Extractive Summarisation of Legal Texts

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

April 26, 2006

Abstract. We describe research carried out as part of a text summarisation project for the legal domain for which we use a new XML corpus of judgments of the UK House of Lords. These judgments represent a particularly important part of public discourse due to the role that precedents play in English law. We present experimental results using a range of features and machine learning techniques for the task of predicting the rhetorical status of sentences and for the task of selecting the most summary-worthy sentences from a document. Results for these components are encouraging as they achieve state-of-the-art accuracy using robust, automatically generated cue phrase information. Sample output from the system illustrates the potential of summarisation technology for legal information management systems and highlights the utility of our rhetorical annotation scheme as a model of legal discourse, which provides a clear means for structuring summaries and tailoring them to different types of users.

Keywords: Automatic Text Summarisation, Legal Discourse, Natural Language Processing, Machine Learning, XML, Knowledge Management

1. Introduction

Legal proceedings are an important part of public discourse among the government and its individual and corporate citizens. While corporate entities generally have teams of lawyers to interface with the legal system, individuals often lack this advantage. Automatic summarisation offers a route to providing important information in a format that is more accessible and understandable to the average individual. We present a study of automatic summarisation of English law reports. These form an especially important part of UK legal discourse due to the role that precedents play in common law, which makes access to them essential for a wide range of people. The research we report investigates an approach to automatic summarisation that has the advantage of providing a clear means of tailoring summaries to different types of users from students and other legal novices to solicitors and judges.

Currently, selected judgments are manually summarised by legal experts. While an ultimate goal of legal summarisation would be to provide clear, non-technical summaries of legal judgments, an automatic system using current technology would already enable immediate access to preliminary summaries, and serve as an assisting technology in manual summarisation. Automatic summaries might also be incorporated to provide dynamic, customised content in information retrieval systems. For example, consider a case database

where the user queries using key words or natural language questions and gets back a list of summaries of possible precedent-setting rulings including an indication of the decision. Alternatively, the whole document could be treated as a query in which case a system could actively search for and summarise documents similar to that which the user is currently viewing. These kinds of systems have great utility both for learning law and especially as a research aid for law professionals.

The automatic summarisation literature makes a distinction between *indicative* and *informative* summaries. The former provides a reference function for selecting documents for more in-depth reading while the latter aims to cover all the salient information in the source at some level of detail (Borko and Bernier, 1975; Mani, 2001). As Mani notes, this distinction was developed as a prescriptive guideline for professional abstractors. The distinction is useful nevertheless for defining the intended use of automatic systems. In a domain such as law, where truth preservation is so important, it would be hard to imagine automatically creating informative summaries with current techniques. However, automatic indicative summaries in legal information retrieval systems would be a great boon to legal research and information management.

In the SUM project we have developed a system for summarising legal judgments that is generic and portable and which maintains a mechanism to account for the rhetorical structure of the argumentation of a case—see Figure 1 for a diagrammatic overview of the SUM system. We have been working with judgments of the House of Lords,[1] a domain we refer to here as HOLJ. HOLJ texts contain a header providing structured information, followed by a sequence of sometimes lengthy 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. While this might constitute some part of a summary, it is also necessary to pick out an appropriate number of relevant informative sentences from the unstructured text in the body of the document. Our system uses a mixture of statistical and linguistic techniques which aid the determination of the function or importance of a sentence. Summaries can then be generated by combining sentences extracted from the document and different kinds and lengths of summary can be generated according to the user's needs.

Previous NLP work in the legal domain addresses Information Retrieval (IR) and the computation of simple features such as word frequency. In order to perform summarisation, it is necessary to look at other features which may be characteristic of texts in general and legal texts in particular. These can then serve to build a model for the creation of legal summaries (Moens and

[1] Accessible on the House of Lords website, `http://www.parliament.uk/judicial_work/judicial_work.cfm`

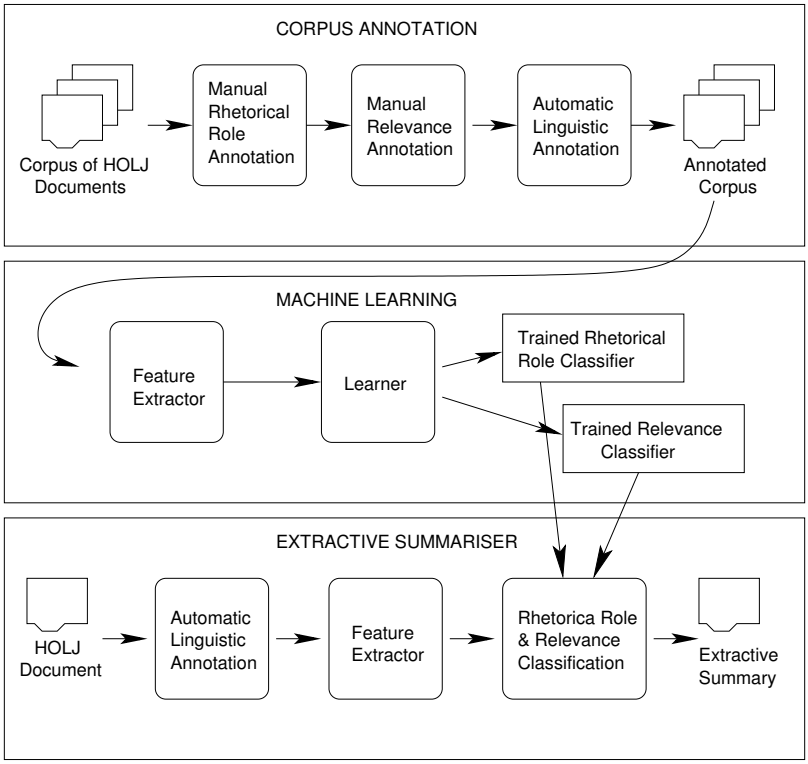

Figure 1. SUM System Architecture

Busser, 2002). In our project, we are developing an automatic summarisation system based on the approach of Teufel and Moens (2002; 1998; 1999; 1997). The core component of this is a statistical classifier which categorises sentences in order that they might be seen as candidate text excerpts to be used in a summary. Useful features include standard IR measures such as word frequency but features which reflect linguistic properties of the sentence are even more informative.

Spärck-Jones (1998) has argued that most practically oriented work on automated summarisation can be classified as either based on *text extraction* or *fact extraction*. When automated summarisation is based on *text extraction*, an abstract will typically consist of sentences selected from the source text, possibly with some smoothing to increase the coherence between the sentences. The advantage of this method is that it is a very general technique, which will work without the system needing to be told beforehand what might be interesting or relevant information. But general methods for identifying abstract-worthy sentences are not very reliable when used in specific domains, and can easily result in important information being overlooked.

When summarisation is based on *fact extraction*, on the other hand, the starting point is a predefined template of slots and possible fillers. These systems extract information from a given text and fill out the agreed template. These templates can then be used to generate shorter texts: material in the source text not of relevance to the template will have been discarded, and the resulting template can be rendered as a much more succinct version of the original text. The disadvantage of this methodology is that the summary only reflects what is in the template.

Teufel and Moens have focused on the domain of scientific articles. This lends itself to automatic text summarisation because documents of this genre tend to be structured in predictable ways and to contain formalised language which can aid the summarisation process (e.g. cue phrases such as 'the importance of', 'to summarise', 'we disagree') (Teufel and Moens 2002, 2000). Their system is an instance of the text extraction approach to summarisation but one which retains a flavour of the fact extraction approach. For long scientific texts it is not feasible to define templates with a wide enough range, however simple sentence selection does not offer much scope for regenerating the text into different types of abstracts. For these reasons, Teufel and Moens experimented with ways of combining the best aspects of both approaches by combining sentence selection with information about *why* a certain sentence is extracted—e.g. is it a description of the main result, or an important criticism of someone else's work? In this way they are able to produce flexible summaries of varying length and for various audiences. 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.

Although there is a significant distance in style between scientific articles and legal texts, we have found it useful to build upon the work of Teufel and Moens and to pursue the methodology of investigating the usefulness of a range of features in determining the argumentative role of a sentence. 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 Teufel and Moens worked with, while remaining challengingly different in many respects.

Although our choice of methodology is designed to test the portability of the Teufel and Moens approach to a new domain, our general aims are comparable with those of the SALOMON project (Moens et al., 1997), which also deals with summarisation of legal texts. The basic scheme of the argumentative structure we define is similar in some ways to that of (Cheung et al., 2001) which was designed for summarisation of Chinese judgment texts (Cheung et al., 2001). The legal summarisation work of Farzindar and

Lapalme (Farzindar and Lapalme, 2004; Farzindar, 2005) is closely related to ours in that they identify thematic structures in legal documents and determine semantic roles of textual units.

