not scale to more refined distinctions. The decisions we made with the annotation scheme reflect a desire to balance quality of annotation against detail. Also, as mentioned earlier, the cost of annotation for these complex legal documents is not insignificant.

2.3 Linguistic Analysis

One of the aims of the SUM project is to create an annotated corpus in the legal domain which will be available to NLP researchers. With this aim in mind we have used the HOLXML format for the corpus and we encode all the results of linguistic processing as XML annotations. Figure 1 shows the broad details of the automatic processing that we perform, with the processing divided into an initial tokenisation module and a later linguistic annotation module. The architecture of our system is one where a range of NLP tools is used in a modular, pipelined way to add linguistic knowledge to the XML document markup.

In the tokenisation module we convert from the source HTML to HOLXML and then pass the data through a sequence of calls to a variety of XML-based tools from the LT TTT and LT XML toolsets (Grover et al., 2000; Thompson et al., 1997). The core program in our pipelines is the LT TTT program *fsgmatch*, a general purpose transducer which processes an input stream and adds annotations using rules provided in a hand-written grammar file. The other main LT TTT program is *ltpos*, a statistical combined part-of-speech (POS) tagger and sentence boundary disambiguation module (Mikheev, 1997). The first step in the tokenisation modules uses *fsgmatch* to segment the contents of the paragraphs into word tokens encoded in the XML as W elements. Once the word tokens have been identified, the next step uses *ltpos* to mark up the sen-

[4] The doubly annotated files were used only for computing kappa. For the experiments, we trained and tested on the 40 annotated files produced by the main annotator.

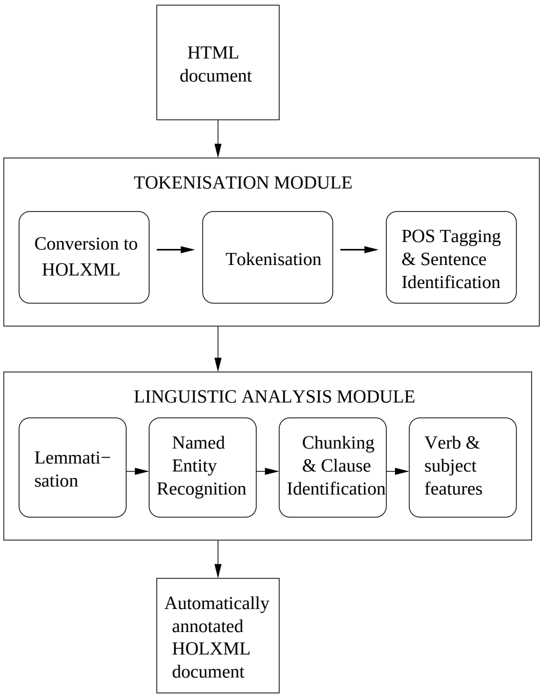
HTML document
TOKENISATION MODULE
Conversion to HOLXML
Tokenisation
POS Tagging & Sentence Identification
LINGUISTIC ANALYSIS MODULE
Lemmati−sation
Named Entity Recognition
Chunking & Clause Identification
Verb & subject features
Automatically annotated HOLXML document

Figure 1: HOLJ processing stages

tences as SENT elements and to add part of speech attributes to word tokens.

The motivation for the module that performs further linguistic analysis is to compute information to be used to provide features for the sentence classifier. However, the information we compute is general purpose, making the data useful for a range of NLP research activities.

The first step in the linguistic analysis module lemmatises the inflected words using Minnen et al.'s (2000) *morpha* lemmatiser. This program is not XML-aware so we use *xmlperl* (McKelvie, 1999) to provide a wrapper so that it can be incorporated in the XML pipeline. We use a similar mechanism for the other non-XML components.

The next stage, described in Figure 1 as Named Entity Recognition, is in fact a more complex layering of two kinds of named entity recognition. The documents in our domain contain the standard kinds of entities familiar from the MUC and CoNLL competitions (Chinchor, 1998; Roth and van den Bosch, 2002; Daelemans and Osborne, 2003), such as person, organisation, location and date. However, they also contain entities which are are specific to the domain. Table 2 shows examples of the entities we have marked up in the corpus (in our annotation scheme these are noun groups (NG) with specific type and subtype attributes). In the top two blocks of the table are examples of domain-specific entities such as courts, judges, acts and judgments, while in the third block we show examples of non-domain-specific entity types. We use different strategies for the identification of the two classes of entities: for the domain-specific ones we use hand-crafted LT TTT rules, while for the non-domain-specific ones we use the C&C named entity tagger (Curran and Clark, 2003) trained on the MUC7 data set. For some entities, the two approaches provide competing analyses and in all cases the domain-specific label is to be preferred since it provides finer-grained information. However, while the rule-based recogniser can operate incrementally over data which already contains some entity markup, the C&C tagger is trained to operate over unlabelled sentences. For this reason we run the C&C tagger first and encode its results as attributes on the words. We then run the domain-specific tagger, encoding its results as XML elements enclosing the words, and finish with a similar encoding of whichever C&C entities can still be realised in the unlabelled subparts of the sentences (these are labelled as subtype='fromCC').

