Automatic Summarisation of Legal Documents

Claire Grover, Ben Hachey, Ian Hughson and Chris Korycinski
School of Informatics
University of Edinburgh
{grover, bhachey, sih, chrisk}@cogsci.ed.ac.uk

ABSTRACT

We report on the SUM project which applies automatic summarisation techniques to the legal domain. We describe our methodology whereby sentences from the text are classified according to their rhetorical role in order that particular types of sentence can be extracted to form a summary. We describe some experiments with judgements of the House of Lords: we have performed automatic linguistic annotation of a small sample set and then hand-annotated the sentences in the set in order to explore the relationship between linguistic features and argumentative roles. We use state-of-the-art NLP techniques to perform the linguistic annotation using XML-based tools and a combination of rule-based and statistical methods. We focus here on the predictive capacity of tense and aspect features for a classifier.

1. INTRODUCTION

Law reports form the most important part of a lawyer's or law student's reading matter. These reports are records of the proceedings of a court and their importance derives from the role that precedents play in English law. They are used as evidence for or against a particular line of legal reasoning. In order to make judgments accessible and to enable rapid scrutiny of their relevance, they are usually summarised by legal experts. These summaries vary according to target audience (e.g. students, solicitors).

Manual summarisation can be considered as a form of information selection using an unconstrained vocabulary with no artificial linguistic limitations. Automatic summarisation, on the other hand, has postponed the goal of text generation *de novo* and currently focuses largely on the retrieval of relevant sections of the original text. The retrieved sections can then be used as the basis of summaries with the aid of suitable smoothing phrases.

In the SUM project we are investigating methods for generating flexible summaries of documents in the legal domain. Our methodology builds and extends the Teufel and Moens [23] approach to automatic summarisation. The work we report on in this paper deals with judgments from the judicial branch of the House of Lords. We have completed a preliminary study using a small sample of judgment documents. We have hand-annotated the sentences in these documents and performed automatic linguistic processing in order to study the link between the argumentative role and linguistic features of a sentence. Our primary focus is on correlations between sentence type and verb group properties (e.g. tense, aspect). To this end, we have used state-of-the-art NLP techniques to distinguish main and subordinate clauses and to find the tense and aspect features of the main verb in each sentence. In this paper we report on our NLP techniques and on the findings of our study. We discuss the implications for the summarisation system that we are in the process of developing.

Section 2 describes the background to our work. In Section 2.1 we review NLP work in the legal domain and introduce the Teufel and Moens approach to summarisation; in Section 2.2 we describe our methods and the annotation scheme we have developed for the House of Lords judgements. Section 3 provides an overview of the tools and techniques we have used in the automatic linguistic processing of the judgements. Our processing paradigm is XML-based and we use specialist XML-aware tools to perform tasks such as tokenisation, part-of-speech tagging and chunking—these are described in Section 3.1. Our primary interest is tense information about individual sentences and to compute this we need to distinguish main from subordinate clauses in order to identify the main verb group. We report on our statistically-based approach to this task in Section 3.2. In Section 3.3 we present the results of our preliminary evaluations based on the small corpus of hand-annotated judgements. Finally, in Section 4, we discuss present work on refining our rhetorical annotation scheme for the legal domain before drawing some conclusions and outlining future work in Section 5.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
ICAIL'03, June 24-28, 2003, Edinburgh, Scotland, UK
Copyright 2003 ACM 1-58113-747-8. $5.00.

2. AUTOMATIC SUMMARISATION

2.1 Background

Much of the previous NLP work in the legal domain concerns 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 [14]. In our project, we are developing an automatic summarisation system based on the approach of Teufel and Moens. 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 might include standard IR measures such as word frequency but other highly informative features are likely to be ones which reflect linguistic properties of the sentences.

The texts we are currently exploring are judgments of the House of Lords, a domain we refer to here as HOLJ[1]. These 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. This paper focuses on the mixture of statistical and linguistic techniques which aid the determination of the function or importance of a sentence.

Previous work on summarisation has concentrated on the domain of scientific papers. This has lent 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') [23], [22]. 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 [23, 21] and to pursue the methodology of investigating the usefulness of a range of features in determining the argumentative role of a sentence.

Sparck Jones (1999) 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.

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

For long scientific texts, it does not seem feasible to define templates with a wide enough range, however sentence selection does not offer much scope for re-generating 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?

This approach can be thought of as a more complex variant of template filling, where the slots in the template are high-level structural or rhetorical roles (in the case of scientific texts, these slots express argumentative roles like *main goal* and *type of solution*) and the fillers are sentences extracted from the source text using a variety of statistical and linguistic techniques exploiting indicators such as cue phrases. With this combined approach the closed nature of the fact extraction approach is avoided without giving up its flexibility: summaries can be generated from this kind of template without the need to reproduce extracted sentences out of context. Sentences can be reordered, since they have rhetorical roles associated with them; some can be suppressed if a user is not interested in certain types of rhetorical roles.

