# Sentence Extraction for Legal Text Summarisation

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

## Abstract

We describe a system for generating extractive summaries of texts in the legal domain, focusing on the relevance classifier, which determines which sentences are abstract-worthy. We experiment with naïve Bayes and maximum entropy estimation toolkits and explore methods for selecting abstract-worthy sentences in rank order. Evaluation using standard accuracy measures and using correlation confirm the utility of our approach, but suggest different optimal configurations.

## 1 Introduction

In the SUM project we are developing a system for summarising legal judgments that is generic and portable and which maintains a mechanism to account for the rhetorical structure of the argumentation of a case. Following Teufel and Moens [2002], we are developing a text extraction system that retains a flavour of the fact extraction approach. This is achieved by combining sentence selection with information about *why* a certain sentence is extracted—e.g. is it part of a judge's argumentation, or does it contain a decision regarding the disposal of the case? In this way we are able to produce flexible summaries of varying length and for various audiences. Sentences can be reordered, since they have rhetorical roles associated with them, or they can be suppressed if a user is not interested in certain types of rhetorical roles.

We have prepared a new corpus of UK House of Lords judgments (HOLJ) for this work which contains two layers of manual annotation: rhetorical role and relevance. The rhetorical roles represent the sentence's 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. In the current version of the corpus there are 69 judgments 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. In the current version of the corpus, 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.

## 2 Classification and Relevance

Following from [Kupiec *et al.*, 1995], 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. For relevance prediction, we performed experiments with publicly available naïve Bayes (NB) and maximum entropy (ME) estimation toolkits. The NB implementation, found in the *Weka* toolkit, is based on John and Langley's [1995] algorithm incorporating statistical methods for nonparametric density estimation of continuous variables. The ME estimation toolkit, written by Zhang Le, contains a C++ implementation of the LMVM [Malouf, 2002] estimation algorithm. For ME, we use the *Weka* implementation of Fayyad and Irani's [1993] MDL algorithm to discretise numeric features.

The features that we have been experimenting with for the HOLJ corpus are broadly similar to those used by Teufel and Moens [2002]. They consist of **location** features encoding the position of the sentence in document, speech and paragraph; a **thematic words** feature encoding the average *tf*idf* weight of the sentence terms; a **sentence length** feature encoding the number of tokens in the sentence; **quotation** features encoding percentage of sentence tokens inside and in-line quote and whether or not the sentence is inside a block quote; **entity** features encoding the presence or absence of named entities in the sentence; and **cue phrase** features.

The term 'cue phrase' covers the kinds of stock phrases which are frequently good indicators of rhetorical status (e.g. phrases such as *The aim of this study* in the scientific article domain and *It seems to me that* in the HOLJ domain). Teufel and Moens invested a considerable amount of effort in building hand-crafted lexicons where the cue phrases are assigned to one of a number of fixed categories. A primary aim of the current research is to investigate whether this information can be encoded using automatically computable linguistic features. If they can, then this helps to relieve the burden involved in porting systems such as these to new domains. Our preliminary cue phrase feature set includes syntactic features of the main verb (voice, tense, aspect, modality, negation). 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

| | NB | | | ME | | |
|---|---|---|---|---|---|---|
| | *P* | *R* | *F* | *P* | *R* | *F* |
| Cue | 34.9 | 21.5 | 26.6 | 66.6 | 15.2 | 24.8 |
| Entities | 30.7 | 26.4 | 28.4 | 66.8 | 15.4 | 25.1 |
| Them. Words | 32.2 | 26.9 | 29.3 | 68.6 | 15.7 | 25.5 |
| Location | 31.6 | 27.2 | 29.2 | 73.4 | 16.4 | 26.9 |
| Quotations | 31.2 | 27.7 | 29.4 | 71.7 | 17.4 | **28.0** |
| Sent. Length | 31.7 | 29.4 | **29.8** | 71.4 | 16.9 | 27.3 |

Table 1: Accuracy measures for *yes* predictions.