Other work in the artificial intelligence and law field tends to favour conventional deep AI techniques. For example, the well known practical and theoretical work of Aleven (1997) and Greenwood et al. (2003) focuses on logical representations of case law with the aim of modelling complex reasoning and argumentation. While the inference possible in this kind of approach allows advanced applications such as sophisticated tutoring systems, it suffers from a development bottleneck at the knowledge acquisition stage.

The SUM system eschews deep AI in favour of a shallow but robust text processing methods which are more akin to the pragmatic approach taken in information retrieval. Our aim is to build a system that can take any in-domain text and create a summary of it without human intervention. Another possible use of the system, then, is as a complementary technology in a legal digital library portal, e.g. Lupo and Batini (2003), where the summaries could be used as concise descriptions of search results.

Apart from the other legal summarisation work mentioned above, our work is most similar to work on automated norm extraction (ANE) and textual case-based reasoning (TCBR). ANE partially automates the task of formalising legislative norms using NLP (van Engers et al., 2004). However, while we employ similar NLP techniques, our system must exist on its own and provide a summary for any document presented. The norm extraction system, on the other hand, works in tandem with humans to ease the task of legal curation. TCBR (Weber et al., 2006) is also similar in underlying techniques. Both the ANE and TCBR work pursue semantic extraction components (i.e. information extraction) that are used to represent propositional content. While we employ named entity recognition and use main verbs to help model the primary propositional content of a sentence, we do not pursue explicit representations of relations.

In the following section we describe our corpus of judgments of the House of Lords and explain the manual and automatic annotation that has been done. In Section 3 we report on the machine learning experiments that we have performed in order to train classifiers for rhetorical role labelling and relevance classification. In Section 4 we explore the issues involved in generating tailored extractive summaries. Finally, in Section 5 we give conclusions and outline a number of directions for future work.

2. Corpus

2.1. Introduction to HOLJ

In this section we describe the corpus of judgments of the House of Lords which we have gathered and annotated. 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.

Our corpus is comprised of 188 judgments from the years 2001–2003 from the House of Lords website. (For a subset of these, manually created summaries are available[2]). 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.[3] 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 are two layers of manual annotation in the corpus. The first is manual annotation of sentences for their rhetorical role and the second is annotation of sentences for 'relevance' as measured by whether they match sentences in hand-written summaries. We take the sentence as the appropriate unit of annotation and processing for both layers—while clause-level annotation would be finer-grained, 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. In the current version of the corpus there are 69 judgments which have been annotated for rhetorical role and a subset of 47 of these have also been annotated for relevance. A third layer of annotation is automatic linguistic annotation, which provides the features which are used by the rhetorical role and relevance classifiers. We

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

[3] While a summarisation system integrated in an electronic publishing framework would benefit from XML standards such as MetaLex (Winkels et al., 2002), in the current work, we see XML primarily as a system-internal data representation.

describe the two manual annotation layers in the following two subsections and then conclude this section with a description of the automatic linguistic annotation.

2.2. MANUAL RHETORICAL STATUS ANNOTATION

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' (2002; 1999) 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 the case of 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.

Figure 2 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

Label	Freq.	Description
FACT	862 (8.5%)	A recounting of 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%)	A description of 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%)	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%)	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 . . . .*
DISPOSAL	935 (9%)	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.*

Figure 2. Rhetorical Annotation Scheme for Legal Judgments

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 Figure 2 shows, PROCEEDINGS, BACKGROUND and FRAMING make up about 75% of the sentences with the other categories being less frequently attested.

The experiments that we report in Section 3 have all been conducted using a subset of 40 of the rhetorically annotated judgments and the frequency figures in Figure 2 are computed over these 40 documents. This subset of the corpus is similar in size to the corpus reported in (Teufel and Moens, 2002): the Teufel and Moens corpus consists of 80 conference articles while ours consists of 40 HOLJ documents. The Teufel and Moens corpus contains 12,188 sentences and 285,934 words while ours contains 10,169 sentences and 290,793 words.

The judgments in our rhetorical role annotated corpus 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 coefficient, N is the number of sentences and k is the number of annotators). This is slightly higher than that reported by Teufel and Moens and above the .80 mark which Krippendorf (1980) suggests is the cut-off for good reliability.

2.3. MANUAL 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. As previously mentioned, our corpus includes handwritten 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.

The literature contains descriptions of a number of methods for automatic alignment of sentences which would be relevant here. These include Teufel and Moens (1997), Mani and Bloedorn (1998), Banko et al. (1999), Marcu (1999) and Jing and McKeown (1999). However, Teufel and Moens (2002) concluded that human annotation was required for their task and thus we chose to perform relevance annotation entirely manually. The resulting aligned corpus, however, is a suitable resource for experimentation with automatic alignment methods. Since we will be making it freely available we hope both to perform experiments of our own and to compare our work with others using the same resource.

To perform the manual annotation, we adjusted our previous use of the NITE XML toolkit annotation tool. In the new task, the summary is converted to XML and each sentence is assigned a unique identifier. The annotator keeps open a view of the summary sentences while interacting with the annotation tool to assign a value to an ALIGN attribute on each document sentence. If a document sentence does not align with a summary sentence then it is left unaltered and it acquires the default assignment ALIGN='NONE'. Note that this method of annotation allows for a summary sentence to be aligned with several document sentences but each document sentence can align with at most one summary sentence. It also allows for the possibility that there may be a summary sentence with which no document sentence aligns.

Kupiec et al. (1995) report similar work in the scientific/technical domain and enumerate ways in which summary sentences may match document sen-

Type	HOLJ Example	
Direct Match	Original:	*Each would exclude a breach of duty that the actor was not aware he was committing.*
	Summary:	*A breach of duty that the actor was not aware he was committing was excluded.*
Direct Join	Original 1:	*Mr Cave received no answer to his letter.*
	Original 2:	*He wrote again on a number of occasions in 1996 but still did not receive an answer.*
	Summary:	*Letters by him to the defendants in 1995 and 1996 had been unanswered.*
Incomplete Match	Original:	*In my judgment, however, the relevant date was the date when the respondent passed its resolution to grant outline planning permission.*
	Summary:	*The better interpretation was that time only ran from the grant of permission.*
Incomplete Join	Original 1:	*It was a claim for damages for being made bankrupt.*
	Original 2:	*PwC are being sued by their own former client, the very person to whom they owed a duty of care.*
	Original 3:	*Ms Mulkerrins' claim is an unusual one, for she complains of PwC's failure to prevent the making of a bankruptcy order against her.*
	Summary:	*LORD MILLET, agreeing with Lord Walker of Gestingthorpe, said that the claimant sought damages from her former professional advisors, the defendants, for having negligently failed to protect her from bankruptcy.*

Figure 3. Document-Summary Sentence Alignment

tences. The simplest case is a direct sentence match where two sentences are identical modulo minor modifications or where they have essentially the same content. Summary sentences are frequently a blend of more than one document sentence, and in simple cases these are direct joins of the source sentences. Examples from our corpus of both of these kinds of direct match are given in the first two rows of Figure 3. Other pairings are less direct and Kupiec et al. describe these as incomplete matches and joins. The second two rows of Figure 3 show examples of incomplete matches from our corpus. Kupiec et al. present statistics showing the distribution of correspondences in their corpus: 79% of their summary sentences have direct matches, 3% are direct joins, 9% are incomplete matches or joins and 9% are summary sentences for which no corresponding sentence can be found.

The task of manually aligning sentences is not an easy one and we did not wish to make it harder by requiring our annotators to record the type of correspondence at the time of annotation. It has, however, proved difficult to make post-hoc categorisations into the classes that Kupiec et al. have defined. The distinction between direct match and incomplete match has proved hard to use with our data, and this may be an indication that the manual summaries in our corpus bear a more complex relationship to the source documents than is the

Sent. 17	*He contends that a blanket policy of requiring the absence of prisoners when their legally privileged correspondence is examined infringes, to an unnecessary and impermissible extent, a basic right recognised both at common law and under the European Convention for the Protection of Human Rights and Fundamental Freedoms, and that the general terms of section 47 authorise no such infringement, either expressly or impliedly.*
Sent. 60	*In principle, such letters are privileged under Article 8.*
Sent. 180	*Article 8.1 gives Mr Daly a right to respect for his correspondence.*
Summary:	*It was similarly in breach of art 8.1 of the Convention for the Protection of Human Rights and Fundamental Freedoms, as scheduled to the Human Rights Act 1998, which gave the applicant a right to respect for his correspondence.*

Figure 4. A Sample Alignment

case with Kupiec et al.'s corpus. One clear source of extra complexity lies in the fact that our source documents are a collection of individual speeches each on the same topic, making the summaries closer to multi-document summaries than is the case with other corpora. Thus one summary sentence will frequently match several document sentences taken from more than one lord's discussion: there may be a direct match with a sentence from one lord but an incomplete match with a sentence from another lord. Typically, such cases arise in sentences which report the overall judgment, i.e. the combined views of all five lords. Even within a single lord's judgment, there is often much repetition with the effect that several document sentences align with a single summary sentence. Thus the summary sentence shown in Figure 4 has been paired by the annotator with sentences 17, 60 and 180 from a single lord's judgment.

