Legal Knowledge and Information Systems
S. Villata et al. (Eds.)

doi:10.3233/FAIA200875

Summarisation with Majority Opinion

Oliver RAY, Amy CONROY, and Rozano IMANSYAH
Department of Computer Science, University of Bristol
e-mail: {csxor, ac16888, lv18263}@bristol.ac.uk

Abstract. This paper introduces a method called SUmmarisation with Majority Opinion (SUMO) that integrates and extends two prior approaches for abstractively and extractively summarising UK House of Lords cases. We show how combining two previously distinct lines of work allows us to better address the challenges resulting from this court's unusual tradition of publishing the opinions of multiple judges with no formal statement of the reasoning (if any) agreed by a majority. We do this by applying natural language processing and machine learning, Conditional Random Fields (CRFs), to a data set we created by fusing together expert-annotated sentence labels from the HOLJ corpus of *rhetorical role summary relevance* with the ASMO corpus of *agreement statement* and *majority opinion*. By using CRFs and a bespoke summary generator on our enriched data set, we show a significant quantitative F1-score improvement in rhetorical role and relevance classification of 10-15% over the state-of-the-art SUM system; and we show a significant qualitative improvement in the quality of our summaries, which closely resemble gold-standard multi-judge abstracts according to a proof-of-principle user study.

Keywords. Legal Summarisation, UK House of Lords (UKHL), Machine Learning.

1. Introduction

The summarisation of legal judgments is a challenging task [1] especially in courts like the UK House of Lords (UKHL) which publish the opinions of multiple judges with no formal statement of the reasoning (if any) agreed by a majority [2,3]. The aim of this work is to automatically generate multi-judge summaries that closely resemble gold-standard abstracts published in the Incorporated Council of Law Reporting (ICLR) Daily Law Reports (DLR). We achieve this goal by integrating and extending two previously independent lines of work applying computational methods to UKHL case law [4,5].

First, we create an enriched data set of UKHL cases by fusing expert-annotated sentence labels from the HOLJ corpus of [4], which marks up the *rhetorical role* and *summary relevance* of sentences, together with the ASMO corpus of [5], which marks up explicit inter-judge *agreement statements* and *majority opinions*. Then we implement a new summary pipeline, called SUmmarisation with Majority Opinion (SUMO) that uses natural language processing and Conditional Random Fields (CRFs) to generate better quality summaries than the previous state-of-the-art system, *SUM* [4].

The main benefits of SUMO over SUM are that: (i) we increase the rhetorical role and relevance classification F1-scores by 10-15% (to about 75% and 40%, respectively); (ii) we supplement extractively generated case abstracts with abstractively generated inter-judge agreement summaries in the DLR style; and (iii) we demonstrate superior quality using both ROUGE metrics and expert feedback from a preliminary user study.

2. Background

The UKHL, or UK Supreme Court (UKSC) since 2009, differs from most other courts by publishing judgments that consist of the seriatim opinions of multiple law lords (usually 5 from a panel of 12) with no accompanying statement of consensus (if one even exists) on the *ratio decidendi*. And, while the judges always return a *majority decision* (to allow or dismiss an appeal), a binding precedent is only set by a *majority opinion* (where more than half also agree on the legal reasons) [2]. Thus, judges usually discuss drafts of their speeches with each other and often state (dis)agreements with their peers in the final judgment. But, in practice, even UKHL/UKSC judges recognise that it can be very hard to determine when a majority opinion exists [3]. As a result of this unique challenge, there is very little prior research on the automatic summarisations of UKHL cases. In fact, we found just two lines of work - that we integrate and extend in this paper.