Features common to information retrieval, which were used successfully in the genre of scientific papers by Teufel and Moens include:[2]

tf∗idf - This is an unsupervised clustering approach originally proposed by Salton [16]. Although this is a crude method of concept identification when applied to free-texts, it could well have a supporting role in topic segmentation and might show major topic shifts. Common techniques which are used to obtain this metric include stopword lists and unsupervised semantic clustering using suffix-stripping algorithms, such as those presented by Porter [15] or Lovins [7].

indicator/cue phrases - These have been used by authors to identify significant areas of text ("... to summarise..."; "...the importance of...") and also to clarify argumentative perspectives ("... on the other hand..."; "... has no basis..."; etc.). These phrases can equally be used by an automatic summarisation system to locate phrases or sentences which correspond to a particular category of argumentative structure [23, 22, 5].

document structure - Texts are far from the 'bag of words' which is sometimes assumed in statistical NLP. Both sentences and paragraphs are carefully structured to try to express the author's meaning. So we can examine a variety of writing features which have been shown to be robust. For example, the sense of the paragraph is often given in the opening sentence, whilst the first sentences in a section may 'set the scene' and the last ones may summarise the section.

[2]We have edited the list to emphasise those features which we judge to be relevant in the law domain. The full list can be consulted in [23].

2.2 House of Lords Judgments

Judgments of the House of Lords are based on facts that have already been settled in the lower courts so they constitute a genre given over to largely unadulterated legal reasoning. Furthermore, being products of the highest court in England[3], they are of major importance for determining the future interpretation of English law. The meat of a 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.

The methodology we implement is based on the argumentative zoning approach to summarisation described above. We are in the early stages which can be described as follows:

Task 1. Decide which rhetorical categories or argumentative moves are of importance in the source text and are of use in the abstract.

Task 2. In a collection of relevant texts, decide for every sentence which rhetorical category best describes it; this process is called "argumentative zoning".

We anticipate that the subsequent steps will be:

Task 3. Build a system which can identify in an unseen text whether sentences express the facts, previous rulings, current ruling, etc.

Task 4. Using sentence selection techniques, select the most abstract-worthy sentences. Use these, together with their rhetorical information, to generate summaries.

Task 5. Evaluate the resulting summaries. Initially we will use the summaries at `http://www.lawreports.co.uk`, but from their very nature, we do not anticipate using them to evaluate an adaptive system.

Our annotation scheme, like our general approach, is motivated by previous successful incorporation of rhetorical information in the domain of scientific articles. Teufel et al. [20] show that regularities in the argumentative structure of a research article follow from the authors' primary communicative goal. In scientific texts, the author's goal is to convince their audience that they have provided a contribution to science. From this goal follow highly predictable sub-goals.

For the legal domain, the communicative goal is slightly different; the author's primary communicative goal is to convince his/her peers that their position is sound, having considered the case with regards to all relevant points of law. A different set of sub-goals follows(refer to Table 1).[4]

[3]To be more specific, the House of Lords hears civil cases from all of the United Kingdom and criminal cases from England, Wales and Northern Ireland.

[4]The basic scheme of the argumentative structure we define turns out to be similar to one which was conceived of for work on legal summarisation of Chinese judgment texts [2].

BACKGROUND - Does the sentence contain generally accepted background knowledge (i.e. sentences containing law, summary of law, history of law, and legal precedents)? **E.g.** "Section 12 (3A) begins with the words: "In determining for the purposes of this section whether to provide assistance by way of residential accommodation to a person...."
CASE - Does the sentence contain a description of the case (i.e. the events leading up to legal proceedings and any summary of the proceedings and decisions of the lower courts)? **E.g.** "Immediately following Mr Fitzgerald's dismissal IMP brought proceedings and obtained a Mareva injunction against him."
OWN - Does the sentence contain statements that can be attributed to the Lord speaking about the case (i.e. include interpretation of BACKGROUND and CASE, argument, and any explicit judgment as to whether the appeal should be allowed)? **E.g.** "For the reasons already given I would hold that VAT is payable in the sum of £1.63 in respect of postage and I would allow the appeal."

Table 1: Description of the basic rhetorical scheme distinguished in our preliminary annotation experiments.

We annotated five randomly selected appeals cases for the purpose of preliminary analysis of our linguistic features. These were marked-up by a single annotator, who assigned a rhetorical label to each sentence. These categories are, effectively, our templates for HOLJ documents (see Section 2.1) and the sentences we extract serve as fillers. This can be viewed as the first stage in the production of adaptive summaries, where certain categories of sentences are more (or less) appropriate to generate a summary suited to a particular type of user.