Due to the difficulty in categorising the matches according to the scheme shown in Figure 3, we are unable to report statistics which are exactly parallel to the ones given in Kupiec et al. (1995). We can however provide some statistics from our corpus to elucidate the relationship between the summary sentences and the source documents, as shown in Figure 5.

Assuming that the 1-1 and 1-2 matches are likely to correspond to Kupiec et al.'s direct match and direct join categories, we have an approximate total of 61% of pairings falling into these categories as against the 82% reported by Kupiec et al. (1995). There is a correspondingly higher incidence of non-direct matches: 34% as against Kupiec et al.'s 9%. The proportion of unmatched sentences is lower (5% as compared to 9%) though this may be a reflection of the fact that our statistics are approximations rather than absolute measurements.

Number of summary-document pairs:	47
Total Number of summary sentences:	688
Total Number of document sentences:	12,939
Number of aligned summary sentences:	656
Number of unaligned summary sentences:	32
Percentage of summary sentences which are aligned:	95.3%
Number of aligned document sentences:	1660
Number of unaligned document sentences:	11,279
Percentage of document sentences which are aligned:	12.8%

Type of match	No. of sentences	% of total summary sentences
1-1	282	41%
1-2	135	20%
1-3	88	13%
1-4	63	9%
1-5	35	5%
1-6	17	2%
1-7 or more	36	5%
no match	32	5%

Figure 5. Alignment Statistics

2.4. 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 6 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., 2004). 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 classi-

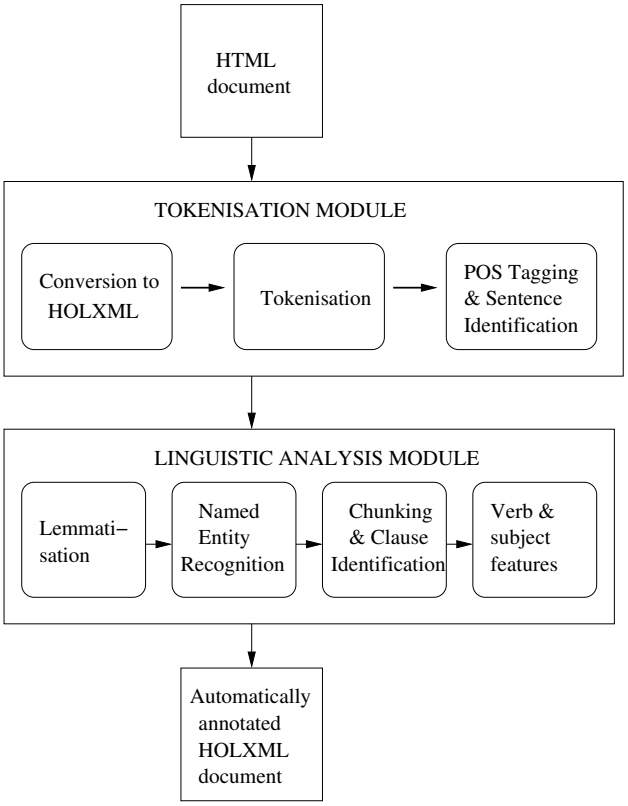

Figure 6. HOLJ Processing Stages

fier. 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 6 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. Figure 7 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 figure 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

<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*

Figure 7. Named Entities in the Corpus

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, 2003b) 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 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 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. Experiments

3.1. Classifying Sentences

Both of our main sub-problems (rhetorical role assignment and sentence extraction/relevance ranking) can be formulated as classification tasks. Following from Kupiec et al. (1995), this has been a standard approach for text extraction summarisation as it provides an empirical method for combining different information sources about the textual unit under consideration (e.g. Teufel and Moens, 1997, Aone et al., 1999). The general processing model is to identify a number of features of sentences and use a corpus to induce an empirical model of how these features interact. Given some new sentence, then, we have a function that takes the feature values as input and outputs the predicted class.

Besides being straightforward to evaluate using standard accuracy measures, classification tasks have the added advantage that there is a range of algorithms for learning and inference available. In the case of rhetorical role assignment, we present experiments with a number of learning algorithms from the Weka package and with maximum entropy models both in a standard classification framework and in a sequence labelling framework. For relevance prediction, we performed experiments with naïve Bayes and maximum

entropy models, adapting the output of the latter to rank sentences for extract-worthiness as well as make hard yes/no decisions about whether a sentence is more like extract or non-extract examples. Before presenting the results of these experiments in sections 3.2 and 3.3, we discuss the various information sources we use as features.

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 Teufel and Moens 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 (2002)). The features that we have been experimenting with for the HOLJ corpus are broadly similar to those used by Teufel and Moens 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 Teufel and Moens's scientific article domain, they did find it to be a useful indicator. Teufel and Moens 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.

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 Teufel and Moens, 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 Teufel and Moens, encodes the percentage of sentence tokens inside an in-line quote and whether or not the sentence is inside a block quote.

Entities. Teufel and Moens 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 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). Teufel and Moens 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 (Grover et al., 2003). 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 Teufel and Moens. A main verb lemma feature simulates Teufel and Moens'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. RHETORICAL STATUS

3.2.1. *Results*

We ran per-feature and cumulative experiments for four classifiers in the *Weka* package: an implementation of Quinlan's (1993) decision tree algorithm (C4.5); an implementation of John and Langley's (1995) algorithm incorporating statistical methods for nonparametric density estimation of continuous variables in a naïve Bayes model (NB); an implementation of Littlestone's (1988) algorithm for mistake-driven learning of a linear separator (Winnow); and an implementation of Platt's (1998) sequential minimal optimisation 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 (Malouf, 2002) and GIS (Darroch and Ratcliff, 1972) estimation al-

[4] Written by Zhang Le: `http://homepages.inf.ed.ac.uk/s0450736/maxent.html`

Table I. Micro-averaged F-score results for rhetorical classification.

	C4.5		**NB**		**Winnow**		**SVM**		**ME**	
	Ind	*Cum*	*Ind*	*Cum*	*Ind*	*Cum*	*Ind*	*Cum*	*Ind*	*Cum*
Cue Phrase	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. Lngth	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. Wds	30.4	59.7	21.2	51.7	25.7	**41.4**	12.0	**60.6**	27.7	**57.5**
Baseline					12.0					

gorithms.[5] We use continuous features for all algorithms except Winnow and maximum entropy. In order to evaluate these, we discretise continuous features using the *Weka* filter based on Fayyad and Irani's (1993) MDL method for discretisation.

Micro-averaged[6] F-scores for each classifier are presented in Table I.[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 Teufel and Moens. 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 because of the skewed distribution of rhetorical categories (67% of sentences fall into the most frequent category).

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).

[5] We used LMVM for early experiments, but all final results 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] F-score is a single measure incorporating precision and recall. All F-scores in this paper weight precision and recall equally to give the harmonic mean, or balanced F-score.

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 Teufel and Moens system achieves an improvement of 18 points. Like Teufel and Moens, 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 (2002) state simply that they experimented with maximum entropy but it did not show significant improvement over naïve Bayes. We hypothesise that this is due to the very carefully constructed feature set optimised for naïve Bayes. Results from Osborne (2002), 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 inter-dependencies 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 has proved highly effective in similar natural language tasks with large, noisy feature sets such as text categorisation, part-of-speech tagging, and named entity recognition. We focus on maximum entropy modelling for the sequencing experiments in the next section.

3.2.2. *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.

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

Table II. Maximum entropy F-score results for rhetorical classification.

	ME		**PL**		**SEQ**	
	Ind	*Cum*	*Ind*	*Cum*	*Ind*	*Cum*
Cue Phrase	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. Lngth	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. Wds	27.7	**57.5**	26.7	**58.1**	31.7	60.8
Baseline			12.0			

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. (Molina and Pla, 2002). In this work, we implement the approach used by Ratnaparkhi (1996; 1998) for part-of-speech tagging and also used by Curran and Clark, Curran and Clark (2003a, 2003b) 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:

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

where $p(y_i|x_i)$ is the normalised 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:

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

where the f_j include the features described in section 3.1 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 (McCallum et al., 2000). However, Ratnaparkhi's (1998) 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.

