The HOLJ Corpus: supporting summarisation of legal texts

Claire Grover **Ben Hachey** **Ian Hughson**
School of Informatics, University of Edinburgh
2 Buccleuch Place
Edinburgh, EH8 9LW
Scotland, UK
{grover,bhachey}@inf.ed.ac.uk, S.I.Hughson@sms.ed.ac.uk

Abstract

We describe an XML-encoded corpus of texts in the legal domain which was gathered for an automatic summarisation project. We describe two distinct layers of annotation: manual annotation of the rhetorical status of sentences and an entirely automatic annotation process incorporating a host of individual linguistic processors. The manual rhetorical status annotation has been developed as training and testing material for a summarisation system based on the work of Teufel and Moens, while the automatic layer of annotation encodes linguistic information as features for a machine learning approach to rhetorical status classification.

1 Project Overview

The main aim of our project is to explore techniques for automatic summarisation of texts in the legal domain. A somewhat simplistic characterisation of the field of automatic summarisation is that there are two main approaches, *fact extraction* and *sentence extraction*. The former uses Information Extraction techniques to fill predefined templates which serve as a summary of the document; the latter compiles summaries by extracting key sentences with some smoothing to increase the coherence between the sentences. Our approach to summarisation is based on that of Teufel and Moens (1999a, 2002, henceforth T&M). T&M work on summarising scientific articles and they use the best aspects of sentence extraction and fact extraction by combining sentence selection with information about *why* a certain sentence is extracted—i.e. its rhetorical role: is it, for example, a description of the main result, or is it a criticism of someone else's work? This approach can be thought of as a more complex variant of template filling, where the slots in the template are high-level structural or rhetorical roles (in the case of scientific texts, these slots express argumentative roles such as *goal* and *solution*) and the fillers are sentences extracted from the source text using a variety of statistical and linguistic techniques. With this combined approach the closed nature of the fact extraction approach is avoided without giving up its flexibility: summaries can be generated from this kind of template without the need to reproduce extracted sentences out of context. Sentences can be reordered or suppressed depending on the rhetorical role associated with them.

Taking the work of T&M as our point of departure, we explore the extent to which their approach can be transferred to a new domain, legal texts. We focus our attention on a subpart of the legal domain, namely 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 scientific articles papers that T&M worked with, while remaining challengingly different in many respects.

In this paper we describe the corpus of legal texts that we have gathered and annotated. 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 pro-

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

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

cessed 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. 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.

There will be three layers of annotation in the final version of our corpus, with work on the first two well under way. The first layer is manual annotation of sentences for their rhetorical role. The second layer is automatic linguistic annotation. The third layer is annotation of sentences for 'relevance' as measured by whether they match sentences in handwritten summaries. We describe the first two layers in Sections 2 and 3, and in Section 4 we discuss possible approaches to the third annotation layer.

2 Rhetorical Status Annotation

2.1 Rhetorical Roles for Law Reports

The rhetorical roles that can be assigned to sentences will naturally vary from domain to domain and will reflect the argumentative structure of the texts in the domain. In designing an annotation scheme, decisions must be made about how fine-grained the labels can be and an optimal balance has to be found between informational richness and human annotator reliability. In this section we discuss some of the considerations involved in designing our annotation scheme.

Teufel and Moens' scheme draws on the CARS (Create a Research Space) model of Swales (1990). A key factor in this, for the purposes of summarisation, is that each rhetorical move or category describes the status of a unit of text with respect to the overall communicative goal of a paper, rather than relating it hierarchically to other units, as in Rhetorical Structure Theory (Mann and Thompson, 1987), for example. In scientific research, the goal is to convince the intended audience that the work reported is a valid contribution to science (Myers, 1992), i.e. that it is in some way novel and original and extends the boundaries of knowledge.

Legal judgments are very different in this regard. They are more strongly performative than research reports, the fundamental act being decision. In particular, the judge aims to convince his professional and academic peers of the soundness of his argument. Therefore, a judgment serves both a declaratory and a justificatory function (Maley, 1994). In truth, it does more even than this, for it is not enough to show that a decision is justified: it must be shown to be proper. That is, the fundamental communicative purpose of a judgment is to *legitimise* a decision, by showing that it derives, by a legitimate process, from authoritative sources of law.