Part of the rule-based entity recognition component builds an 'on-the-fly' lexicon from names found in the header of the document. Here the names of the lords who are judging the case are listed as well as the names of the respondent and appellant. Since instances of these three entities occurring in the body of the judgment are likely to be distributed differently across sentences with different rhetorical roles, it is useful to mark them up explicitly. We create an expanded lexicon from the 'on-the-fly' lexicon containing entries for consecutive substrings of the original entry in order to perform a more flexible lexical look-up. Thus the entity *Commission* is recognised as an appellant substring entity in the document where *Northern Ireland Human Rights Commission* has been identified as an appellant entity.

As future work, we plan to create a named entity gold standard for the HOLJ domain and evaluate the named entity recognition we are performing. For now, we can use rhetorical status classification as a task-based evaluation to estimate the utility of entity recognition. The generic C&C entity recognition together with the hand-crafted rules for the HOLJ domain prove to be the third most effective feature set after the cue phrase and location features (Table 3).

The next stage in the linguistic analysis module performs noun group and verb group chunking using *fsgmatch* with the specialised hand-written rule sets which were the core part of LT CHUNK (Finch and Mikheev, 1997). The noun group and verb group mark-up plus POS tags provide the relevant

<NG type='enamex-pers' subtype='committee-lord'>	*Lord Rodger of Earlsferry* *Lord Hutton*
<NG type='caseent' subtype='appellant'> <NG type='caseentsub' subtype='appellant'>	*Northern Ireland Human Rights Commission* *Commission*
<NG type='caseent' subtype='respondent'> <NG type='caseentsub' subtype='respondent'>	*URATEMP VENTURES LIMITED* *Uratemp Ventures*
<NG type='enamex-pers' subtype='judge'>	*Collins J* *Potter and Hale LJJ*
<NG type='enamex-org' subtype='court'>	*European Court of Justice* *Bristol County Court*
<NG type='legal-ent' subtype='act'>	*Value Added Tax Act 1994* *Adoption Act 1976*
<NG type='legal-ent' subtype='section'>	*section 18(1)(a)* *para 3.1*
<NG type='legal-ent' subtype='judgment'>	*Turner J [1996] STC 1469* *Apple and Pear Development Council v Commissioners of Customs and Excise (Case 102/86) [1988] STC 221*
<NG type='enamex-loc' subtype='fromCC'>	*Oakdene Road* *Kuwait Airport*
<NG type='enamex-pers' subtype='fromCC'>	*Irfan Choudhry* *John MacDermott*
<NG type='enamex-org' subtype='fromCC'>	*Powergen* *Grayan Building Services Ltd*

Table 2: Named entities in the corpus

features for the next processing step. In a previous paper we showed that a range of information about the main verb group of the sentence was likely to provide important clues as to the rhetorical status of the sentence (e.g. a present tense active verb will correlate more highly with BACKGROUND or DISPOSAL sentences while a simple past tense sentence is more likely to be found in a FACT sentence). In order to find the main verb group of a sentence, however, we need to establish its clause structure. We do this with a clause identifier (Hachey, 2002) built using the CoNLL-2001 shared task data (Sang and Déjean, 2001). Clause identification is performed in three steps. First, two maximum entropy classifiers (Berger et al., 1996) are applied, where the first predicts clause start labels and the second predicts clause end labels. In the the third step clause segmentation is inferred from the predicted starts and ends using a maximum entropy model whose sole purpose is to provide confidence values for potential clauses.

The final stages of linguistic processing use hand-written LT TTT components to compute features of verb and noun groups. For all verb groups, attributes encoding tense, aspect, modality and negation are added to the mark-up: for example, *might not have been brought* is analysed as <VG tense='pres', aspect='perf', voice='pass', modal='yes', neg='yes'>. In addition, subject noun groups are identified and lemma information from the head noun of the subject and the head verb of the verb group are propagated to the verb group attribute list.

3 The Sentence Classifier

3.1 Feature Sets

The feature set described in Teufel and Moens (2002) 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 *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, 2002). The features that we have been experimenting with for the HOLJ domain are broadly similar to those used by T&M and are described in the remainder of this section.

Location. For sentence extraction in the news 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, sentence number after the beginning of the LORD element, sentence number before the end of the LORD, sentence number after the beginning of the paragraph, and sentence number before the end of the paragraph.

