Automatic Legal Text Summarisation: Experiments with Summary Structuring

Ben Hachey
School of Informatics, University of Edinburgh
2 Buccleuch Place
Edinburgh EH8 9LW, UK
bhachey@inf.ed.ac.uk

Claire Grover
School of Informatics, University of Edinburgh
2 Buccleuch Place
Edinburgh EH8 9LW, UK
grover@inf.ed.ac.uk

ABSTRACT

We describe a set of experiments using machine learning techniques for the task of extractive summarisation. The research is part of a summarisation project for which we use a corpus of judgments of the UK House of Lords. We present classification results for naïve Bayes and maximum entropy and we explore methods for scoring the summary-worthiness of a sentence. We present sample output from the system, illustrating the utility of rhetorical status information, which provides a means for structuring summaries and tailoring them to different types of users.

Keywords

Automatic summarisation, Discourse, Natural language, Machine learning

1. INTRODUCTION

In this paper we report on a set of experiments to classify sentences for relevance, that is whether they should be part of an extractive summary or not. The sentence extraction task forms part of an automatic summarisation system in the legal domain. The experiments described are part of an ongoing endeavour to determine the best classification techniques and the best feature sets for the task. In the SUM project[1], we are exploring methods for generating flexible summaries of legal documents. Our approach to summarisation is described in detail in [5, 7, 8] and takes as a point of departure the work of Teufel and Moens [17, 16]. The Teufel and Moens approach is an instance of what is known as the *text extraction* method of summarisation. In this approach a summary typically consists of sentences selected from the source text, with some smoothing (e.g reordering, anaphora resolution) to increase the coherence between them. Following Teufel and Moens, we go beyond simple sentence selection and classify source sentences according to their rhetorical status (e.g. a description of background facts in the case, a reference to a point of law, etc.). With sentences classified in this manner, different kinds of summaries can be generated with prominence given to particular kinds of sentence. The main focus of this paper is the sentence extraction task and methods of structuring summaries.

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

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 for the sentence extraction task. 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 House of Lords judgments which we have gathered and annotated. These texts contain a header providing structured information, followed by a sequence of 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 judgments, 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 comprises 188 judgments from the years 2001–2003 from the House of Lords website. (For a subset of these, manually created abstracts 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. 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. The rhetorical roles represent the sentence contribution to the overall communicative goal of the document. In the case of HOLJ texts, the communicative goal for each lord is to convince their peers of the soundness of their argument. Table 1 provides an overview of the rhetorical annotation scheme that we have developed for the HOLJ corpus. In the current version of the corpus there are 69 judg-

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

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

Table 1: Rhetorical Annotation Scheme for Legal Judgments

ments which have been annotated for rhetorical role. The second manual layer is annotation of sentences for 'relevance' as measured by whether they match sentences in hand-written summaries. 47 of the 69 judgments which have been annotated for rhetorical role 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. The rhetorical role annotation and automatic linguistic annotation are described in previous work [5, 7, 8]. We describe the manual relevance annotation in the following subsection.

2.2 Manual Relevance Annotation

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

The literature contains a number of methods for automatic alignment of sentences which would be relevant here (e.g. [15, 13, 2, 14, 9]). However, [17] 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 and 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 used a NITE XML Toolkit annotation tool [3]. 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.

[11] report similar work in the scientific/technical domain and enumerate ways in which summary sentences may match document sentences. 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 Table 2. Other pairings are less direct and Kupiec et al. describe these as incomplete matches and joins. The second two rows of Table 2 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 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 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.

Due to the difficulty in categorising the matches according to the scheme shown in Table 2, we are unable to report statistics which are exactly parallel to the ones given in [11]. We can however provide some statistics from our corpus to elucidate the relation-

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.*

Table 2: Document-Summary Sentence Alignment

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 sum sents
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%

Table 3: Alignment Statistics

ship between the summary sentences and the source documents, as shown in Table 3.

Assuming that the summary sentences which are matched with one (1-1) and (1-2) 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 [11]. 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.

3. EXPERIMENTS

3.1 Relevance Classification

Following from [11], machine learning has been the standard approach to text extraction summarisation as it provides an empirical method for combining different information sources about the textual unit under consideration (e.g. [15, 1]). 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.

As well as 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. For relevance prediction, we performed experiments with publicly available naïve Bayes (NB) and maximum entropy (ME) estimation toolkits. The naïve Bayes implementation, found in the *Weka* toolkit, is based on John and Langley's [10] algorithm incorporating statistical methods for nonparametric density estimation of continuous variables. The maximum entropy estimation toolkit, written by Zhang Le, contains a C++ implementation of the LMVM [12] estimation algorithm. For ME, we use the *Weka* implementation of Fayyad and Irani's [4] MDL algorithm to discretise numeric features.