Table II gives the results for sequencing (SEQ) as well as results for a model incorporating previous labels but no search (PL) and results on the

Table III. Per-category precision, recall and balanced F scores for rhetorical classification using the sequencing model.

Rhet Role	***P*(SEQ)**	***R*(SEQ)**	***F*(SEQ)**	**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	00.0	00.0	00.0	0.5	0.0
Micro Average	61.4	60.9	61.2	–	–

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. Previous label features without search have not improved scores significantly, though, as they did for Teufel and Moens.

Further improvements might be gained by using a search that incorporates following predictions as well as previous predictions or a reranking method, e.g. (Collins, 2000). We might also improve the performance using methods with a different underlying model. The conditional random fields (Lafferty et al., 2001) framework, for instance, avoids biases of directed graphical models such as the approach here and MEMMs by removing the simplifying Markov assumption.

Table III 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 much larger proportion of the average summary than of the source document.

The system performance is roughly equal for all but FACT and TEXTUAL sentences. It performs very well on TEXTUAL sentences because sentences having to do with document structure tend to be formulaic and easy to identify. Also, the average sentence length for TEXTUAL sentences ($\sim$ 8.3) is a reliable indicator, falling far below the overall average of $\sim$ 29.6 words. Conversely, for FACT sentences, the performance suffers because of the het-

erogeneity 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).

3.3. RELEVANCE

3.3.1. *Results*

Evaluation of summaries is a complex and contentious issue. In this section, we present a quick overview of the difficulties of evaluation and some solutions from the literature. We then present a preliminary evaluation using standard accuracy measures. Results reported in this section are obtained from a subset of 47 documents annotated both for rhetorical status and relevance with seven randomly chosen documents withheld for testing.

Despite a long history of summarisation research, the community has not come up with an agreed best practise or produced fully automatic methods for reliable intrinsic evaluation. In fact, it is probably safe to say that the latter cannot be solved without solving the problem of automatic summarisation itself. Detailed evaluation efforts generally incorporate manual scoring of summaries according to a number of qualitative criteria such as coverage of propositional content with penalties for repetition, and linguistic well-formedness (e.g. presence of antecedents for pronouns, proper use of discourse connectives, correct ordering of text units).[9]

While IR accuracy measures are insufficient for evaluating all aspects of the summarisation task, they do allow for a quick, automatic approximation of system performance for extractive summaries that will help us to choose which learning algorithm to work with. Table IV contains precision (*P*), recall (*R*) and F-scores (*F*) for two classifier models: naïve Bayes and maximum entropy. The baseline is created by selecting sentences from the end of the document as described in the section 4.2.

[9] The Document Understanding Conferences `http://duc.nist.gov/` run by the American National Institute of Standards and Technology and the Text Summarization Challenge `http://lr-www.pi.titech.ac.jp/tsc/index-en.html` run by the NII-NACSIS Test Collection for IR Systems Project in Japan are examples of large-scale, formal evaluations.

Table IV. Precision, recall and balanced F scores for NB and ME *Yes* (i.e. sentence is a good summary sentence) predictions.

	***P*(NB)**	***R*(NB)**	***F*(NB)**	***P*(ME)**	***R*(ME)**	***F*(ME)**
Cue Phrase	55.1	3.3	6.2	0.0	0.0	0.0
Location	32.9	23.0	27.1	75.1	15.8	26.1
Entities	31.3	27.2	29.1	76.3	16.0	26.5
Sent. Lngth	30.5	28.9	29.7	73.3	15.9	26.1
Quotations	30.2	29.3	29.7	71.8	16.7	27.1
Them. Wds	31.7	30.7	**31.2**	71.4	16.9	**27.3**
Baseline			46.7/16.0/23.8			

Though none of the feature sets perform well individually, all contribute positively to the cumulative scores with the exception of sentence length for maximum entropy and quotation for naïve Bayes. Both classifiers perform significantly better than baseline and F-scores for the best feature combinations are roughly similar to the partial results reported in (Teufel and Moens, 2002). While the best naïve Bayes F-score is higher, precision (30.3%) is far lower than the best maximum entropy model (71.4%). As high precision is a desirable characteristic when we consider the fact (discussed in the next section) that relevance prediction is perhaps better conceived of as a ranking task than a classification task, we use ME for the remaining experiments.

3.3.2. *Prediction Versus Ranking*

A basic aspect of summarisation system design, especially a system that needs to be flexible enough to suit various user types, is that the size of the summary will be variable. For instance, students may need a 20 sentence summary, containing e.g. quite detailed background information, to get the same information a judge would get from a 10 sentence summary. Furthermore, any given user might want to request a longer summary for a certain document.

One way to achieve this is to apply some sort of ranking to document sentences rather than a binary decision over each sentence. In our case, we want to give a rating of *how* extract-worthy a sentence is instead of making a hard yes/no decision about whether each sentence is an extract sentence or not. We can then use this rating to add the highest ranking sentences to the summary first.

We modified a maximum entropy classifier so that its positive prediction values can be directly compared by outputting the value from the exponential equation $exp(\sum_j \lambda_j f_j(x_i, y_i))$ without multiplying by the normalising factor

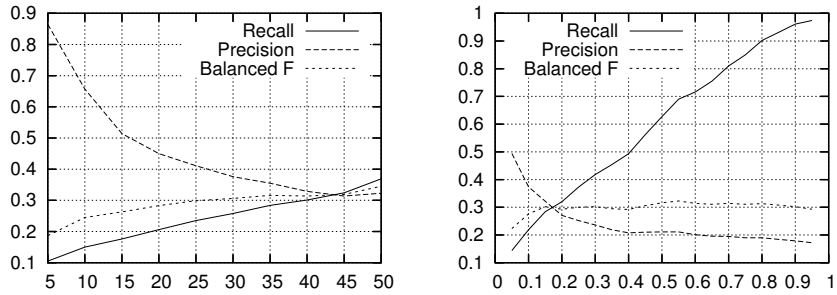

Figure 8. Accuracy plotted against summary size

Table V. Precision, recall and balanced F scores by rhetorical category.

Rhet Role	***P*(ME)**	***R*(ME)**	***F*(ME)**	**DocDist**	**SumDist**
FACT	0.0	0.0	0.0	8.5	10.3
PROCEEDINGS	39.3	12.4	18.8	24.0	18.4
BACKGROUND	0.0	0.0	0.0	27.5	10.2
FRAMING	25.0	6.0	9.6	23.0	30.0
DISPOSAL	79.2	48.7	60.3	9.0	31.1
TEXTUAL	33.3	100	50.0	7.5	0.2
OTHER	0.0	0.0	0.0	0.5	0.0
Micro Average	51.4	17.6	26.3	–	–

$\frac{1}{Z(x_i)}$. Figure 8 shows how precision, recall and F-score performance varies for different absolute summary sizes (top) and for different compression rates (as a percentage of the total source document size in sentences, bottom).

Note that this version of the system does not exert any explicit control over the number of sentences of each rhetorical category that will appear in the summary. As we saw in Table III, the distribution of rhetorical categories in the gold standard extractive summaries is not uniform nor is it the same as the distribution in the source documents. Table V gives a breakdown of scores for each rhetorical category with an absolute summary size of 15. The source document and summary distributions of rhetorical categories are repeated in the rightmost columns.

The scores for FACT and BACKGROUND sentences in Table V illustrate a problem with ranking alone and help to motivate the rhetorical classification. The primary reason for the low scores for FACT and BACKGROUND sentences is the fact that these tend to get low relevance ranking with respect to e.g. DISPOSAL sentences. The design for the final system includes rhetor-

ical profiles that indicate the amount of each type of sentence a user needs. Thus, the rhetorical classification will help us ensure that the summary has some background material if necessary. This will also improve recall scores for categories like BACKGROUND and FACT.

The sentence extraction system performs very well on the most important rhetorical category, DISPOSAL, which makes up nearly one third of the gold standard extracts. DISPOSAL sentences are more important than their document distribution might suggest as they contain the final decisions concerning the appeal. Table V also helps to illustrate the utility of rhetorical status classification. Clearly, ranking alone is not enough as some rhetorical categories are inherently more extract-worthy according to our measure (e.g. FACT and FRAMING both get very low recall). Rhetorical status information will allow us to create a template that will help get the correct distribution of discourse information in the template.

Finally, we have performed preliminary experiments with lemmatised token and hypermym cue phrase features. NB attains F-scores of 26.5 and 27.4 respectively with the addition of lemmas and hypernyms to the cue phrase features, while ME attains F-scores of 21.6 and 22.0. These results are promising and suggest that the overall performance will be improved with further engineering of features based on lexical items. Again, as for the results of rhetorical classification, we find these results encouraging given that the cue phrases consist of fully automatic, largely domain-independent linguistic information.