Table1 provides an overview of the rhetorical annotation scheme that we have developed for our corpus. The set of labels follows almost directly from the above observations about the communicative purpose of a judgment. The initial parts of a judgment typically restate the facts and events which caused the initial proceedings and we label these sentences with the rhetorical role FACT. By the time the case has come to the House of Lords it will have passed through a number of lower courts and there are further details pertaining to the previous hearings which also need to be restated: these sentences are labelled PROCEEDINGS. In considering the case the law lord discusses precedents and legislation and a large part of the judgment consists in presenting these authorities, most frequently by direct quotation. We use the label BACKGROUND for this rhetorical role. The FRAMING rhetorical role captures all aspects of the law lord's chain of argumentation while the DISPOSAL rhetorical role is used for sentences which indicate the lord's agreement or disagreement with a previous ruling: since this is a court of appeal, the lord's actual decision, either allowing or dismissing the appeal, is annotated as DISPOSAL. The TEXTUAL rhetorical role is used for sentences which indicate structure in the ruling, while the OTHER category is for sentences which cannot be fitted into the annotation scheme. As the frequency column in Table 1 shows, PROCEEDINGS, BACKGROUND and FRAMING make up about 75% of the sentences with the other categories being less frequently attested.

2.2 Manual Annotation of Rhetorical Status

The manual annotation of rhetorical roles is work in progress and so far we have 40 documents fully annotated. The frequency figures in Table 1 are taken from this manually annotated subset of the corpus and the classifiers described in Section 2.3 have been trained and evaluated on the same subset. This 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

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

contains 12,188 sentences and 285,934 words while ours contains 10,169 sentences and 290,793 words.

The 40 judgments in our manually annotated subset were annotated by two annotators using the NITE XML toolkit annotation tool (Carletta et al., 2003). Annotation guidelines were developed by a team including 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 (1980) suggests is the cut-off for good reliability.

2.3 Experiments with Rhetorical Role Classification

Using the manually annotated subset of the corpus we have performed a number of preliminary experiments to determine which classifier and which feature set would be appropriate for rhetorical role classification. A brief summary of the micro-averaged F-score[3] results is given in Table 2 (Detailed results in Hachey and Grover, 2004).

The features with which we have been experimenting for the HOLJ domain are broadly similar to those used by T&M and include 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 feature set includes such features as: (**L**) location of a sentence within the document and its subsections and paragraphs; (**C**) cue phrases; (**E**) whether the sentence contains named entities; (**S**) sentence length; (**T**) average *tf∗idf* term weight; and (**Q**) whether the sentence contains a quotation or is inside a block quote.

Classifier	Features	F-score
C4.5	L	65.4
NB	LCESQ	51.8
Winnow	LCESQT	41.4
SVM	LCESQT	60.6

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

While we still expect to achieve gains over these preliminary scores, our system already exhibits an improvement over baseline similar to that achieved by the T&M system, which is encouraging given that we have not invested any time in developing the hand-crafted cue phrases that proved the most useful feature for T&M, but rather have attempted to simulate these through fully automatic, largely domain-independent linguistic information.

We plan further experiments to investigate the effect of other cue phrase features. For example, subject and main verb hypernyms should allow better

[3] Micro-averaging weights categories by their prior probability as opposed to macro-averaging which puts equal weight on each class regardless of how sparsely populated it might be.