The features described in [17] include many of the features which are typically used in sentence extraction approaches to automatic summarisation as well as certain other features developed specifically for rhetorical role classification. Briefly, the 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 [17]). 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 [17] 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 feature which is a count of the number of tokens in the sentence.

Quotation. This feature, which does not have a direct match in Teufel and Moens, encodes the proportion of sentence tokens inside an in-line quote and whether the sentence is inside a block quote.

Entities. We recognise a range of named entities [5] and generate binary-valued entity type features which take the value 0 or 1 indicating the presence or absence of each 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 handcrafted 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 [6]. 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 handcoded 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 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. Detailed evaluation efforts for automatic summarisation 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).[3]

[3]Cf e.g. `http://duc.nist.gov/` and `http://lr-www.pi.titech.ac.jp/tsc/index-en.html`.

Yes	**NB**			**ME**		
	P	*R*	*F*	*P*	*R*	*F*
Cue	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. Length	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. Words	31.7	30.7	**31.2**	71.4	16.9	**27.3**
Baseline	46.7/16.0/23.8					

Table 4: Precision, recall and balanced F scores for YES predictions.

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 4 contains precision (*P*), recall (*R*) and F-scores (*F*) for the naïve Bayes (NB) and maximum entropy (ME) classifiers. These are incrementally cumulative starting at the top with just cue phrase features. The baseline is created by selecting sentences from the end of the document as described in the section 4.2.[4]

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. Both classifiers perform significantly better than baseline and F-scores for the best feature combinations are roughly similar to the partial results reported in [17]. 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 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, for example, 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.

Since we need a ranking rather than a *yes/no* classification, this might actually be considered a regression task. However, due to the way the corpus was annotated, the target attribute is in fact binary. As both of our classifiers are probabilistic, we use $p(y = yes|\vec{x})$ as a way to rank sentences. We also remove the normalising factor $\frac{1}{Z(\vec{x})}$ from the maximum entropy classifier so that the values

[4]Note that, in anticipation of the next section and because we are really interested in the summary and not the source, this is a strict evaluation that counts only YES predictions. Micro- and macro-averaging over YES and NO predictions gives F-scores of 87.6 and 67.3 respectively.

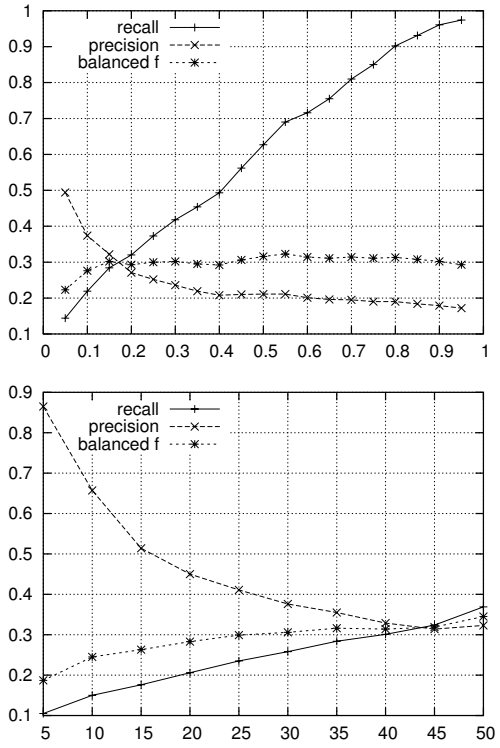

Figure 1: Accuracy plotted against summary size

from the exponential equation $exp(\sum_j \lambda_j f_j(\vec{x}, y))$ can be directly compared. Figure 1 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, at this point, the system does not have an explicit model of the number of sentences of each rhetorical category that should appear in the summary. Table 5 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 given in the rightmost columns. Importantly, the distribution of rhetorical categories in the gold standard extractive summaries is not the same as the distribution in the source documents.

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 5 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.

Besides the logical well-formedness of the summaries, we believe that the accuracy measures will also be improved when we start to control the number of sentences of each rhetorical category that

Rhet Role	*P*	*R*	*F*	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	–	–

Table 5: Precision, recall and balanced F scores by rhetorical category.

end up in the summary. For example, FACT and BACKGROUND sentences tend to get low relevance ranking relative to DISPOSAL sentences and thus are not included until the summary size is quite large. If the summary is made to include 10% FACT and BACKGROUND sentences as in the gold standard extracts, we believe this would improve accuracy within this category. Furthermore, while individual results for the current cue phrase encoding may seem low, preliminary experiments including lemmatised token and hypernym cue phrase features are promising and suggest that we will be able to improve the overall performance with better features based on lexical items:

	NB	ME
Lemmas	38.9/20.1/26.5	63.2/13.0/21.6
Hypernyms	23.6/32.7/27.4	60.3/13.4/22.0

The addition of lemma and hypernym information both give improvements of about 20 points for NB and ME compared to the individual cue phrase scores in table 4. This suggests that our cue phrases are successful, a very encouraging result given that these consist of fully automatic, largely domain-independent linguistic information. Also, as maximum entropy does not model feature interactions particularly well, explicitly conjoined features are likely to improve scores for ME.