4. Summary Strategies

4.1. Manual Summaries from ICLR

In Section 2.3 we described the manual annotation for relevance where sentences in the source documents were paired with sentences in the manually produced summaries from the ICLR website. In Table 5 we showed some statistics about the relative sizes of the documents and their summaries and about the ways in which the sentences matched one another. We left it until this section to comment on the nature of the ICLR summaries and to discuss the kinds of summaries that an automatic system might produce as compared to the manually created ones.

The ICLR summaries are on average 15.5 sentences long and the average sentence length is 38 words. This compares to an average number of sentences in the source documents of 525 sentences with an average sentence length of 29 words. From this we can see that the summaries are highly compressed versions of the originals which tend to pack information into longer than average sentences. The manual summaries also tend to follow a

highly stylised format, especially for the first two or three sentences. The opening sentence(s) make an assertion of fact and the following sentence starts with a stock expression which is usually a variant of "*The House of Lords so held in allowing/dismissing an appeal ...*". The remainder of this key sentence contains a very compressed synopsis of all of the court cases and decisions which are precursors to the House of Lords judgment. These compressed synopses are often extremely hard for a lay person to follow. An example of this structure can be seen in the first four sentences of the ICLR abstractive summary of the case used as an example in the appendix:

> *"The House of Lords had jurisdiction to entertain an appeal against a refusal of the Court of Appeal, on a renewed application under RSC Ord 59, r 14(3), of permission to apply for judicial review. Grounds for applying for judicial review of a planning permission first arose, under RSC Ord 53, r 4(1), on the grant of permission rather than on the resolution to grant it. The House of Lords so held in allowing an appeal by Sonia Maria Burkett from the Court of Appeal (Sedley, Ward and Jonathan Parker LJJ) which had on 13 December 2000 dismissed a renewed application by her and her late husband for permission to apply for judicial review of a grant by the local planning authority to the interested party, St George West London Ltd, of an outline planning permission. Richards J had refused their application on the ground of delay."*

When we examine how these opening sentences of the summaries are paired in our annotated corpus with source document sentences, we see that these are sentences which map to a high number of source sentences, usually from more than one lord's speech.

The main body of a manual summary tends to be simpler and the pairings between summary and source sentences are more likely to be one-to-one. The type of match is also more likely to be a direct or close match and the sentences tend to be taken from the main lord's speech in the order in which they occur in the source. Thus, this middle part of a summary bears the closest resemblance to the type of extractive summary which our system is designed to produce.

The final few sentences of a manual summary tend to provide an overview of the opinions of the lords whose judgments were very short. The final two sentences of the summary of the case quoted above and in the appendix are as follows:

> *"LORD HOPE OF CRAIGHEAD delivered an opinion agreeing with Lord Steyn in allowing the appeal. LORD MILLETT and LORD PHILLIPS OF WORTH MATRAVERS agreed with Lord Slynn and Lord Steyn."*

From this brief description of the manual summaries, it is clear that an automatic extractive system will not be capable of producing summaries in

exactly the same style. However, a decomposition of some of the more compressed parts of a manual summary into an uncompressed list of extracted sentences might be just as indicative of content and occasionally more comprehensible to a non-expert reader.

In Section 3.3 we gave statistics for the relative distribution of rhetorical roles among the sentences that are aligned with summary sentences. From Table V, it can be seen that DISPOSAL sentences are much more frequent in summaries than in the source documents (31% in summaries as compared to 9% in the source documents). FACT sentences and FRAMING sentences also occur more frequently, while PROCEEDINGS and BACKGROUND sentences occur less frequently. We can use information about these comparative distributions to inform the design of templates for generating different kinds of extractive summary, as discussed in the remainder of this section.

4.2. Preliminary Discourse Structuring

The questions that need to be addressed when creating an extractive summary strategy can be roughly separated into issues having to do with *the size of the summary*, *the way sentences are selected*, and *how the summary is structured*. We start this section by presenting several summaries before discussing some of the alternatives in creating and structuring our summaries. This section presents an example summary as a running example and discusses the various system design issues.

Appendix Sections A, B and C contain a gold standard extractive summary, a baseline summary, and a summary from our preliminary system. The gold standard summary contains all sentences that are aligned with summary sentences as described in Section 2.3 and these are ordered to reflect the order of the corresponding sentences in the manual abstractive summaries. The baseline and system summaries are formed respectively by selecting a number of sentences from the end of each speech and by selecting sentences according to the ranking method described above in section 3.3.2. The reader may also want to refer to the system architecture diagram (Figure 1).

There are some obvious problems in the system summary, especially in the area of discourse smoothing. Sentence number 183, for example, details an aspect of a previous hearing on the case, but also serves to introduce a quotation. However, though the discursive fit is not quite right, we do glean useful and important information about the decision on this case. Furthermore, the improvement over the baseline is evident (refer to the speech of Lord Hope of Craighead in the appendix for a concise example) and illustrates the potential of this type of application within a legal information retrieval and document management system, even without being discursively smooth.

With respect to *the size of the summary*, as with the other summary strategy choices, ultimately we want to base our decision on some measure of

utility for the target users. A glance at the compression plots in Figure 8, though, allows some interesting observations. We can see from the plots that precision and recall are balanced at around 45 sentences in terms of absolute size or around 17% in terms of the proportion of the source document. However, this illustrates the contention alluded to earlier between automatic evaluation measures such as precision and recall and the fact that the final system needs to be optimised with specific users and tasks in mind. While still providing the potential for a substantial time savings to the user, a summary of 45 sentences is on the long side e.g. for an indicative summary that might be used as a snippet returned from a query of a legal database.

For the current work, we have chosen an absolute summary length of 15. This is somewhat arbitrary; however, it is approximately the average number of sentences in the manual abstracts. And, while this is probably too short to capture all of the information in the gold standard abstracts due to the fact that abstract sentences are sometimes aligned with more than one document sentences containing different propositional content, it suits the current illustrative purposes in that it is not too long, a constraint which will be equally important in the final system design. We chose an absolute summary length as opposed to a summary length relative to the original document size because the length of the manual abstracts is highly uniform relative to the size of the source documents and because this is a desirable property for the initial text presented by an information retrieval system.

With respect to *the way sentences are selected*, both the system we present and the baseline select sentences first from lords that have longer speeches. They both ensure that at least one sentence is selected for each lord. And they select sentences from each lord in proportion to the size of the speech in the source document. The method of selection is the biggest variable in this category. Our best system summaries to date come from the ranking approach based on the unnormalised yes-prediction value from the maximum entropy model that is described in section 3.3.2.

We have also considered several baseline selection methods. One possible baseline for automatic summarisation is random selection. However, due to the correlation between logical structuring and order of presentation in most types of formal prose, a baseline that simply selects sentences from the periphery of certain easily identified text units (e.g. documents, paragraphs) provides a baseline that in some domains, especially newswire, proves difficult to improve on. Though simple, this approach is reliable enough to be incorporated into popular enterprise systems (e.g. Wasson, 1998).

While putting a synopsis of the document in the first paragraphs (the news 'lead') is not an explicit composition strategy in writing legal judgments, the most important sentences in our corpus do tend to occur at the document periphery. Almost without exception, law lords finish their speeches with a few paragraphs containing an explicit statement of whether the appeal should

be allowed. Therefore, our working baseline is to take sentences from the end of the lord's speeches.

A further important option for selection that we have not yet implemented is to select sentences according to some prescribed distribution of rhetorical categories, an obvious choice being the distribution from the gold standard summaries. As mentioned above (Section 3.3.2), sentences from different rhetorical categories have different levels of extract-worthiness. Having the rhetorical categories separated will allow us to create summaries with differing amounts of sentences from given rhetorical categories with a single model of relevance. Conversely, it also makes it possible to create different models of relevance for different rhetorical categories.

Finally, with respect to *how the summary is structured*, the baseline and system summaries both order speeches according to size, presenting speeches with more sentences first. This is a logical choice as the discourse between the judges is such that there is normally one primary speech (or a couple of primary speeches). The other lords generally have a chance to read a draft of this speech and, subsequently, their speeches are in some sense responses either agreeing with or arguing against the 'main' speech (or speeches)

As alluded to in Section 3.3.2, there is also the possibility of grouping and ordering sentences by rhetorical status. Lord Hope of Craighead's speech in Appendix Section C is an example where rhetorical status information provides the means to create a logically more coherent summary. Regardless of the fact that the DISPOSAL sentence came first in the source document, we have been able to move this concluding remark to its prototypical location at the end of the speech. This will become even more important when rhetorical templates are used to control the distribution of the argumentative zones in the summaries and when user- and task-focused summaries are considered.

