An Evaluation Framework for Legal Document Summarization

Ankan Mullick♠, Abhilash Nandy♠♢, Manav Nitin Kapadnis♠, Sohan Patnaik♠, R Raghav♠, Roshni Kar♠

♠Indian Institute of Technology Kharagpur ♢ L3S Research Center, Leibniz Universität Hannover

Background

A law practitioner has to go through numerous lengthy legal case proceedings for their practices of various categories, such as land dispute, corruption, etc. Hence, it is important to summarize these documents, and ensure that summaries contain phrases with intent matching the category of the case.

Motivation

- To the best of our knowledge, no evaluation metric that evaluates a summary based on its intent

Our Contributions

- We propose an automated intent-based summarization metric, which shows better agreement with human evaluation as compared to other automated metrics like BLEU, ROUGE-L etc.
- We curated a dataset by annotating intent phrases in two different sets of legal documents
- We also show a proof of concept as to how this system can be automated, with the help of a demo website

Dataset

Data Collection

- We scrape 5000 legal documents from CommonLII using 'selenium'
- 101 documents from the categories of Corruption, Murder, Land Dispute, and Robbery are randomly sampled from this larger set for the Indian dataset (ID)
- For the Australian dataset (AD) we downloaded the Legal Case Reports Dataset from the UCI ML repository and annotated 59 relevant documents

Data Annotation

1. **Initial filtering:** 2 annotators filter out sentences that convey an intent matching the category of the document at hand.
2. **Intent Phrase annotation** 2 other annotators extract a span from each sentence, so as to exclude details not contributing to the intent (e.g. name of the person, date of incident etc.), and only include words expressing corresponding intent. Resulting spans are the intent phrases. Overall Inter-annotator agreement (Cohen κ) is 0.79.

Category	No. of docs		Avg. no. of words/doc		Avg. no. of sentences/doc		Avg. no. of words/intent phrase	
	ID	AD	ID	AD	ID	AD	ID	AD
Corruption	19	15	2542	4613	197	264	6	6
Land Dispute	27	14	2461	11508	196	579	5	6
Murder	32	15	1560	3008	149	183	6	5
Robbery	23	15	1907	7123	162	449	4	5

Metrics

- As mentioned earlier, we propose an automated intent-based summarization metric that shows better correlation with human evaluation as compared to other evaluation metrics such as BLEU, ROUGE-L etc.
- We report the average intent-based F1 score over all the documents.
- **closePair:** A pair of intent phrase and a sentence from the summary, such that, the intent phrase is contained in the sentence is defined as a *closePair*.
- In order to derive the intent-based F1 score, we first calculate the precision and recall.
  1. **Precision:** The fraction of sentences in the summary that form a 'closePair' with atleast one intent phrase gives precision.
  2. **Recall:** The fraction of intent phrases that form a 'closePair' with atleast one sentence from the summary gives recall.
- Finally, F1 score is simply the harmonic mean of the precision score and the recall score.
- **Similarity:** Given a document, the corresponding set P of M intent phrases and output summary O consisting of N sentences, a similarity score s_{ij} between i^{th} intent phrase (P_i) and j^{th} sentence in the summary (O_j) is 1 if P_i is a phrase contained in O_j and 0 otherwise, $\forall i \in \{1, 2, .., M\}$ and $\forall j \in \{1, 2, .., N\}$.
- Mathematically,

$$s_{ij} = \begin{cases} 1, & \text{if } \exists k, P_i = O_j[k : k + \text{length}(P_i)] \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

- Similarly, the mathematical expressions of intent-based precision, recall and F1 score are as follows.

$$P_{int} = \frac{\sum_{j=1}^{N} \mathbf{1}_{\left[\sum_{i=1}^{M} s_{ij} > 0\right]}}{N} \quad (2) \qquad R_{int} = \frac{\sum_{i=1}^{M} \mathbf{1}_{\left[\sum_{j=1}^{N} s_{ij} > 0\right]}}{M} \quad (3)$$

$$F1_{int} = \frac{2.P_{int}.R_{int}}{P_{int} + R_{int}} \quad (4)$$

- In addition to the task of extractive summarization, we also validate our metrics i.e., precision, recall and F1 score on document classification task.

Experiments And Results

We carry out the following experiments in order to validate our proposed intent-based evaluation metric.

1. We use four types of summarization techniques i.e., based on Graphical Model, Letsum, Legal-Longformer Encoder Decoder, and BERT.
2. For the task of Document Classification, we observed that boosting algorithms such as AdaBoost and domain pre-trained transformer models such as LEGAL-BERT outperforms all the other models in terms of Accuracy and Macro F1-score in both the ID and AD datasets.
3. For the task of intent classification, we train JointBERT and its variants to validate our proposed evaluation metric.
4. We report automated metrics such as BLEU, METEOR, ROUGE-L, Sentence and Word Mover Similarity (S + WMS) and BERTScore along with our proposed metric for the task of extractive summarization. Some conclusions are mentioned below.
   - Graphical Model tends to preform the best for lexical metrics such as BLEU, METEOR, ROUGE-L.
   - BERT Extractive Summarizer gives the best BERTScore
   - Legal-LED performs better on ID compared to AD.
   - In case of ID, LetSum performs the best as per Intent Metric and S+WMS, while in case of AD, all models perform almost equally well w.r.t these metrics:
5. We also carry out human evaluation to validate our proposed evaluation metrics. The details of the survey are mentioned in our paper.

Model Name	BLEU		METEOR		ROUGE-L F1		BERT Score		S+WMS		Intent Metric	
	ID	AD	ID	AD	ID	AD	ID	AD	ID	AD	ID	AD
Relevance	-0.09	**-0.03**	-0.14	-0.09	0.06	-0.32	0.03	-0.18	0.25	-0.59	**0.42**	-0.05
Human Score	-0.02	**0.09**	-0.03	**0.09**	0.18	-0.21	-0.04	0.04	0.19	-0.57	**0.34**	-0.04

Fig. 2: Spearman Rank Correlation of automated metrics with human evaluation metrics

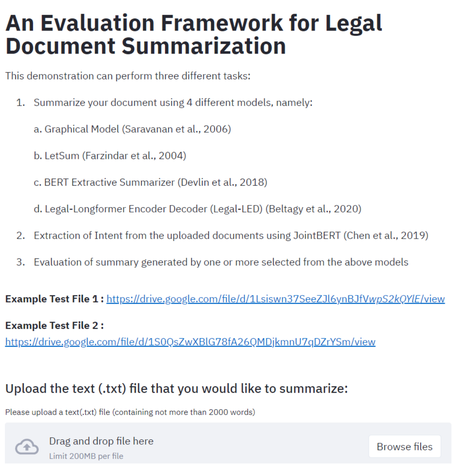

Fig. 3: Demonstration Website