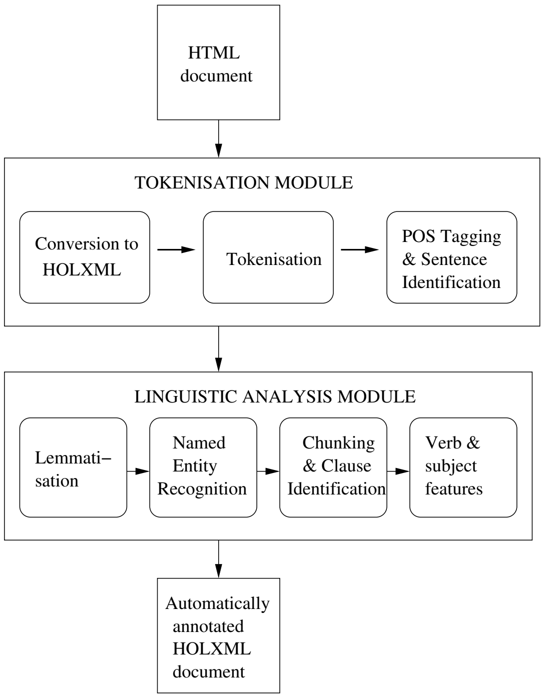

Figure 1: HOLJ Processing Stages

generalisation over *agent* and *type of action* information. We will also experiment with maximum entropy–a machine learning method that allows the integration of a very large number of diverse information sources and has proved highly effective in other natural language tasks–in both classification and sequence modelling frameworks.

3 Automatic Linguistic Markup

One of the aims of our project is to create an annotated corpus of legal texts which will be available to NLP researchers. We encode all the results of linguistic processing as HOLXML 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 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 elements. Once the word tokens have been identified, the next step uses *ltpos* to mark up the sentences and 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 and makes 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. As *morpha* is not XML-aware, we use *xmlperl* (McKelvie, 1999) as a wrapper to incorporate it in the XML pipeline. We use a similar method for other non-XML components.

The next stage, described in Figure 1 as Named Entity Recognition (NER), is in fact a more complex layering of two kinds of NER. Our documents contain the standard kinds of entities familiar from the MUC and CoNLL competitions (Chinchor, 1998; Daelemans and Osborne, 2003), such as person, organisation, location and date but they also contain domain-specific entities. Table 3 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 MUC-7 data set. For some entities, the two approaches provide competing analyses, in which case the domain-specific label is to be preferred since it provides finer-grained information. Wherever there is no competition, C&C entities are marked up and labelled as subtype='fromCC').

During the rule-based entity recognition phase, an 'on-the-fly' lexicon is built from the document header. This includes the names of the lords judging the case as well as the respondent and appellant and it is useful to mark these up explicitly when they occur elsewhere in the document. We create an expanded lexicon from the 'on-the-fly' lexicon containing ordered substrings of the original entry

<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 3: Named Entities in the Corpus

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* occurs in the header as an appellant entity.

The next stage in the linguistic analysis module performs noun group and verb group chunking using *fsgmatch* with specialised hand-written rule sets. The noun group and verb group mark-up plus POS tags provide the relevant features for the next processing step. Elsewhere (Grover et al., 2003), we showed that information about the main verb group of the sentence may provide clues to the rhetorical status of the sentence (e.g. a present tense active verb correlates with BACKGROUND or DISPOSAL). 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 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.

4 Automatic Relevance Annotation

In addition to completing the annotation of rhetorical status, in order to make this a useful corpus for sentence extraction, we also need to annotate sentences for relevance. Many approaches to relevance annotation use human judges, but there are some automatic approaches which pair up sentences from manually created abstracts with sentences in the source text. We survey these here.

As mentioned earlier, our corpus includes hand-written summaries from domain experts. This means that we have the means to relate one to the other to create a gold standard relevance-annotated corpus. The aim is to find sentences in the document that correspond to sentences in the summary, even though they are likely not to be identical in form.

Table 4 summarises five approaches to automatic relevance annotation from the literature. The approaches fall into three basic paradigms based on the methods they use to match abstract content to sentences from the source document: longest com-

Authors	Paradigm	Level
Teufel and Moens (1997)	Longest common subsequence matching	Sentence
Mani and Bloedorn (1998)	IR ($\lvert$overlapping words$\rvert$ + cosine-based similarity metric)	Sentence
Banko et al. (1999)	IR ($\lvert$overlapping words$\rvert$ w/ extra weight for proper nouns)	Sentence
Marcu (1999)	IR (cosine-based similarity metric)	Clause
Jing and McKeown (1999)	HMM (prefers ordered, contiguous words, sentences)	Word

Table 4: Methods for automatic alignment of abstracts with their source documents.

mon subsequence matching, IR-based matching, and hidden Markov model (HMM) decoding. The approaches also differ in the basic alignment unit. The first three operate at the sentence level in the source document, while the fourth and fifth operate at clause level and word level respectively, but can be generalised to sentence-level annotation.