5. Conclusions and Future Work

We have presented work on the automatic summarisation of legal texts. In the context of English law, legal proceedings are an extremely important part of public discourse and automatic summarisation offers a route for providing important information in a format that is more accessible and understandable. While the automatic creation of clear, non-technical, linguistically well-formed summaries is still in the future, the system we present already enables a number of useful applications for managing legal information such as immediate access to preliminary summaries of judgments, tools to assist manual summarisation, and dynamic, customised information retrieval.

We have introduced a new corpus designed for research into legal text summarisation and legal discourse with 3 levels of annotation: rhetorical status, relevance and linguistic markup. The novelty and utility of this resource

lies in the fact that it provides the text summarisation community with a new common resource allowing comparable research in an interesting and valuable domain.

We presented experiments for the sentence-level classification of rhetorical status using a number of machine learning algorithms that have previously shown good performance on natural language tasks. Among these, support vector machines and maximum entropy sequence models prove to be the best suited to our task. Results are especially encouraging considering the fact that this achieves state-of-the performance without hand-crafted cue phrase features. We introduced a robust and generic method for capturing cue phrase information based on widely available linguistic analysis tools.

We presented favourable sentence extraction results in classification and ranking frameworks. The classification system achieves a significant improvement over the baseline. A breakdown of sentence extraction scores by rhetorical category shows that rhetorical information is an important means of controlling argumentative distribution of sentences in an extractive summarisation system. Preliminary scores for cue phrase feature sets including lemma and hypernym information promise further improvements in accuracy.

We also discussed the structure of the manual abstractive summaries from ICLR. We presented an example of the extractive gold standard, the baseline and the system summaries. Comparison shows the potential of the extractive approach to summarisation for applications including immediate access to preliminary case summaries, assisting in manual summarisation and providing automatic indicative summaries for information retrieval systems allowing the legal researcher to quickly locate relevant precedents.

In current work, we are developing a user study which will allow us to assess value of our system for the information retrieval task referred to throughout this paper. A hypothetical case will be presented to subjects, with a number of possible precedent-setting cases. The possible precedents will be presented in various formats, including: our system summaries (tailored to different types of reader and visualised in various ways); the original full text; and the gold standard summaries. Levels of agreement between subjects, and between subjects' and experts' classifications of cases, will allow us to quantify the utility of our system for a group of real users.

Finally, we suggest a number of ideas for research that can be carried out using the HOLJ corpus. As mentioned in Section 2.3, while we have not yet performed experiments with automatic evaluation, the relevance annotations on our corpus are a valuable resource for this line of research. The corpus can also be used for further experiments with discourse structuring such as Lapata's (2003) probabilistic approach, the corpus can be adapted to investigate methods for minimising repetition of propositional content in extracted summary sentences, and it can be used for named entity bootstrapping using a noisy seed set.

Appendix

A. Gold Standard Extractive Summary

This section contains a sample gold standard extractive summary formed by selecting all document sentences that were aligned with a sentence from the manual abstractive summary. It is ordered to reflect the order of the corresponding abstract sentences. The columns contain the gold standard rhetorical role assigned by the annotators, the sentence number in the source document, the relevance ranking assigned by our system and the sentence text.

The case is *Regina v. London Borough of Hammersmith and Fulham and Others, Ex P Burkett and Another*, heard on 23 May 2002. The original document is available at `http://www.publications.parliament.uk/pa/ld200102/ldjudgmt/jd020523/burket-1.htm`.

Rhet Role	Sent	Rank	Text
			Lord Steyn
DISPOSAL	376	1.02	For all these reasons I am satisfied that the words " from the date when the grounds for the application first arose " refer to the date when the planning permission was granted .
DISPOSAL	378	24	It follows that in my view the decisions of Richards J and the Court of Appeal were not correct .
DISPOSAL	398	2.91	For these reasons , as well as the reasons given by my noble and learned friend Lord Slynn of Hadley , I would allow the appeal and remit the matter for decision by the High Court on the substantive issues .
FACT	151	0.29	On 6 April 2000 Mr and Mrs Burkett submitted an application for permission to apply the judicial review .
FACT	163	0.29	Acting on the authority of the resolution of 15 September 1999 the director of the environment Department of the local authority granted outline planning permission on the same day .
PROCEEDINGS	183	3.08	In the judgment of the court (Ward , Sedley and Jonathan Parker LJJ) , given on 13 December , this argument is dismissed on the following ground (paragraph 8) :
PROCEEDINGS	35	0.19	Mrs Burkett and her late husband applied for judicial review .
PROCEEDINGS	39	0.23	After a full inter partes hearing the Court of Appeal refused permission to seek judicial review on grounds of delay and dismissed the appeal .

Rhet Role	Sent	Rank	Text
			Lord Steyn (continued)
PROCEEDINGS	167	0.41	On 29 June 2000 Richards J accepted after reading what he described as detailed skeleton arguments from the local authority and the developer , but without hearing oral arguments from them , that the grounds for judicial review were , on the merits , arguable but refused permission on the grounds of delay .
PROCEEDINGS	37	0.99	He refused permission on the grounds of delay .
BACKGROUND	57	0.21	Lord Hoffmann observed , at p 18B , that a renewed application to the Court of Appeal under RSC Ord 59 , r 14 (3) is a true appeal with a procedure adapted to its ex parte nature .
DISPOSAL	71	0.28	It follows that the House has jurisdiction to grant leave to appeal against a refusal by the Court of Appeal of permission to apply for judicial review .
FRAMING	66	0.31	A material difference , however , is that in the present case the Court of Appeal granted leave to appeal and heard the appeal .
FRAMING	67	0.24	It would be extraordinary if in such a case the House had no jurisdiction .
DISPOSAL	70	0.34	In my view the conclusion is inescapable that Lord Diplock 's extempore observation was not correct .
FRAMING	335	0.31	It weighs in favour of a clear and straightforward interpretation which will yield a readily ascertainable starting date .
FRAMING	367	0.14	By contrast if the better interpretation is that time only runs under Ord 53 , r 4 (1) , from the grant of permission the procedural regime will be certain and everybody will know where they stand .
FRAMING	337	0.19	Secondly , legal policy favours simplicity and certainty rather than complexity and uncertainty .
FRAMING	345	0.29	Unfortunately , the judgment in the Greenpeace case and the judgment of the Court of Appeal , although carefully reasoned , do not produce certainty .
PROCEEDINGS	172	0.23	In my judgment , however , the relevant date was the date when the respondent passed its resolution to grant outline planning permission .
			Lord Slynn of Hadley
DISPOSAL	13	0.33	It seems to me clear that because someone fails to challenge in time a resolution conditionally authorising the grant of planning permission , that failure does not prevent a challenge to the grant itself if brought in time , i e from the date when the planning permission is granted .
DISPOSAL	20	1.35	I would accordingly allow the appeal and remit the substantive question to the High Court for decision .
FACT	7	0.41	On 12 May 2000 planning permission was actually granted .
PROCEEDINGS	6	0.39	On 6 April 2000 the appellant applied for leave to move for judicial review of that decision .

Rhet Role	Sent	Rank	Text
			Lord Slynn of Hadley (continued)
PROCEEDINGS	10	1.20	Richards J and the Court of Appeal refused permission on the ground that the application was out of time .
DISPOSAL	12	0.42	In my opinion , for the reasons given by Lord Steyn , where there is a challenge to the grant itself , time runs from the date of the grant and not from the date of the resolution .
			Lord Hope of Craighead
PROCEEDINGS	411	0.27	The fact that the Court of Appeal granted permission to the applicants to appeal from the decision of Richards J shows that the decision of the judge to refuse permission was not treated as final and conclusive and without appeal in that court .
FRAMING	402	0.27	Subject only to some observations which I should like to add to what he has said on the questions of jurisdiction and promptitude , I agree with it .
DISPOSAL	403	1.25	I too would allow the appeal .
			Lord Phillips of Worth Matravers
DISPOSAL	457	0.75	For the reasons they give I too would allow the appeal .
			Lord Millet
DISPOSAL	453	0.75	For the reasons they give I too would allow the appeal .

B. Baseline Summary (Last Sentences from Each Speech)

This section contains a sample baseline summary formed by selecting a number of sentences from the end of each lords' speech. Speeches are ordered by their number of sentences they contain in the summary and sentences within lords are left in their document order. The columns contain the predicted rhetorical role assigned by our system, the sentence number in the source document, the relevance ranking assigned by our system and the sentence text.

The case is *Regina v. London Borough of Hammersmith and Fulham and Others, Ex P Burkett and Another*, heard on 23 May 2002. The original document is available at `http://www.publications.parliament.uk/pa/ld200102/ldjudgmt/jd020523/burket-1.htm`.