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 (2002) records whether a sentence contains one or more of the 18 highest scoring words. In our system we summarise 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 implement 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 recognise a wide range of named entities and generate binary-valued entity type features which take the value 0 or 1.

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 compiling lists of such cue phrases and 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 the effects of T&M's cue phrase features can be achieved 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 to be correlated with rhetorical status (Grover et al., 2003). We also use features indicating 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.

3.2 Classifier Results and Discussion

We ran experiments for four classifiers in the *Weka* package:[5] C4.5 decision trees, naïve Bayes (NB) incorporating nonparametric density estimation of continuous variables, the Winnow[6] algorithm for mistake-driven learning of a linear separator, and the sequential minimal optimization algorithm for training support vector machines (SVM) using polynomial kernels. Default parameter settings are used for all algorithms.

Micro-averaged[7] F-scores for each classifier are presented in Table 3. The I columns contain individual scores for each feature type and the C columns contain cumulative 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. 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 T&M, where they achieve a micro-averaged F-score of 72. However, the picture is slightly different when we consider the systems in the context of their respective baselines. Teufel and Moens (2002) 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

[5] `http://www.cs.waikato.ac.nz/ml/weka/`

[6] To evaluate Winnow, we use the *Weka* implementation of the MDL discretization method which recursively splits intervals at the cut-off point that minimizes entropy.

[7] Micro-averaging weights categories by their prior probability. By contrast, macro-averaging puts equal weight on each class regardless of how sparsely populated it might be.

	C4.5		**NB**		**Winnow**		**SVM**	
	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
Location	**65.4**	54.9	34.9	47.5	34.2	40.2	35.9	55.0
Entities	35.5	54.4	32.6	48.8	26.0	40.2	33.1	56.5
Sent. Length	27.2	55.1	20.0	49.1	27.0	40.4	12.0	56.8
Quotations	28.4	59.5	29.7	**51.8**	23.3	41.1	27.8	60.2
Them. Words	30.4	59.7	21.2	51.7	25.7	**41.4**	12.0	**60.6**
Baseline	12.0							

Table 3: Micro-averaged F-score results for rhetorical classification

because of the skewed distribution of rhetorical categories (67% of sentences fall into the most frequent category).[8]

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 optimal Winnow and C4.5) 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 very encouraging given that we have not invested any time in developing the hand-crafted cue phrase features that proved most useful for T&M, but rather have attempted to simulate these through fully automatic, largely domain-independent linguistic information.

The fact that C4.5 decision trees outperform all algorithms on location features led us to believe we might be using an inferior representation for location features. To test this, we encoded our location features in the same way as T&M. This gave improved F scores for SVMs (41.5) and naïve Bayes (41.0) but worse scores for Winnow and dramatically worse scores for C4.5, indicating, as one would expect, that the discrete T&M location features lose information present in our non-discretized location features.

Maximum entropy (ME) modelling is another machine learning method which allows the integration of diverse information sources. ME approaches explicitly model the dependence between features and have proven highly effective in similar natural language tasks such as text categorisation, part-of-speech tagging, and named entity recognition. The next step in our research will be to experiment with maximum entropy modelling and to compare it with the techniques reported here.

4 Conclusions and Future Work

We have presented new work on the 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 and sentence extraction annotation.

We have effectively laid the ground work for detailed experiments with robust and generic methods for capturing cue phrase information. This is favourable as it can be automatically ported to new text summarisation domains where the tools are available for linguistic analysis, as opposed to relying on cue phrases which need to be hand-crafted for each domain. Hand-crafted cue phrase lists are necessarily more fragile and more susceptible to over-fitting in large-scale applications.

Future experiments will use maximum entropy modelling to incorporate our diverse range of sparse linguistic and textual features. We plan to experiment with maximum entropy for sentence-level rhetorical status prediction in both standard classification and sequence modelling frameworks.

We also intend to incorporate bootstrapped named entity recognition systems. While generic linguistic analysis tools (e.g. part-of-speech tagging, chunking) are easy to come by in many languages, domain-specific named entity recognition is not. We have invested a considerable amount of time in writing named entity rules by hand for the HOLJ domain. However, current research is investigating methods for bootstrapping named entity systems from small amounts of seed data. Effective methods will make our linguistic features fully domain-independent for domains and languages where linguistic analysis tools are available.

[8] T&M use macro-averaging in order to down-weight their largest category which was the least interesting for their summaries. With our more uniform distribution of rhetorical categories and without any reason, as yet, to expect the number of summary sentences coming from any one category to be far out of proportion, we believe it better to report micro-averaged scores. If we compare macro-averaged F scores, the SVM classifier achieves a score (52) a bit higher than T&M (50). C4.5 outperforms both by a considerable amount, achieving a macro-averaged F score of 58.