3. LINGUISTIC ANALYSIS

3.1 Processing with XML›Based Tools

As described in Section 2.2, the sentences in our small pilot corpus were hand annotated with labels reflecting their rhetorical type. This annotation was performed on XML versions of the original HTML texts which were downloaded from the House of Lords website. In this section we describe the use of XML tools in the conversion from HTML and in the linguistic annotation of the documents.

A wide range of XML-based tools for NLP applications lend themselves to a modular, pipelined approach to processing whereby linguistic knowledge is computed and added as XML annotations in an incremental fashion. In processing the HOLJ documents we have built a pipeline using as key components the programs distributed with the LT TTT and LT XML toolsets [3, 24] and the *xmlperl* program [11]. The overall processing stages contained in our pipeline are shown in Figure 1.

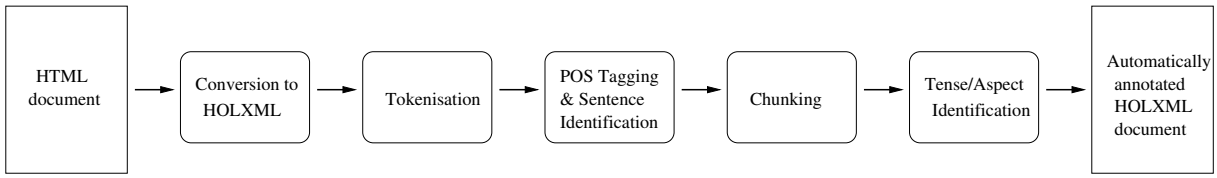

Figure 1: Processing Stages

In the first stage of processing we convert from the source HTML to an XML format defined in a DTD, *hol.dtd*, which we refer to as HOLXML in Figure 1. The DTD defines a House of Lords Judgment as a J element whose BODY element is composed of a number of LORD elements. Each LORD element contains the judgement of one individual lord and is composed of a sequence of paragraphs (P elements).

Once the document has been converted to this basic XML structure, we start the linguistic analysis by passing the data through a pipeline composed of calls to a variety of XML-based tools from the LT TTT and LT XML toolsets. The core program in our pipelines is the LT TTT program *fsgmatch*, a general purpose transducer which processes an input stream and rewrites it using rules provided in a hand-written grammar file, where the rewrite usually takes the form of the addition of XML mark-up. Typically, *fsgmatch* rules specify patterns over sequences of XML elements or use a regular expression language to identify patterns inside the character strings (PCDATA) which are the content of elements. The other main LT TTT program is *ltpos*, a statistical combined part-of-speech (POS) tagger and sentence identifier [12].

The first step in the linguistic annotation process uses *fsgmatch* to segment the contents of the paragraphs into word tokens encoded in the XML as W elements. Once the word tokens have been identified, the next step uses *ltpos* to mark up the sentences as SENT elements and to add part-of-speech attributes to word tokens (e.g. `<W C='NN'>opinion</W>` is a word of category noun). Note that the tagset used by *ltpos* is the Penn Treebank tagset [9].

The following step performs a level of shallow syntactic processing known as "chunking". This is a method of partially identifying constituent structure which stops short of the fully connected parse trees which are typically produced by traditional syntactic parsers/grammars. The output of a chunker contains "noun groups" which are similar to the syntactician's "noun phrases". It also includes "verb groups" which consist of contiguous verbal elements such as modals, auxiliaries and main verbs. To illustrate, the sentence "I would allow the appeal and make the order he proposes" is chunked in this way:[5]

```
<NG>I</NG> <VG>would allow</VG> <NG>the appeal</NG>
and <VG>make</VG> <NG>the order</NG> <NG>he</NG>
<VG>proposes</VG>
```

[5]Judgments - In re Kanaris (Respondent)(application for a writ of Habeas Corpus)(on appeal from the Administrative Court of the Queen's Bench Division of Her Majesty's High Court of Justice), heard on 30 January 2003, paragraph 2

	TENSE	ASPECT	VOICE	MOD
proposes	PRES	SIMPLE	ACT	NO
was brought	PAST	SIMPLE	PASS	NO
would supersede	PRES	SIMPLE	ACT	YES
to grant	INF	SIMPLE	ACT	NO
might have occurred	PRES	PERF	ACT	YES
had been cancelled	PAST	PERF	PASS	NO

Table 2: Tense, Aspect, Voice and Modality Features

The method we use for chunking is another use of *fsgmatch*, utilising a specialised hand-written rule set for noun and verb groups. Once verb groups have been identified we use another *fsgmatch* grammar to analyse the verb groups and encode information about tense, aspect, voice and modality in attributes on the VG elements. Table 2 gives some examples of verb groups and their analysis.