Rhet Role	Sent	Rank	Text
			Lord Steyn
FRAMING	389	0.20	Secondly , there is at the very least doubt whether the obligation to apply " promptly " is sufficiently certain to comply with European Community law and the Convention for the Protection of Human Rights and Fundamental Freedoms (1953) (Cmd 8969) .
FRAMING	390	0.30	It is a matter for consideration whether the requirement of promptitude , read with the three months limit , is not productive of unnecessary uncertainty and practical difficulty .
FRAMING	391	0.16	Moreover , Craig , Administrative Law , 4th ed , has pointed out , at p 794 :
BACKGROUND	392	0.23	" The short time limits may , in a paradoxical sense , increase the amount of litigation against the administration .
BACKGROUND	393	0.17	An individual who believes that the public body has acted ultra vires now has the strongest incentive to seek a judicial resolution of the matter immediately , as opposed to attempting a negotiated solution , quite simply because if the individual forbears from suing he or she may be deemed not to have applied promptly or within the three month time limit "
FRAMING	394	0.18	And in regard to truly urgent cases the court would in any event in its ultimate discretion or under section 31 (6) of the 1981 Act be able to refuse relief where it is appropriate to do so : see Craig , Administrative Law , 4th ed , 794 .

Rhet Role	Sent	Rank	Text
			Lord Steyn (continued)
FRAMING	395	0.22	The burden in such cases to act quickly would always be on the applicant : see Jones and Phillpot , " He Who Hesitates is Lost : Judicial Review of Planning Permissions " [2000] JPL 564 , at 589 .
TEXTUAL	396	0.19	XIII .
TEXTUAL	397	0.19	Disposal .
DISPOSAL	398	2.91	For these reasons , as well as the reasons given by my noble and learned friend Lord Slynn of Hadley , I would allow the appeal and remit the matter for decision by the High Court on the substantive issues .
			Lord Hope of Craighead
FRAMING	448	0.37	They provide a sufficiently clear and workable rule for the avoidance of undue delay in the bringing of these applications , as experience of the operation of judicial review in Scotland has shown .
DISPOSAL	449	0.27	I do not think that it would be incompatible with his Convention rights for an applicant who must be taken to have acquiesced in the decision which he seeks to bring under review , or whose delay has been such that another interested party may be prejudiced , to be told that his application cannot proceed because he has delayed too long in bringing it .
			Lord Phillips of Worth Matravers
DISPOSAL	457	0.75	For the reasons they give I too would allow the appeal .
			Lord Millet
DISPOSAL	453	0.75	For the reasons they give I too would allow the appeal .
			Lord Slynn of Hadley
DISPOSAL	20	1.35	I would accordingly allow the appeal and remit the substantive question to the High Court for decision .

C. System Summary (Ranking by YES Confidence Score)

This section contains a sample system summary formed by selecting the sentences with the highest relevance ranking from each lord. Lords' speeches are ordered by their size and sentences within speeches are ordered by a rhetorical structuring strategy that puts FACT sentences first. It groups PROCEEDINGS, BACKGROUND and FRAMING next as BACKGROUND can be used in support of both PROCEEDINGS and FRAMING sentences. And finally, DISPOSAL sentences are presented. The columns contain the predicted rhetorical role assigned by our system, the sentence number in the source document, the relevance ranking assigned by our system and the sentence text.

The case is *Regina v. London Borough of Hammersmith and Fulham and Others, Ex P Burkett and Another*, heard on 23 May 2002. The original document is available at `http://www.publications.parliament.uk/pa/ld200102/ldjudgmt/jd020523/burket-1.htm`.

Rhet Role	Sent	Rank	Text
			Lord Steyn
PROCEEDINGS	40	1.38	The Court of Appeal refused leave to appeal to the House of Lords .
PROCEEDINGS	43	1.58	In In re Poh the judge had refused leave to apply for judicial review .
PROCEEDINGS	44	1.37	The applicant appealed ex parte by originating motion to the Court of Appeal who refused leave .
PROCEEDINGS	166	1.07	On 18 May 2000 Newman J refused permission to apply for judicial review on the papers in respect of both delay and merits .
PROCEEDINGS	178	2.06	In the circumstances , and particularly in the absence of a clear warning by the applicants to the local authority , the judge refused to extend time .
PROCEEDINGS	183	3.08	In the judgment of the court (Ward , Sedley and Jonathan Parker LJJ) , given on 13 December , this argument is dismissed on the following ground (paragraph 8) :
PROCEEDINGS	194	5.08	The Court of Appeal [2001] JPL 775 dismissed the appeal and refused leave to appeal to the House of Lords .
FRAMING	302	2.45	And in strict law it could be dismissed .
DISPOSAL	376	1.02	For all these reasons I am satisfied that the words " from the date when the grounds for the application first arose " refer to the date when the planning permission was granted .

Rhet Role	Sent	Rank	Text
			Lord Steyn (continued)
DISPOSAL	398	2.91	For these reasons , as well as the reasons given by my noble and learned friend Lord Slynn of Hadley , I would allow the appeal and remit the matter for decision by the High Court on the substantive issues .
			Lord Hope of Craighead
FRAMING	437	1.32	But decisions as to whether a petition should be dismissed on the ground of delay are made in the light of the circumstances in which time was allowed to pass .
DISPOSAL	403	1.25	I too would allow the appeal .
			Lord Phillips of Worth Matravers
DISPOSAL	457	0.75	For the reasons they give I too would allow the appeal .
			Lord Millet
DISPOSAL	453	0.75	For the reasons they give I too would allow the appeal .
			Lord Slynn of Hadley
DISPOSAL	20	1.35	I would accordingly allow the appeal and remit the substantive question to the High Court for decision .

References

Aleven, V.: 1997, 'Teaching Case-Based Argumentation through a Model and Examples'. Ph.D. thesis, University of Pittsburgh, Pittsburgh, PA, USA.

Aone, C., M. E. Okurowski, J. Gorlinsky, and B. Larsen: 1999, 'A Trainable Summarizer with Knowledge Acquired from Robust NLP Techniques'. In: I. Mani and M. T. Maybury (eds.): *Advances in Automatic Text Summarization*. Cambridge, Massechusetts: MIT Press, pp. 71–80.

Banko, M., V. Mittal, M. Kantrowitz, and J. Goldstein: 1999, 'Generating Extraction-Based Summaries from Hand-Written Summaries by Aligning Text Spans'. In: *Proceedings of the 4th Meeting of the Pacific Association for Computational Linguistics*. Waterloo, Ontario, Canada.

Borko, H. and C. L. Bernier: 1975, *Abstracting concepts and methods*. New York: Academic Press.

Carletta, J., S. Evert, U. Heid, J. Kilgour, J. Robertson, and H. Voormann: 2003, 'The NITE XML Toolkit: flexible annotation for multi-modal language data'. *Behavior Research Methods, Instruments, and Computers, special issue on Measuring Behavior* **35**(3), 353–363.

Cheung, L., T. Lai, B. Tsou, F. Chik, R. Luk, and O. Kwong: 2001, 'A Preliminary Study of Lexical Density for the Development of XML-based Discourse Structure tagger'. In: *Proceedings of the 1st NLP and XML Workshop*. Tokyo, Japan.

Chinchor, N. A.: 1998, *Proceedings of the 7th Message Understanding Conference*. Fairfax, Virginia.

Collins, M.: 2000, 'Discriminative Reranking for Natural Language Parsing'. In: *Proceedings of the 17th International Conference on Machine Learning*. Stanford University, CA, USA.

Curran, J. R. and S. Clark: 2003a, 'Investigating GIS and Smoothing for Maximum Entropy Taggers'. In: *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*. Budapest, Hungary.

Curran, J. R. and S. Clark: 2003b, 'Language Independent NER using a Maximum Entropy Tagger'. In: *Proceedings of the Conference on Computational Natural Language Learning*. Edmonton, Alberta, Canada.

Daelemans, W. and M. Osborne: 2003, *Proceedings of the Conference on Computational Language Learning*. Edmonton, Alberta, Canada.

Darroch, J. N. and D. Ratcliff: 1972, 'Generalized Iterative Scaling for Log-Linear Models'. *The Annals of Mathematical Statistics* **43**(5), 1470–1480.

Farzindar, A.: 2005, 'Résumé Automatique de Textes Juridiques'. Ph.D. thesis, Université de Montréal and Université Paris-Sorbonne.

Farzindar, A. and G. Lapalme: 2004, 'Legal Text Summarization by Exploration of the Thematic Structure and Argumentative Roles'. In: *Proceedings of the ACL-2004 Text Summarization Branches Out Workshop*. Barcelona, Spain.