For future work, we are considering active learning and co-training. Active learning (Cohn et al., 1994) would seem the appropriate starting point for our task as we currently have no gold standard data but we do have annotation resources. We may also benefit from co-training (Blum and Mitchell, 1998) and rule induction (Riloff and Jones, 1999) with the seed data set from the initial annotation for active learning.

We have also performed a preliminary experiment with hypernym features for subject and verb lemmas which should allow better generalisation over cue phrase information. This is a rather noisy feature as we are not performing word sense disambiguation, but adding all WordNet hypernyms of the first three senses as features. Nevertheless, this has shown an improvement with the naïve Bayes classifier from 24.75 for the cue phrase features sets (minus lemma features) to 27.45 when hypernyms are included. Future work will further investigate hypernym features.

Acknowledgments

This work is supported by EPSRC grant GR/N35311. The corpus annotation was carried out by Vasilis Karaiskos and Hui-Mei Liao using the NITE XML Toolkit (NXT), with assistance from Jonathan Kilgour in configuring NXT. We are grateful to Stephen Clark and three anonymous reviewers for their valuable comments.

References

Adam Berger, Stephen Della Pietra, and Vincent Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, Madison, Wisconsin.

Nancy A. Chinchor. 1998. *Proceedings of the Seventh Message Understanding Conference (MUC-7)*. Fairfax, Virginia.

David Cohn, Les Atlas, and Richard Ladner. 1994. Improving generalization with active learning. *Machine Learning*, 15(2):201–221.

James R. Curran and Stephen Clark. 2003. Language independent NER using a maximum entropy tagger. In *Proceedings of CoNLL-2003*, pages 164–167. Edmonton, Canada.

Walter Daelemans and Miles Osborne. 2003. *Proceedings of the Seventh Workshop on Computational Language Learning (CoNLL-2003)*. Edmonton, Canada.

Steve Finch and Andrei Mikheev. 1997. A workbench for finding structure in texts. In *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP-97)*. Washington D.C.

Claire Grover, Colin Matheson, Andrei Mikheev, and Marc Moens. 2000. LT TTT—a flexible tokenisation tool. In *LREC 2000—Proceedings of the 2nd International Conference on Language Resources and Evaluation*, pages 1147–1154.

Claire Grover, Ben Hachey, and Chris Korycinski. 2003. Summarising legal texts: Sentential tense and argumentative roles. In *HLT-NAACL 2003 Workshop: Text Summarization (DUC03)*, pages 33–40, Edmonton, Canada.

Ben Hachey. 2002. Recognising clauses using symbolic and machine learning approaches. Master's thesis, University of Edinburgh.

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

David McKelvie. 1999. Xmlperl 1.0.4 XML processing software. http://www.cogsci.ed.ac.uk/~dmck/xmlperl.

Andrei Mikheev. 1997. Automatic rule induction for unknown word guessing. *Computational Linguistics*, 23(3):405–423.

Guido Minnen, John Carroll, and Darren Pearce. 2000. Robust, applied morphological generation. In *Proceedings of 1st International Natural Language Generation Conference (INLG'2000)*.

Marie-Francine Moens, Caroline Uyttendaele, and Jos Dumortier. 1997. Abstracting of legal cases: The SALOMON experience. In *Proceedings of the Sixth International Conference on Artificial Intelligence and Law, ACM*, pages 114–122.

Ellen Riloff and Rosie Jones. 1999. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the 16th National Conference on Artificial Intelligence (AIII-99)*, Orlando, Florida.

Dan Roth and Antal van den Bosch. 2002. *Proceedings of the Sixth Workshop on Computational Language Learning (CoNLL-2002)*. Taipei, Taiwan.

Erik Tjong Kim Sang and Hervé Déjean. 2001. Introduction to the CoNLL-2001 shared task: clause identification. In *Proceedings of the Fifth Workshop on Computational Language Learning*, pages 53–57.

Karen Spärck-Jones. 1998. Automatic summarising: factors and directions. In *Advances in Automatic Text Summarisation*, pages 1–14. MIT Press.

Simone Teufel and Marc Moens. 1999a. Argumentative classification of extracted sentences as a first step towards flexible abstracting. In *Advances in Automatic Text Summarization*, pages 137–175. MIT Press.

Simone Teufel and Marc Moens. 1999b. Discourse-level argumentation in scientific articles: human and automatic annotation. In *Towards Standards and Tools for Discourse Tagging*, pages 84–93. ACL Workshop.

Simone Teufel and Marc Moens. 2002. Summarising scientific articles—experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445.

Henry Thompson, Richard Tobin, David McKelvie, and Chris Brew. 1997. LT XML. software API and XML toolkit. http://www.ltg.ed.ac.uk/software/.