The final stage in the process is the step described in detail in Section 3.2, namely the process of identifying which verb group is the main verb group in the sentence. We call this process from our pipeline using *xmlperl* to pass each sentence in turn to the main verb identifier and to receive its verdict back and encode it in the XML as the value of the MV (main verb) attribute on sentence elements. Figure 2 shows a small part of one of our documents after it has been fully processed by the pipeline.[6]

3.2 Clause and Main Verb Identification

The primary method for identifying the main verb and thus the tense of a sentence is through the clause structure. We employ a probabilistic clause identifier induced from sections 15-18 of the Penn Treebank [10], built as part of post-conference research [4] into the CoNLL-2001 shared task [17]. This section gives an overview of the clause identification system and then describes how this information is incorporated into the main verb identification algorithm.

CoNLL (Conference on Natural Language Learning) is a yearly meeting of researchers interested in using machine learning to solve problems in natural language processing. Each year an outstanding issue in NLP is the focus of the shared task portion of the conference. The organisers make some data set available to all participants and specify how they are to be evaluated. This allows a direct comparison of a number of different learning approaches to a specific problem.

[6]Judgments - Robertson (AP) v Fife Council, heard on 25 July 2002, paragraph 1

```
<LORD>
<P>
<SENT MV='0' sid='1'><NG><W C='NNP'>LORD</W>
<W C='NNP'>SLYNN</W></NG> <W C='IN'>OF</W> <NG>
<W C='NNP'>HADLEY</W></NG></SENT>
</P>
<P>
<SENT MV='0' sid='2'><NG><W C='PRP$'>My</W>
<W C='NNS'>Lords</W></NG><W C=',',>,</W></SENT>
</P>
<P no='1'>
<SENT MV='1' sid='3'><NG><W C='PRP'>I</W></NG>
<VG ASP='PERF' MODAL='NO' TENSE='PRES' VOICE='ACT'
vgid='1'><W C='VBP'>have</W> <W C='VBN'>had</W></VG>
<NG><W C='DT'>the</W> <W C='NN'>advantage</W></NG>
<W C='IN'>of</W> <W C='VBG'>reading</W> <NG>
<W C='DT'>the</W> <W C='NN'>draft</W></NG>
<W C='IN'>of</W> <NG><W C='DT'>the</W> <W C='NN'>
opinion</W></NG> <VG ASP='SIMPLE' MODAL='NO'
TENSE='INF' VOICE='PASS' vgid='2'><W C='TO'>to</W>
<W C='VB'>be</W> <W C='VBN'>given</W></VG>
<W C='IN'>by</W> <NG><W C='PRP$'>my</W>
<W C='JJ'>noble</W> <W C='CC'>and</W> <W C='JJ'>
learned</W> <W C='NN'>friend</W> <W C='NNP'>Lord</W>
<W C='NNP'>Hope</W></NG> <W C='IN'>of</W> <NG>
<W C='NNP'>Craighead</W></NG><W C='.'>.</W>
</SENT> .....
</P>
....
</LORD>
```

Figure 2: A Sample of Annotated HOLJ

The clause identification task is divided into three phases. The first two are classification problems similar to part-of-speech tagging where a label is assigned to each word depending on the sentential context. In phase one, we predict for each word whether it is likely that a clause starts at that position in the sentence. In phase two, we predict clause ends. In the final step, phase three, an embedded clause structure is inferred from the start and end predictions.

The first two phases are approached as straightforward classification in a maximum entropy framework with relative position of contextual information encoded in the features. The maximum entropy algorithm produces a distribution $p_*(\vec{x}, c)$ based on a set of labelled training examples, where $\vec{x}$ is the vector of active features. In evaluation mode, we select the class label c that maximises p_*.

The features we use in the first two phases include words, part-of-speech tags, and chunk tags within a set window as well as features that encode long distance dependencies and sequence information. Consider the task of predicting whether a clause starts at the word *which* in the following sentence:[7]

> Part IV ... is of obvious importance if the Act is to have the teeth which Parliament doubtless intended it should.

[7]Judgments - Anyanwu and Other v. South Bank Student Union and Another And Commission For Racial Equality, heard on 22 March 2001, paragraph 4

System	Precision	Recall	F
CoNLL 1st	84.82	73.28	78.63
Our system	*83.74*	*71.25*	*76.99*
CoNLL Ave	72.46	60.00	65.64

Table 3: Scores for our clause identification system on the Penn Treebank compared to the best and average CoNLL-2001 scores.

The fact that there is this subordinating conjunction at the current position followed by a verb group (*intended*) to the right gives much stronger evidence than if we only looked at the word and its immediate context.

The more difficult part of the task is inferring the proper segmentation. This does not translate to a straightforward classification task as the resulting structure must be a properly embedded, non-overlapping clause structure (e.g. "[The question is [whether the direction [which it contains] applies ...] .]"). To deal with this, we created a maximum entropy model whose sole purpose was to provide confidence values for potential clauses. This model uses features similar to those described above to assign a probability between zero and one for each clause candidate (defined as all ordered combinations of phase one start points and phase two end points). The actual segmentation algorithm then chooses clause candidates one-by-one in order of confidence. After each choice is made, all remaining candidates with crossing brackets are removed from consideration.