The first strand of work is the SUM system [4] which generates extractive summaries by classifying sentences according to their rhetorical role (**Facts**, **Proceedings**, **Background**, **Framing**, **Disposal**, **Textual** and **Other**) and classifying sentences as *relevant* to the summary. They introduced the HOLJ corpus which marks up the sentences of 47 UKHL cases with expert-annotated labels indicating their main rhetorical role and to which (if any) of the DLR gold-standard summary sentences they most closely align. The sentences are also marked up with machine-generated labels denoting linguistic features like sentence length and location, named entities, quoted text, thematic words and cue phrases. These were used to train two classifiers which achieved F1-scores of 61.2% for role and 31.2% for relevance; and these predictions were then used to extract summary sentences more effectively than a variety of baseline methods.

The second strand of work is the ASMO system [5] which identifies explicit inter-judge (dis)agreement statements and uses them to infer the existence of incontestable majority opinions. They introduced the ASMO corpus which marks up the sentences in a superset of 300 UKHL cases with expert-annotated labels identifying **acknowledgements**, **outcomes**, various types of (dis)agreement (**Full**, **Partial**, **Order**, **Generic** and **Self**), along with the set of judges (if any) whose reasoning forms the **majority opinion**. The sentences are also marked up with machine-generated labels (inspired by HOLJ) denoting length and location, unigrams and POS tags, named entities and a set of hand-crafted cue phrases. These were used train a classifier which detects full agreement statements with an F1-score of 94.3% and uses them to infer incontestable majority opinions with an F1-score of 81%.

3. Summarisation by Majority Opinion (SUMO)

We began by combining the expert labels from HOLJ and ASMO to create an enriched UKHL corpus. Due to the differences in case identification and sentence splitting, this required a non-trivial alignment and merging process [6]. We used normalised variants of sentence length and location, and quotations and cue phrases as our feature-sets. We also identified generic named entities using spaCy[1] and legal entities using ICLR&D's Blackstone[2]. This resulted in 7 feature-sets which we used to train our rhetorical and

[1] https://spacy.io/ [2] https://github.com/ICLRandD/Blackstone

relevance classifiers (using predicted role as an extra feature when training the relevance classifier).

We developed our *SUMO* pipeline in Python using a combination of shallow natural language processing and supervised machine learning [7]. Our approach uses a multi-class rhetorical classifier (to predict the role of a sentence) as well as a binary relevance classifier (to predict if it aligns to a sentence in the summary). We trained the model by splitting the corpus in to self-contained speeches rather than whole judgments, as we hypothesised this would help our sequential modelling method to exploit the overall structure of each lord's speech without being confused by transitions between speeches.

We performed the classifications tasks using the novel approach of applying CRFs to summarise legal texts, previously attempted by only one piece of work [8]. CRFs avoid biases evident in other sequence models such as Maximum-entropy Markov models by using a single exponential model to determine the probability of the entire sequence of the labels. We extract the marginal probability from the relevance classifier to assign a ranking to each sentence as to *how* summary-worthy the sentence is. This give us more flexibility to create summaries of arbitrary lengths depending on the needs of the user. We combine this data with the rhetorical role to output structured summaries in the same style as the DLR gold standard summaries.

In order to replicate the manually written statements from the DLR summaries that indicate agreement between lords, we use the data from the ASMO system to identify the agreements as well as who formed the majority opinion (see [7]). This meant that our summaries include representative sentences such as: *"...***LORD SLYNN** *and* **LORD STEYN. LORD MILLETT** *and* **LORD PHILLIPS** *delivered an opinion agreeing with* **LORD SLYNN** *and* **LORD STEYN. LORD HOPE** *did not agree with the line of reasoning..."* We combine this information with the rhetorical roles predicted by our system to select the highest ranking sentences and create a structured summary in the same style as the ICLR gold standard. This goes beyond the simple ranking only summary produced by the SUM system.

4. Results and Evaluation

Using our methodology we are able to achieve a weighted average F1-score for our rhetorical classifier of 77.8%, with RandomizedSearchCV utilised to validate our results. This is a 16.6% increase over SUM's rhetorical classifier. Our relevance classifier achieves a binary-averaged F1-score of 42.1%, validated using the same methodology as our rhetorical classifier. This is a 10.9% increase over the SUM system's relevance classifier.