Fayyad, U. and K. Irani: 1993, 'Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning'. In: *Proceedings of the 13th International Joint Conference on Artificial Intelligence*. Chambéry, France.

Greenwood, K., T. Bench-Capon, and P. McBurney: 2003, 'Towards a Computational Account of Persuasion in Law'. In: *Proceedings of the 9th International Conference on Artificial Intelligence and Law*. Edinburgh, Scotland.

Grover, C., B. Hachey, I. Hughson, and C. Korycinski: 2003, 'Automatic Summarisation of Legal Documents'. In: *Proceedings of the 9th International Conference on Artificial Intelligence and Law*. Edinburgh, Scotland.

Grover, C., C. Matheson, A. Mikheev, and M. Moens: 2000, 'LT TTT—A Flexible Tokenisation Tool'. In: *Proceedings of the 2nd International Conference on Language Resources and Evaluation*. Athens, Greece.

Hachey, B.: 2002, 'Recognising Clauses Using Symbolic and Machine Learning Approaches'. Master's thesis, University of Edinburgh, Edinburgh, Scotland.

Jing, H. and K. R. McKeown: 1999, 'The Decomposition of Human-Written Summary Sentences'. In: *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Berkeley, CA, USA, pp. 129–136.

John, G. H. and P. Langley: 1995, 'Esitmating Continuous Distributions in Bayesian Classifiers'. In: *Proceedings of the 11th Annual Conference on Uncertainty in Artificial Intelligence*. Montréal, Québec, Canada.

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

Kupiec, J., J. Pedersen, and F. Chen: 1995, 'A Trainable Document Summarizer'. In: *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Seattle, WA, USA.

Lafferty, J., A. McCallum, and F. Pereira: 2001, 'Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data'. In: *Proceedings of the 18th International Conference on Machine Learning*. Williams College, MA, USA.

Lapata, M.: 2003, 'Probabilistic Text Structuring: Experiments with Sentence Ordering'. In: *Proceedings of the 41st Meeting of the Association of Computational Linguistics*. Sapporo, Japan.

Littlestone, N.: 1988, 'Learning Quickly when Irrelevant Attributes Abound: A New Linear Threshold Algorithm'. *Machine Learning* **2**, 285–318.

Lupo, C. and C. Batini: 2003, 'A Federative Approach to Laws Access by Citizens: The "Normeinrete" System'. In: *Proceedings of the 2nd International Conference on Electronic Governance*. Prague, Czech Republic.

Maley, Y.: 1994, 'The Language of the Law'. In: J. Gibbons (ed.): *Language and the Law*. London: Longman, pp. 11–50.

Malouf, R.: 2002, 'A Comparison of Algorithms for Maximum Entropy Parameter Estimation'. In: *Proceedings of the Conference on Computational Natural Language Learning*. Taipei, Taiwan.

Mani, I.: 2001, *Automatic Summarization*. John Benjamins.

Mani, I. and E. Bloedorn: 1998, 'Machine Learning of Generic and User-focused Summarization'. In: *Proceedings of the 15th National Conference on Aartificial Intelligence*. Madison, WI, USA.

Mann, W. C. and S. A. Thompson: 1987, 'Rhetorical Structure Theory: Description and construction of text structures'. In: G. Kempen (ed.): *Natural Language Generation: New Results in Artificial Intelligence, Psychology, and Linguistics*. Dordrecht, NL: Marinus Nijhoff Publishers, pp. 85–95.

Marcu, D.: 1999, 'The Automatic Construction of Large-Scale Corpora for Summarization Research'. In: *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Berkeley, CA, USA.

McCallum, A., D. Freitag, and F. Pereira: 2000, 'Maximum Entropy Markov Models for Information Extraction and Segmentation'. In: *Proceedings of the 17th International Conference on Machine Learning*. Stanford University, CA, USA.

McKelvie, D.: 1999, 'XMLPERL 1.0.4 XML processing software'. `http://www.cogsci.ed.ac.uk/~dmck/xmlperl`.

Mikheev, A.: 1997, 'Automatic Rule Induction for Unknown Word Guessing'. *Computational Linguistics* **23**(3), 405–423.

Minnen, G., J. Carroll, and D. Pearce: 2000, 'Robust, Applied Morphological Generation'. In: *Proceedings of 1st International Natural Language Generation Conference*. Mitzpe Ramon, Israel.

Moens, M.-F. and R. D. Busser: 2002, 'First Steps in Building a Model for the Retrieval of Court Decisions'. *International Journal of Human-Computer Studies* **57**(5), 429–446.

Moens, M. F., C. Uyttendaele, and J. Dumortier: 1997, 'Abstracting of Legal Cases: The SALOMON Experience'. In: *The 6th International Conference on Artificial Intelligence and Law*. Melbourne, Victoria, Australia.

Molina, A. and F. Pla: 2002, 'Shallow Parsing Using Specialized HMMs'. *The Journal of Machine Learning Research* **2**, 595–613.

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

Osborne, M.: 2002, 'Using Maximum Entropy for Sentence Extraction'. In: *Proceedings of the ACL-2002 Automatic Summarization Workshop*. Philadelphia, PA, USA.

Platt, J. C.: 1998, 'Fast Training of Support Vector Machines using Sequential Minimal Optimization'. In: B. Schölkopf, C. J. Burges, and A. J. Smola (eds.): *Advances in Kernel Methods: Support Vector Learning*. MIT Press.

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

Ratnaparkhi, A.: 1996, 'A Maximum Entropy Part-Of-Speech Tagger'. In: *Proceedings of the 1st Conference on Empirical Methods in Natural Language Processing*. Philadelphia, PA, USA.

Ratnaparkhi, A.: 1998, 'Maximum Entropy Models for Natural Language Ambiguity Resolution'. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA, USA.

Sang, E. T. K. and H. Déjean: 2001, 'Introduction to the CoNLL-2001 Shared Task: Clause Identification'. In: *Proceedings of the Conference on Computational Language Learning*. Toulouse, France.

Spärck-Jones, K.: 1998, 'Automatic Summarising: Factors and Directions'. In: I. Mani and M. T. Maybury (eds.): *Advances in Automatic Text Summarisation*. Cambridge, Massechusetts, pp. 1–14.

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

Teufel, S., J. Carletta, and M. Moens: 1999, 'An Annotation Scheme for Discourse-Level Argumentation in Reserach Articles'. In: *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics*. Bergen, Norway.

Teufel, S. and M. Moens: 1997, 'Sentence Extraction as a Classification Task'. In: *Proceedings of the ACL/EACL'97 Workshop on Intelligent and Scalable Text Summarization*. Madrid, Spain.

Teufel, S. and M. Moens: 1998, 'Argumentative Classification of Extracted Sentences as a First Step Towards Flexible Abstracting'. In: I. Mani and M. T. Maybury (eds.): *Advances in Automatic Text Summarization*. Cambridge, Massechusetts, pp. 137–175, MIT Press.

Teufel, S. and M. Moens: 1999, 'Discourse-Level Argumentation in Scientific Articles: Human and Automatic Annotation'. In: *Proceedings of the ACL-1999 Workshop Towards Standards and Tools for Discourse Tagging*. College Park, MD, USA.

Teufel, S. and M. Moens: 2000, 'What's Yours and What's Mine: Determining Intellectual Attribution in Scientific Text'. In: *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*. Hong Kong.

Teufel, S. and M. Moens: 2002, 'Summarising Scientific Articles – Experiments with Relevance and Rhetorical Status'. *Computational Linguistics* **28**(4), 409–445.

Thompson, H., R. Tobin, D. McKelvie, and C. Brew: 2004, 'LT XML version 1.2.7'. `http://www.ltg.ed.ac.uk/software/xml`.

van Engers, T. M., R. van Gog, and K. Sayah: 2004, 'A Case Study on Automated Norm Extraction'. In: *Proceedings of the 17th Annual Conference on Legal Knowledge and Information Systems*. Berlin, Germany.

Wasson, M.: 1998, 'Using Leading Text for News Summaries: Evaluation Results and Implications for Commercial Summarization Applications'. In: *Proceedings of the Joint 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*. Montréal, Québec, Canada.

Weber, R. O., K. D. Ashley, and S. Brninghaus: 2006, 'Textual case-based reasoning'. *Knowledge Engineering Review*. To appear, Preprint accessed 25 April 2006 from `http://www.geocities.com/bruninghaus/papers/weberashleybruninghaus-text%ualcbr.pdf`.

Winkels, R., A. Boer, and R. Hoekstra: 2002, 'MetaLex: An XML Standard for Legal Documents'. In: *Proceedings of the XML Europe Conference*. London, England.