Table 3 compares precision, recall, and F scores (a single measure incorporating *Precision* and *Recall*; all F scores in this paper weight *Precision* and *Recall* equally to give the harmonic mean) for our system with CoNLL-2001 results. The results are well above the average scores, failing to surpass only the top CoNLL-2001 system, which obtained F scores some 10 points higher than the second runner-up.[8]

Once clause boundaries have been determined, they are used to identify a sentence's main verb group. A verb group that is at the top level according to the clause segmentation is considered a stronger candidate than any embedded verb group (i.e. a matrix/main clause verb group is preferred over verb groups found in subordinate clauses). In addition, there are several other heuristics encoded in the algorithm. These sanity checks watch for cases in which the complex clause segmenting algorithm described above misses certain strong formal indicators of subordination. Specifically, we consider whether or not a verb group is preceded by a subordinating conjunction (e.g. *that*, *which*). We also consider whether a verb group starts with a participle or infinitive *to* (e.g. *provided* in "accommodation provided for the purpose

[8]For the current work, we obtained a further improvement by training on hand-annotated POS and chunk data from the Treebank. (This wasn't available to the shared task systems as they were mimicking the situation where this labour-intensive information is not available and noisy, automatic approaches must be employed.) In reality, this information is present in the Treebank and using it improved our F score from 73.94 to 76.99. Note that with or without this improvement our system's rank falls between the first and second best CoNLL-2001 systems.

of restricting liberty", *to* in "counted as a relevant period to be deducted"). These heuristics are in the following ranked order (those closer to the beginning of the list being more likely characteristics of a main verb group):

1. *Does not* occur within an embedded clause, *is not* preceded by a subordinating conjunction, *does not* start with a participial or infinitive verb form.
2. *Does* occur within an embedded clause, *is not* preceded by a subordinating conjunction, *does not* start with a participial or infinitive verb form.
3. *Does not* occur within an embedded clause, *is* preceded by a subordinating conjunction.
4. *Does not* occur within an embedded clause, *does* start with a participial or infinitive verb form.
5. *Does* occur within an embedded clause, *is* preceded by a subordinating conjunction.
6. *Does* occur within an embedded clause, *does* start with a participial or infinitive verb form.

We also observed in the corpus that verb groups closer to the beginning of a sentence are more likely to be the main verb group. Therefore, where there are multiple verb groups at a given heuristic level, we prefer those closer to the beginning of a sentence. Scores for main verb group identification are presented in the results section below.

3.3 Results

As mentioned above, the current work has concentrated on identifying the rhetorical structure of the HOLJ domain. In studying this structure, we have begun looking for formal indicators of rhetorical categories as well. The linguistic analysis described in the previous sections is motivated by an observation that tense may be a useful feature. Here, we report a preliminary analysis of this observation short of implementing a classifier. An empirical study of the annotated files reported in section 2.2 provides the starting point for these tasks.

Our identification of the inflection for a sentence depends on the tools described in sections 3.1 and 3.2 above. These consist of (1) identifying the tense of verb groups, and (2) identifying the main verb group. Results for these two steps of automatic linguistic analysis calculated from a sample of 100 sentences from the HOLJ corpus are summarised in Table 4.

For the evaluation of verb group tense identification, we report scores for identifying past and present, defined by the tense, aspect, and modality features on verb groups as follows:

past – TENSE=PAST, ASPECT=SIMPLE, MOD=NO
pres – TENSE=PRES, ASPECT=SIMPLE, MOD=NO

The source of errors for tense identification is mainly due to errors in the POS and chunking phases. In the case of past tense, the POS tagger has difficulty identifying past participles because of their similarity to simple past tense verbs. To illustrate, the word *obtained* in the sentence "IMP brought proceedings and obtained a Mareva injunction against him" is a simple past tense verb. The same word could be used as a past or passive participle in a verb phrase (e.g. *have obtained*, *was obtained*) or the passive participle can head a reduced relative clause (e.g. "The Mareva injunction, obtained immediately after Mr Fitzgerald's dismissal, was brought by IMP"). Performance for present tense verbs is lower than that for past tense verbs because they are more easily mistaken for, say, nouns with the same spelling. For example, there were two errors in our sample where the verb *falls* was tagged as a noun and assigned to a noun group chunk instead of a verb group.

		PRECISION	RECALL	F
1.	*(past)*	97.78	88.00	92.63
	(pres)	81.58	93.93	87.32
2.		90.80	84.04	87.29

Table 4: Performance results on a sample from the HOLJ corpus for (1) tense identification and (2) main verb group identification.