The first approach (Teufel and Moens, 1997) uses a simple surface similarity measure (longest common subsequence of non-stop-list words) for matching abstract sentences with sentences from the source document. This approach accounts for order and length of match, but does not incorporate semantic weight of match terms. Nor does it allow for reordering of matched terms.

Mani and Bloedorn (1998) discuss an IR method using a cosine-based similarity metric over *tf∗idf* scores with an additional term that counts the number of abstract words present in a source sentence. The entire abstract is treated as a query, effectively circumventing the level-of-alignment question during matching. In the end the top *c*% of sentences are labelled as relevant extract sentences.

The third approach (Banko et al., 1999) treats terms from abstract sentences as query terms. These are matched against sentences from the original document. Proper nouns get double weight when computing overlap. This method was found to be better than a version which took the top *tl∗tf*[4] words from each summary sentence and used relative frequency to identify matching source sentences.

Marcu (1999) also describes an IR-based approach. Like Mani and Bloedorn, a cosine-based similarity metric is used. However, the score is used to determine the similarity between the full abstract and the extract. The extract is initialised as the full source document, then clauses are removed in a greedy fashion until the maximum agreement between the abstract and the extract is achieved. Additionally, several post-matching heuristics are applied to remove over-generous matches (e.g. clauses with less than three non-stop words).

[4] *tl∗tf* (*term length term frequency*) is used as an efficient estimation of *tf∗idf* based on the assumption that frequent words in any language tend to be short (i.e. *term length* is proportional to *inverse document frequency*).

A shortcoming of the bag-of-words IR approaches is the fact that they do not encode order preferences. Another approach that accounts for ordering is reported by Jing and McKeown (1999), who use an HMM with hand-coded transition weights. On the other hand, Jing and McKeown do not include features encoding semantic weight such as *tf∗idf*. Like Marcu, correction heuristics are employed to remove certain matches (e.g sentences that contribute less than 2 non-stop words).

Before drawing conclusions, we should consider how appropriate our data is for automatic relevance annotation. Teufel and Moens (2002) discuss the difference between abstracts created by document authors and those created by professional abstractors noting that the former tend to be less systematic and more "deep generated" while the latter are more likely to be created by sentence extraction. T&M quantify this effect by measuring the proportion of abstract sentences that appear in the source document (either as a close variant or in identical form). They report 45% for their corpus of author-created abstracts. Kupiec et al. (1995), by contrast, report 79% for their corpus of professional abstracts. Our summaries are not author created, so we would expect a higher rate of close-variant matches. On the other hand, though our summaries are created by domain expects, they are not necessarily professional abstractors so we might expect more variation in summarisation strategy.

Ultimately, human supervision may be required as in Teufel and Moens (2002), however we can make some observations about the automatic annotation methods above. While IR approaches and approaches that model order and distance constraints have proved effective, it would be interesting to test a model that incorporates both a measure of the semantic weight of matching terms and surface constraints. Since we have named entities marked in our corpus, we could modify the Banko (1999) method by matching terms at the entity-level or simply apply extra weighting to terms inside entities. We might also match entity types or even grounded entities in the case of appellant and respondent.

Also, it may be desirable to annotate sentences with a weight indicating the degree of relevance.

Then, a numerical prediction method might be used in place of classification, avoiding information loss in the model due to discretisation. Also, if the annotation is weighted, then we might incorporate degree of relevance into the evaluation metric.