Finally, it should be noted that the legal domain appears to be more complex than scientific articles and especially news, the most commonly reported domains in the automatic summarisation literature. This is evidenced in characteristics of legal discourse such as the longer average sentence lengths, longer average document lengths, and the sometimes convoluted and philosophical nature of legalese where there is not an absolute logical template and there is a looser notion of topic which lends itself to a less centralised focus.

4. SUMMARY STRATEGIES

4.1 Manual Summaries from ICLR

In Section 2.2 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 3 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 House of Lords 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 difficult 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:[5]

> *"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 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 (e.g. `http://www.lawreports.co.uk/hlpcmayb0.4.htm`) into an uncompressed list of extracted sentences (e.g. Appendix A) might be just as indicative of content and occasionally more comprehensible to a non-expert reader.

In Section 3 we gave statistics for the relative distribution of rhetorical roles among the sentences that are aligned with summary sentences. From Table 5, 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 and discusses various summary template design issues. First, we present an summary as a running example for this section. Appendices A, B and C show a gold standard extractive summary, a baseline summary, and a summary from our preliminary system respectively. The case is *Regina v. London Borough of Hammersmith and Fulham and Others, Ex P Burkett and Another*, heard on 23 May 2002.[6]

The gold standard extractive summary in Appendix A was formed by selecting all document sentences that were aligned with a sentence from the manual abstractive summary as described in Section 2.2. These are ordered to reflect the order of the corresponding sentences in the manual abstractive summary. 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 baseline summary in Appendix B was formed by selecting the final sentences from each lord. Lords' speeches are ordered by the 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 system summary in Appendix C was formed by selecting sentences the sentences with the highest relevance ranking from each lord, as determined by the ranking method described above in section 3.3. 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

[5]The manual abstractive summary is available from the ICLR web site at `http://www.lawreports.co.uk/hlpcmayb0.4.htm`

[6]The original document is available from the House of Lords web site at `http://www.publications.parliament.uk/pa/ld200102/ldjudgmt/jd020523/burket-1.htm`.

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.

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 1, 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 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 and 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 ME model that is described in section 3.3.

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. [18]).

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), 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 here present summary speeches containing 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, there is also the possibility of grouping and ordering sentences by rhetorical status. Lord Hope of Craighead's speech in Appendix 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.

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.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented work on the automatic summarisation of legal texts. We use 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 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. As with previous work on rhetorical classification [7, 8], we

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

Finally, we 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 performing another evaluation that looks at the correlation between 1/0 numerical representation of whether the sentence was annotated as relevant or not and the ranking score from the classifier. Preliminary results suggest than the normalised probability from the maximum entropy classifier may actually be a better ranking score that the unnormalised score described in section 3.3.

We are also developing a user study which will allow us to assess the value of our system for the information retrieval task referred to throughout this paper. Briefly, this will present a hypothetical case to the subjects with a number of possible precedent-setting cases. The possible precedents will be presented in various formats including our system summaries, the original full text and the gold standard summaries, allowing us to quantify the utility of our system for various real users.

6. REFERENCES

[1] C. Aone, M. E. Okurowski, J. Gorlinsky, and B. Larsen. A trainable summarizer with knowledge acquired from robust NLP techniques. In I. Mani and M. T. Maybury, editors, *Advances in Automatic Text Summarization*, pages 71–80. MIT Press, Cambridge, Massechusetts, 1999.

[2] M. Banko, V. Mittal, M. Kantrowitz, and J. Goldstein. Generating extraction-based summaries from hand-written summaries by aligning text spans. In *Proceedings of the Pacific Association for Computational Linguistics*, 1999.

[3] J. Carletta, S. Evert, U. Heid, J. Kilgour, J. Robertson, and H. Voormann. The nite xml toolkit: flexible annotation for multi-modal language data. *Behavior Research Methods, Instruments, and Computers, special issue on Measuring Behavior*, 35(3), 2003.

[4] U. Fayyad and K. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, 1993.

[5] C. Grover, B. Hachey, and I. Hughson. The HOLJ corpus: supporting summarisation of legal texts. In *Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora*, Geneva, Switzerland, 2004.

[6] C. Grover, B. Hachey, I. Hughson, and C. Korycinski. Automatic summarisation of legal documents. In *Proceedings of the 9th International Conference on Artificial Intelligence and Law*, 2003.