| | NB | | ME | |
|---|---|---|---|---|
| | **I** | **C** | **I** | **C** |
| Cue | 0.187 | 0.187 | 0.208 | 0.208 |
| Entities | 0.103 | 0.211 | 0.056 | 0.219 |
| Them. Words | 0.016 | 0.211 | 0.000 | **0.227** |
| Location | 0.104 | 0.229 | -0.031 | 0.166 |
| Quotations | 0.092 | 0.233 | 0.093 | 0.187 |
| Sent. Length | 0.069 | **0.235** | 0.000 | 0.175 |

Table 2: Point-biserial correlation coefficients.

the hand-coded *agent* features of Teufel and Moens. A main verb lemma feature simulates Teufel and Moens's *type of action* and a feature encoding the part-of-speech after the main verb is meant to capture basic subcategorisation information.

## 3 Experimental Results

Table 1 contains cumulative precision (*P*), recall (*R*) and f-scores (*F*) for the naïve Bayes (NB) and maximum entropy (ME) classifiers on the relevance classification task.[1] Though only the cue phrase feature set performs well individually, all feature sets contribute positively to the cumulative scores with the exception of sentence length for ME and quotation for NB. Both classifiers perform significantly better than a baseline created by selecting sentences from the end of the document, which obtains *P*, *R* and *F* scores of 46.7, 16.0 and 23.8. F-scores for the best feature combinations are similar to the partial results reported in Teufel and Moens [2002]. Taking the f-score as the best metric to optimise would lead us to choose NB.

However, 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. So, what we actually want to do is rate *how* relevant/extract-worthy a sentence is in such a way that will allow us to select sentences in rank order. Bearing this in mind, precision is probably the more important metric given that recall will be controlled by the size of the summary. So, ME with all but sentence length features actually appears to be the better approach for sentence extraction.

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. To evaluate the ranking methods with respect to our binary gold standard, we use the point-biserial correlation coefficient ($r_{pb}$). Table 2 contains correlation coefficients between the gold standard *yes/no* classification and $p(y = yes|\vec{x})$ for naïve Bayes (NB) and maximum entropy (ME).[2] The I column has scores for the individual feature sets and the C column has cumulative scores. The correlation results are strikingly different for NB and ME. While NB successfully incorporates all features ($r_{pb} = 0.235$), ME performs best using only cue phrase, entity and thematic word features ($r_{pb} = 0.208$). For ME, the location feature set actually gives a negative correlation. Judging by these results, we would again be likely to choose NB.

## 4 Conclusions and Future Work

In this paper, we have presented work on the automatic summarisation of legal texts for which we have compiled a new corpus with annotation of rhetorical status, relevance and linguistic markup. We presented sentence extraction results in classification and ranking frameworks. Naïve Bayes and maximum entropy classifiers achieve significant improvements over the baseline according to standard accuracy measures. We have also used the point-biserial correlation coefficient for quantitative evaluation of our extraction system, the results of which suggest diffierent optimal configurations. In current work, we are developing a user study that will help determine empirically whether correlation coefficients are a better evaluation metric than precision and recall accuracy measures.

## References


[Fayyad and Irani, 1993] U. Fayyad and K. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In *IJCAI*, 1993.

[John and Langley, 1995] G. H. John and P. Langley. Esitmating continuous distributions in bayesian classifiers. In *UAI*, 1995.

[Kupiec *et al.*, 1995] J. Kupiec, J. Pedersen, and F. Chen. A trainable document summarizer. In *SIGIR*, pages 68–73, 1995.

[Malouf, 2002] R. Malouf. A comparison of algorithms for maximum entropy parameter estimation. In *CoNLL*, 2002.

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

[Wolf and Gibson, 2004] F. Wolf and E. Gibson. Paragraph-, word-, and coherence-based approaches to sentence ranking: A comparison of algorithm and human performance. In *ACL*, 2004.


[1] Note that this is a strict evaluation that counts only *yes* predictions. Micro- and macro-averaging over *yes* and *no* predictions give e.g. f-scores of 87.6 and 67.3 respectively for ME.

[2] It has been argued that this is actually a better evaluation than standard accuracy measures, which do not account for degree of agreement [Wolf and Gibson, 2004].