Evaluation of automatically generated texts and in particular of summaries can be very difficult, largely due to the subjective nature of summaries. We use the ROUGE 2.0 toolkit[3] to quantitatively evaluate the summaries produced by our system. We compare the results of the *SUMO* system with a summary generated using the same methodology as the SUM system. The ROUGE-1 F1-score results indicate that the summary produced by SUMO (48.9%) perform better than summaries produced using the SUM methodology (37.6%) as well as the baseline summary (41.9%). Our use of the majority opinion to abstractively generate the agreement sentences that closely resembles the manually written summaries likely contributes to a higher F1-score.

[3] https://github.com/kavgan/ROUGE-2.0

As the ROUGE metrics are not necessarily indicative of a good summary, we balanced this evaluation with a user study. We recruited 8 experts (individuals with UK legal experience, either as an LLB student or graduate and/or as a legal professional), and 10 non-experts to complete our study, which was an online survey. The study evaluated our *SUMO* summary compared to the corresponding ICLR summary across three randomly selected judgments, evaluated using questions in the form of 7-point Likert scales. 81.5% of our participants agreed that our summary was a valid replacement for the ICLR gold standard, and 83.3% agreed that it contained the most important aspects of the case.

One notable comment from one of our evaluators indicated confusion regarding our use of the word agreement. While the summary states that the lord did not agree with the *line of reasoning* of his fellow lords, the first disposal sentence we extracted from him details that he agreed with his fellow lords that the outcome should be dismissed. This shows an interesting observation between the agreement as to the outcome and agreement of the line of reasoning of his fellow lords, a distinction that indicates whether the line of reasoning forms a precedent in common law systems or not.

5. Conclusion

The *SUMO* system introduced in this paper sets a new benchmark for the automatic summarisation of legal judgments in the UK. By applying CRFs to summarise legal texts, as well as introducing a new type of ASMO feature, we improve the F1-scores of the rhetorical role and summary relevance prediction tasks by 10-15% over previous research. We further exploited ASMO features in order to abstractively generate parts of the summary, which based on the ROUGE metrics and positive user feedback indicate a close resemblance to the gold-standard text.

For future work we are developing an NLP method for inferring the decisions of individual sentences from outcome statements (which an analysis of numerous problematic cases shows is not as trivial as it may first seem). This could help us address another important task, revealed by our user feedback, of automatically resolving the ambiguity often associated with different intended uses of the word ‘agreement’: such as in the DLR summaries where it is used loosely, variously referring to reasons, outcomes and orders, or just facts and issues.

References

[1] A Kanapala et al. Text summarization from legal documents: a survey. *Artif Intell Rev (2019) 51: 371–402.*

[2] G Williams. *Learning the Law.* Sweet & Maxwell, 14 edition, 2010.

[3] B Hale. Judgment Writing in the Supreme Court – UK Supreme Court Blog (October), 2010.

[4] B Hachey and C Grover. Extractive Summarisation of Legal Texts. *AI and Law*, 14(4):305–345, 2006.

[5] J Valvoda et al. Using Agreement Statements to Identify Majority Opinion in UKHL Case Law. In *Proc. 31st Int. Conf. on Legal Knowledge and Info. Sys.*, Frontiers in AI and Applications (313): 141-150, 2018.

[6] R Imansyah. Predicting the Role and Relevance of Sentences in UK House of Lord Judgements. Master’s thesis, University of Bristol, Bristol, UK, 2019.

[7] A Conroy. SUMO: A System for Automatically Summarising UK House of Lords (UKHL) Judgments using Majority Opinion. Master’s thesis, University of Bristol, Bristol, UK, 2020, submitted.

[8] M Saravanan et al. Improving Legal Document Summarization Using Graphical Models. *Frontiers in AI and Applications (152): 51-59*, 152:51, 2006.