For the evaluation of main verb group identification, we ignore sentences that are not properly segmented (i.e. part of a sentence is missing or more material is included in a sentence than there should be). In these cases, the actual main verb group may or may not be present when the main verb identification algorithm is run. Segmentation is an interesting problem in its own right and is the subject of much research interest. We thought it appropriate to correct faulty segmentation to avoid confounding errors in segmentation with errors in main verb identification. A state-of-the-art approach is included in our XML pipeline [13] and though we may get slightly better performance if we tailor the segmentation algorithm to our domain, in fact there were only 4 cases of bad segmentation in a random sample of 100 sentences.

The main verb group identification algorithm considers only verb groups assigned by the chunker. One obvious problem is that the algorithm is thus not capable of identifying a verb group as being main if the chunker does not identify it at all. The primary source of errors in the remaining sentences are also propagated from earlier stages in the pipeline. The six cases where the algorithm did not identify the main verb group can be attributed to bad part-of-speech tags, bad chunk tags, or poor clause segmentation.

In their work on argumentative zoning, Teufel and Moens [20] do not explicitly use tense information in their heuristic categories.[9] They also point out that their process of identifying indicator phrases is completely manual. Our integration of techniques for automatic linguistic analysis of legal texts allows us to automate the availability of certain linguistic features we think will be useful in sentence extrac-

[9]Note that our linguistic analysis not only makes available information about the tense of the main verb, but all the acquired annotation from intermediate steps: part-of-speech tags, chunk tags, clause structure, and tense information for all verb groups.

Sentences / Tense	Rhetorical Category BACK	CASE	OWN	Total
past	63	346	112	**521**
pres	119	145	254	**518**
Total	**182**	**491**	**366**	**1039**

Table 5: Contingency table comparing Rhetorical Category and Tense.

	BACKGROUND	CASE	OWN
past	-0.135	0.356	-0.261
pres	0.105	-0.301	0.228

Table 6: *Phi* Coefficient between the categories in our basic rhetorical scheme and sentential tense information.

tion and rhetorical classification. To illustrate the utility of tense information, we will look at the relationship between our main rhetorical categories and simple present and past tense.

We present several statistics of 'related-ness'. First, the χ^2 statistic compares distributions of categorical variables and determines the significance of differences between distributions. Table 5 presents the table of frequencies that this calculation is based on. It is immediately observable that past tense sentences take the CASE rhetorical role more often than BACKGROUND and OWN. The significance value p associated with χ^2 will tell us whether this is simply due to the distribution of rhetorical roles or if the variables are truly dependent.

When computing χ^2, the null hypothesis is normally a statement of no difference. In our case, H_0: "The distribution of rhetorical categories does not differ with respect to tense". When we calculate χ^2 from Table 5, we get a value of 154.60. This translates to significance value far beyond 0.0001 meaning we can confidently reject the null hypothesis and accept the alternative hypothesis that there is a relation between a sentence's rhetorical category and tense.

This confirms the observation that there is a relationship, but it does not give us an idea as to the degree of that relationship. Nor does it describe the specific relationship we observed between past tense and the CASE rhetorical category. The first of these issues is addressed by Cramer's V. Cramer's V standardises χ^2 for sample size and table shape giving a scaled measure of the degree of dependence. Where the range of Cramer's V is $[0, 1]$, higher values indicate stronger dependence. From Table 5, we get a value of 0.39. This indicates that tense information will indeed help to determine the argumentative role of a sentence in conjunction with other standard features.

The specific relationship between past tense and CASE can be explored using the *Phi* coefficient. The *Phi* coefficient is a statistical measure of 'related-ness' for binomial variables that is interpreted like correlation. Values fall in the range $[-1, 1]$, where *positive Phi* means the variables tend to be the same, 0 *Phi* means the variables are not correlated, and *negative Phi* means the variables tend to be opposite. Table 6 presents *Phi* coefficient scores comparing rhetorical categories from our basic scheme with past and present tense variables.

For illustrative purposes, we will focus on identifying the CASE rhetorical move. Past tense and the CASE rhetorical move have a moderate positive *Phi* coefficient. Also, present tense and the CASE rhetorical move have a moderate negative *Phi* coefficient. This suggests two features based on our linguistic analysis that will help a statistical classifier identify the CASE rhetorical move: (1) the sentence is past tense, and (2) the sentence is not present tense. Furthermore, comparing rows indicates that these are both good discriminative indicators. In the case of past tense, there is a positive *Phi* coefficient with the CASE rhetorical move while there is a very weak negative *Phi* coefficient with BACKGROUND and a slightly stronger negative *Phi* coefficient with OWN.

Finally, these results also illustrate the complexity of tense information. In order to identify simple past tense sentences, we must examine three separate mark-up attributes on verb group elements. We check that the TENSE attribute of the main verb group has the value PAST, the ASPECT attribute has the value SIMPLE and the MODAL attribute has the value NO. Feature construction techniques offer a means to automatic discovery of complex features of higher relevance to the concept being learned. Employing machine learning approaches that are capable of modelling dependencies among features (e.g. maximum entropy) is another way to deal with this.