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

[8] B. Hachey and C. Grover. Sequence modelling for sentence classification in a legal summarisation system. In *Proceedings of the 2005 ACM Symposium on Applied Computing*, Santa Fe, New Mexico, 2005.

[9] H. Jing and K. R. McKeown. The decomposition of human-written summary sentences. In *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval*, 1999.

[10] G. H. John and P. Langley. Esitmating continuous distributions in bayesian classifiers. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, 1995.

[11] J. Kupiec, J. Pedersen, and F. Chen. A trainable document summarizer. In *Proceedings of the 18th International Conference on Research and Development in Information Retrieval*, 1995.

[12] R. Malouf. A comparison of algorithms for maximum entropy parameter estimation. In *Proceedings of the 6th Conference on Natural Language Learning*, 2002.

[13] I. Mani and E. Bloedorn. Machine learning of generic and user-focused summarization. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, 1998.

[14] D. Marcu. The automatic construction of large-scale corpora for summarization research. In *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval*, 1999.

[15] S. Teufel and M. Moens. Sentence extraction as a classification task. In *ACL-1997 Workshop on Intelligent and Scalable Text Summarization*, 1997.

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

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

[18] M. Wasson. Using leading text for news summaries: evaluation results and implications for commercial summarization applications. In *Proceedings of the joint 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics*, 1998.

APPENDIX
A. GOLD STANDARD EXTRACT

Rhet	Sent	Rank	Text
			Lord Steyn
DISP	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 .
DISP	378	24	It follows that in my view the decisions of Richards J and the Court of Appeal were not correct .
DISP	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 .
PROC	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) :
PROC	35	0.19	Mrs Burkett and her late husband applied for judicial review .
PROC	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 .
PROC	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 .
PROC	37	0.99	He refused permission on the grounds of delay .
BACK	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 .
DISP	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 .
FRAM	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 .
FRAM	67	0.24	It would be extraordinary if in such a case the House had no jurisdiction .
DISP	70	0.34	In my view the conclusion is inescapable that Lord Diplock 's extempore observation was not correct .
FRAM	335	0.31	It weighs in favour of a clear and straightforward interpretation which will yield a readily ascertainable starting date .
FRAM	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 .
FRAM	337	0.19	Secondly , legal policy favours simplicity and certainty rather than complexity and uncertainty .
FRAM	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 .
PROC	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
DISP	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 .
DISP	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 .
PROC	6	0.39	On 6 April 2000 the appellant applied for leave to move for judicial review of that decision .
PROC	10	1.20	Richards J and the Court of Appeal refused permission on the ground that the application was out of time .
DISP	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
PROC	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 .
FRAM	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 .
DISP	403	1.25	I too would allow the appeal .
			Lord Phillips of Worth Matravers
DISP	457	0.75	For the reasons they give I too would allow the appeal .
			Lord Millet
DISP	453	0.75	For the reasons they give I too would allow the appeal .

B. BASELINE EXTRACT

Rhet	Sent	Rank	Text
			Lord Steyn
FRAM	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) .
FRAM	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 .
FRAM	391	0.16	Moreover , Craig , Administrative Law , 4th ed , has pointed out , at p 794 :
BACK	392	0.23	" The short time limits may , in a paradoxical sense , increase the amount of litigation against the administration .
BACK	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 "
FRAM	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 .
FRAM	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 .
TEXT	396	0.19	XIII .
TEXT	397	0.19	Disposal .
DISP	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
FRAM	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 .
DISP	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
DISP	457	0.75	For the reasons they give I too would allow the appeal .
			Lord Millet
DISP	453	0.75	For the reasons they give I too would allow the appeal .
			Lord Slynn of Hadley
DISP	20	1.35	I would accordingly allow the appeal and remit the substantive question to the High Court for decision .

C. SYSTEM EXTRACT

Rhet	Sent	Rank	Text
			Lord Steyn
PROC	40	1.38	The Court of Appeal refused leave to appeal to the House of Lords .
PROC	43	1.58	In In re Poh the judge had refused leave to apply for judicial review .
PROC	44	1.37	The applicant appealed ex parte by originating motion to the Court of Appeal who refused leave .
PROC	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 .
PROC	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 .
PROC	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) :
PROC	194	5.08	The Court of Appeal [2001] JPL 775 dismissed the appeal and refused leave to appeal to the House of Lords .
FRAM	302	2.45	And in strict law it could be dismissed .
DISP	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 .
DISP	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
FRAM	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 .
DISP	403	1.25	I too would allow the appeal .
			Lord Phillips of Worth Matravers
DISP	457	0.75	For the reasons they give I too would allow the appeal .
			Lord Millet
DISP	453	0.75	For the reasons they give I too would allow the appeal .
			Lord Slynn of Hadley
DISP	20	1.35	I would accordingly allow the appeal and remit the substantive question to the High Court for decision .