5 Conclusions

We have presented a new corpus of UK House of Lords judgments which we are in the process of producing. The current version of the corpus can be downloaded from `http://www.ltg.ed.ac.uk/SUM/`. The final version will contain three layers: rhetorical status annotation, detailed linguistic markup, and relevance annotation. The linguistic markup is fully automatic and we anticipate that relevance annotation can be achieved automatically with a relatively high reliability.

Our rhetorical status annotation derives from Swales' (1990) CARS model where each category describes the status of a unit of text with respect to the overall communicative goal of a document. Preliminary experiments using automatic linguistic markup to extract cue phrase features for rhetorical role classification give encouraging results.

Acknowledgments

This work is supported by EPSRC grant GR/N35311. The corpus annotation was carried out by Vasilis Karaiskos and Hui-Mei Liao with assistance using the NITE XML Tolkit (NXT) from Jonathan Kilgour.

References

Michele Banko, Vibhu Mittal, Mark Kantrowitz, and Jade Goldstein. 1999. Generating extraction-based summaries from hand-written summaries by aligning text spans. In *Proceedings of PACLING'99*.

Jean Carletta, Stefan Evert, Ulrich Heid, Jonathan Kilgour, Judy Robertson, and Holger Voormann. 2003. The NITE XML Toolkit: flexible annotation for multimodal language data. *Behavior Research Methods, Instruments, and Computers, special issue on Measuring Behavior*, 35(3).

Nancy A. Chinchor. 1998. *Proceedings of MUC-7*. Fairfax, Virginia.

James R. Curran and Stephen Clark. 2003. Language independent ner using a maximum entropy tagger. In *Proceedings of CoNLL-2003*.

Walter Daelemans and Miles Osborne. 2003. *Proceedings of CoNLL-2003*. Edmonton, Canada.

Claire Grover, Colin Matheson, Andrei Mikheev, and Marc Moens. 2000. LT TTT—a flexible tokenisation tool. In *Proceedings of LREC-2000*.

Claire Grover, Ben Hachey, Ian Hughson, and Chris Korycinski. 2003. Automatic summarisation of legal documents. In *Proceedings of ICAIL'03*.

Ben Hachey and Claire Grover. 2004. A rhetorical status classifier for legal text summarisation. In *Proceedings of the ACL-2004 Text Summarization Branches Out Workshop*.

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

Hongyan Jing and Kathleen R. McKeown. 1999. The decomposition of human-written summary sentences. In *Proceedings SIGIR'99*.

Klaus Krippendorff. 1980. *Content analysis: An Introduction to its Methodology*. Sage Publications.

Julian Kupiec, Jan O. Pedersen, and Francine Chen. 1995. A trainable document summarizer. In *Proceedings of SIGIR'95*.

Yon Maley. 1994. The language of the law. In John Gibbons, editor, *Language and the Law*. Longman.

Inderjeet Mani and Eric Bloedorn. 1998. Machine learning of generic and user-focused summarization. In *Proceedings of AAAI'98*.

William Mann and Sandra Thompson. 1987. Rhetorical structure theory: Description and construction of text structures. In Gerard Kempen, editor, *Natural Language Generation*. Marinus Nijhoff Publishers.

Daniel Marcu. 1999. The automatic construction of large-scale corpora for summarization research. In *Proceedings of SIGIR'99*.

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).

Guido Minnen, John Carroll, and Darren Pearce. 2000. Robust, applied morphological generation. In *Proceedings of INLG'2000*.

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

Erik Tjong Kim Sang and Hervé Déjean. 2001. Introduction to the CoNLL-2001 shared task: clause identification. In *Proceedings of CoNLL-2001*.

John M. Swales. 1990. *Genre Analysis: English in Academic and Research Settings*. Cambridge University Press.

Simone Teufel and Marc Moens. 1997. Sentence extraction as a classification task. In *Proceedings of the ACL/EACL'97 Workshop on Intelligent and scalable Text summarization*.

Simone Teufel and Marc Moens. 1999. Argumentative classification of extracted sentences as a first step towards fexible abstracting. In Inderjeet Mani and Mark T. Maybury, editors, *Advances in Automatic Text Summarization*. MIT Press.

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

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