4. PRESENT WORK

The investigations reported here took as their starting point the basic version of the rhetorical annotation scheme for scientific papers presented by Teufel et al. [20]. This tripartite analysis is neither rich nor targeted enough to provide a sound basis for proceeding with more extensive studies of legal judgments. Current work on the SUM project is focusing on refining our rhetorical annotation scheme to better reflect the main communicative goal of HOLJ texts.

Legal judgments are very different from articles reporting scientific research as regards communicative goals. They are more strongly performative than research reports, the fundamental act being decision. The reason they are not as brief as that would suggest is that public justice demands that the reasoning leading to a decision be laid open to scrutiny by all. In particular, the judge aims to convince his professional and academic peers of its soundness. Therefore, a judgment serves both a declaratory and a justificatory function [8]. 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.

Teufel and Moens [22] found that a significant part of creating and occupying a scientific niche [19] consists of re-

searchers making clear which assertions they make are their own, which are taken from the work of specific other researchers, and which are drawn from generally accepted background knowledge in the field. The three rhetorical categories which make up the basic version of their annotation scheme are direct reflections of this. The rhetorical scheme used to annotate the full SUM corpus should likewise reflect the fundamental communicative function of judgments. The act of legitimising the decision must therefore be operationalised in order to derive more appropriate rhetorical categories.

Beyond simply attributing work appropriately, Teufel and Moens [23] found that expressing its relationship to other work was also a significant feature of scientific argument. Specifically, scientific researchers make clear where their own work adopts, builds on, or extends that of others and where it differs or departs from others'. The judicial situation is in part similar and in part not. On the one hand, as has already been observed, the role of the judge as lawmaker is downplayed, so there is no equivalent of 'own work' to be assessed in this manner. On the other hand, in demonstrating the provenance of his position, a judge's argument seeks to align the case and the authorities he favours while distancing them from those he does not.

In fact, it appears there are two overlapping aspects of legal judgments which must be teased apart. They might be termed *formulating* and *favouring*. The formulating aspect is that in which the judge seeks to derive a statement of the law that can be applied to the current case. This is, at least on the face of it, an objective and impartial exercise, intended to arrive at the 'correct' interpretation of the law by imposing some kind of order on the body of authorities under consideration. In the favouring aspect, meanwhile, the judge has to make a stark, either-or choice between the competing parties. He must be swayed toward one side and away from the other, by finding the argument of the one stronger or weightier than that of the other. These two aspects, unlike attribution and comparison in scientific writing, seem to be in tension, so the shape of an annotation scheme based on them is not yet clear.

The current working version (see Table 7) reflects this move toward a finer distinction among *formulating* and *favouring* kinds of sentences. The PROXIMATION and DISTANCING categories reflect the notion of *favouring* and FRAMING reflects the notion of *formulating*. DISPOSAL forms another integral part of a judge's argument in which they give their opinion. The CASE category is broken down into FACT and PROCEEDINGS categories and the BACKGROUND rhetorical move stays largely the same.

It is worth noting that these theoretical issues have been brought into focus by experimenting with prototype annotation schemes and, once they have been settled, challenges of a more practical nature will continue to inform the work. In particular, the approach to summarisation being investigated here is a sentence-based one which assumes that a single rhetorical category can be assigned to every sentence in a text. There is no doubt that the sentences produced by judges are often of unusual length and complexity. They can and do sometimes encapsulate complete indirect arguments

FACT - Does the sentence recount one or more of the events or circumstances which gave rise to the legal proceedings?
PROCEEDINGS - Does the sentence describe claims, arguments, or rulings from previous (lower court) hearings of the case?
BACKGROUND - Is the sentence an unqualified recitation or summary of *source of law* material?
PROXIMATION - Does the sentence serve to position the case closer to *source of law* material?
DISTANCING - Does the sentence serve to position the case farther from *source of law* material?
FRAMING - Does the sentence attempt to define the primary question of the case or set forward non-source-of-law principles?
DISPOSAL - Does the sentence present an opinion as to whether the legal action succeeds or fails? Does the sentence detail damages or otherwise give instructions on how the case should be taken forward?

Table 7: The current working version of the rhetorical annotation scheme.

of the type mentioned above. They also can and do perform multiple functions, such as disposing in different ways of several precedents which have been under consideration. Whether or not it is possible to categorise their rhetorical status in unitary terms in any scheme remains to be seen.

5. CONCLUSIONS AND FUTURE WORK

The work reported forms the initial stages in the development of an automatic text summarisation system for judicial transcripts from the House of Lords. We have presented an initial annotation scheme for the rhetorical structure of the domain, assigning a label indicating the argumentative role of each sentence in a portion of the corpus. A number of sophisticated linguistic tools have been described that identify tense information. Finally, various statistical measures of 'related-ness' were presented, illustrating the utility of this information.

We are also interested in improving the tools we use to identify tense features. One way to do this is to retrain the clause identifier. The legal language of the HOLJ domain is considerably different than the expository newspaper text from the Penn Treebank. Furthermore, the Penn Treebank is American English. Ideally, we would like to hand-annotate a portion of the legal judgments with syntactic parse information and train a clause identifier from this. However, this kind of work is very labour intensive and a more realistic approach to ensuring that the training data is slightly more representative might be to retrain the clause identifier on a corpus of British English like the British National Corpus [1].

Finally, as mentioned above, we are specifically interested in

employing feature construction and selection techniques for identifying the relationship between tense features. We are also interested in employing feature mining techniques for automatically identifying cue phrases within sentences. This could be similar to [6], where sequential features are mined from the textual context for a context-sensitive approach to spelling correction.

Acknowledgments

We would like to express out thanks to Beatrice Alex for annotating the texts used in this study and providing invaluable feedback regarding the rhetorical categories. We would also like to thank the three anonymous referees whose comments were very helpful for improving this paper. Finally, we are grateful to the NITE project (`http://nite.nis.sdu.dk/`) for providing us with high-quality, customised annotation tools for our task. This research is supported by EPSRC grant GR/N35311.

6. REFERENCES

[1] G. Burnage and D. Dunlop. Encoding the British National Corpus. In J. Aarts, P. de Haan, and N. Oostdijk, editors, *Design, Analysis and Exploitation, Papers from the 13th international conference on English Language research on computerized corpora*, 1992.

[2] L. Cheung, T. Lai, B. Tsou, F. Chik, R. Luk, and O. Kwong. A preliminary study of lexical density for the development of xml-based discourse structure tagger. In *Proceedings of the 1st NLP and XML Workshop*, pages 63–70, 2001.

[3] C. Grover, C. Matheson, A. Mikheev, and M. Moens. Lt ttt—a flexible tokenisation tool. In *LREC 2000—Proceedings of the second international conference on language resources and evaluation*, pages 1147–1154, 2000.

[4] B. Hachey. Recognising clauses using symbolic and machine learning approaches. Master's thesis, University of Edinburgh, 2002.

[5] A. Knott. *A data-driven methodology for motivating a set of coherence relations*. PhD thesis, Department of Informatics, University of Edinburgh, 1995.

[6] N. Lesh, M. Zaki, and M. Ogihara. Mining features for sequence classification. In *Proceedings of the 5th International Conference on Knoledge Discovery and Data Mining*, pages 342–346, 1999.

[7] J. B. Lovins. Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11(1–2):22–31, 1968.

[8] Y. Maley. The language of the law. In J. Gibbons, editor, *Language and the Law*, pages 11–50. Longman, London, 1994.

[9] M. Marcu, G. Kim, M. A. Marcinkiewicz, and R. MacIntyre. The penn treebank: annotating predicate argument structure. In *ARPA human language technologies workshop*, 1994.

[10] M. Marcus, B. Santorini, and M. A. Marcinkiewicz. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.

[11] D. McKelvie. Xmlperl 1.0.4 xml processing software. *http://www.cogsci.ed.ac.uk/~dmck/xmlperl*, 1999.

[12] A. Mikheev. Automatic rule induction for unknown work guessing. *Computational Linguistics*, 23(3):405–423, 1997.

[13] A. Mikheev. Periods, capitalized words, etc. *Computational Linguistics*, 28(3):289–318, 2002.

[14] M. F. Moens and R. D. Busser. First steps in building a model for the retrieval of court decisions. *Int. J. Human-Computer Studies*, 57(5):429–446, 2002.

[15] M. F. Porter. An algorithm for suffix stipping. *Program*, 14(3):130–137, 1980.

[16] G. Salton. *A Theory of Indexing*. SIAM, Philadelphia, 1975.

[17] E. T. K. Sang and H. Déjean. Introduction to the CoNLL-2001 shared task: clause identification. In *Proceedings of The 5th Workshop on Computational Language Learning*, pages 53–57, 2001.

[18] K. Sparck-Jones. Automatic summarising: factors and directions. In *Advances in automatic text summarisation*, pages 1–14. MIT Press, 1998.

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

[20] S. Teufel, J. Carletta, and M. Moens. An annotation scheme for discourse-level argumentation in reserach articles. In *Proceedings of the 9th conference of the European chamber of the ACL*, pages 110–117. ACL, 1999.

[21] S. Teufel and M. Moens. Sentence extraction as a classification task. In *Workshop 'Intelligent and scalable Text summarization*, pages 58–65. ACL/EACL, 1997.

[22] S. Teufel and M. Moens. What's yours and what's mine: Determining intellectual attribution in scientific text. In *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 84–93, 2000.

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

[24] H. Thompson, R. Tobin, D. McKelvie, and C. Brew. Lt xml. software api and toolkit for xml processing. *http://www.ltg.ed.ac.uk/software/*, 1997.