Abstractive Summarization of Dutch Court Verdicts Using Sequence-to-sequence Models

Nick van de Luijtgaarden, Daniël Prijs, Marijn Schraagen*, Floris Bex
Utrecht University, The Netherlands

Abstract

With the legal sector embracing digitization, the increasing availability of information has led to a need for systems that can automatically summarize legal documents. Most existing research on legal text summarization has so far focused on *extractive* models, which can result in awkward summaries, as sentences in legal documents can be very long and detailed. In this study, we apply two *abstractive* summarization models on a Dutch legal domain dataset. The results show that existing models transfer quite well across domains and languages: the ROUGE scores of our experiments are comparable to state-of-the-art studies on English news article texts. Examining one of the models showed the capability of rewriting long legal sentences to much shorter ones, using mostly vocabulary from the source document. Human evaluation shows that for both models hand-made summaries are still perceived as more relevant and readable, and automatic summaries do not always capture elements such as background, considerations and judgement. Still, generated summaries are valuable if only a keyword summary or no summary at all is present.

1 Introduction

Given the increasing availability of legal information and the fact that many legal documents are often relatively long and dense, there is an increasing need for systems that can automatically summarize these documents. Such summaries can help not only lawyers and judges, but also citizens, companies and researchers to process case law.

Two key approaches exist for automatic summarization: *extractive summarization* involves identifying important text spans from the document and combining them into a summary, and *abstractive summarization* involves generating new sentences that explain in more general terms what the text is about (Hahn and Mani, 2000). Abstractive summaries are potentially more readable and more efficient than extractive summaries. For example, consider the following sentence:

> By letter of 18 June 2012, the appellant addressed a request to the defendant to take enforcement action against [A] Inc. and [B] Inc. for (alleged) violation of the provisions of the Quarantine Facilities for Live Bivalve Molluscs Regulation 2007.

An abstractive model can retain only the information that the appellant requested enforcement action based on the Quarantine Facilities for Live Bivalve Molluscs Regulation 2007, while an extractive model would retain the full sentence. In this paper we apply a reinforcement learning approach with a biLSTM (referred to as **RL**) as well as a deep learning approach based on the BART (Lewis et al., 2019) transformer model (referred to as **BART**) to abstractive summarization of the Dutch case verdict database Rechtspraak.nl. We show that generated summaries are useful, but not yet on par with human-generated summaries.

The rest of this paper is structured as follows. In Section 2 we consider current state-of-the-art models for legal summarization, Section 3 describes our dataset, and Section 4 discusses the design and implementation of the deep learning pipelines. Here we also highlight the different evaluation methods used: in addition to the common ROUGE metric, we also look at abstractiveness (See et al., 2017), i.e., the amount of novelty introduced in the wording of the summary, and perform a human evaluation on the aspects of summary relevance and readability. Finally, Section 5 and 6 will elaborate upon the results and implications of this research.

*Corresponding author, M.P.Schraagen@uu.nl.
Code and data are available via https://git.science.uu.nl/n.vandeluijtgaarden/legal-text-summarization (RL models) and https://github.com/prijsdf/dutch-legal-summarization (BART models).

Proceedings of the Natural Legal Language Processing Workshop 2022, pages 76 - 87

2 Related work

Legal text differs from common document types such as news articles (Kanapala et al., 2017), which has prompted the development of custom word embeddings for legal vocabulary in English (Chalkidis and Kampas, 2019; Chalkidis et al., 2020). However, general pre-trained embeddings or embeddings trained on-the-fly are also commonly used. Early approaches used pattern-based heuristic segmentation approaches (Uyttendaele et al., 1998; Farzindar and Lapalme, 2004). Machine learning was used by Hachey and Grover (2006) to classify sentences as different legal rhetorical structures (Moens and de Busser, 2002) such as fact, proceedings or background. Saravanan et al. (2006) applied probablistic graphical models based on Conditional Random Fields (CRF) to segment and label a legal decision given various rhetorical roles (e.g., argument or final decision). Yousfi-Monod et al. (2010) used a Naive Bayes algorithm with a set of heuristic features to identify sections (introduction, context, reasoning, conclusion) and create a summary. Instead of identifying specific sections or elements, Galgani et al. (2012) use a rule-based approach, where rules created by domain experts are used to identify important phrases in a decision. More recently, Zhong et al. (2019) create summaries by classifying sentences in a decision as, e.g., issues, decision, etc. Similarly, Xu et al. (2020) use a number of different machine learning techniques to classify the issues, conclusions and reasons in a legal verdict. Liu and Chen (2019) use an LSTM classifier on sentences from the 'reasoning' section of Taiwanese Supreme Court judgements to determine which sentences belong to the 'gist' of the judgement, achieving an F1-score of around 0.9. Eidelman (2019) used a combination of supervised sentence-scoring and TF-IDF in an ensemble method on their BillSum dataset. Regarding abstractive approaches, Bhattacharya et al. (2019) use the pointer model by See et al. (2017) on their Supreme Court of India dataset. Zhang et al. (2020) use their pre-trained PEGASUS language model to generate abstractive summaries on the BillSum dataset of Eidelman (2019). Previous work on abstractive summarization of UK court verdicts was performed by Ray et al. (2020).

3 Data

For the current research, data from the Dutch judicial system is used. On average, around 1.6M cases are handled in The Netherlands every year, of which a small percentage is published on the official website Rechtspraak.nl. For the RL experiments in this research a pre-processed version of the Rechtspraak data provided by Pandora Intelligence[1] is used, providing easy access to the type, summary and verdict of each case. For the BART experiments a separate preprocessing pipeline is used that exposes only the case text and the summary. In total, this dataset contains around 430K legal court cases. 94% of these cases contain a summary, and we included only these cases in the data exploration discussion in the current section. An example document can be found in Table A1.

On average, case texts contain $\sim$ 650 words with summaries of length $\sim$ 50. However, a significant amount of summaries has less than 25 words, containing only keywords or a single sentence. A small amount of summaries is over 250 words long. To provide more uniform data to the models, for training we used only cases that have a summary containing between 40 and 150 words, and consisting of a minimum of three and a maximum of six sentences with at least 5 words in each sentence. Note that very short summaries are reintroduced in the dataset for human evaluation.

4 Research Method

We use two deep learning pipelines on the Rechtspraak.nl data: a hybrid reinforcement learning method and a transformer-based method.

4.1 RL model

Chen and Bansal (2018) have proposed a hybrid extractive-abstractive model that first selects important sentences (similar to extractive summarization) and then rewrites them abstractively. First, sentences are represented using a temporal convolutional model and words are converted to a distributed vector representation using word embeddings. Sequences of word vectors are fed through the layers of the model to capture the dependencies of nearby words. Selection of sentences from the sentence representations is then done by training a pointer network based on a set of features (Vinyals et al., 2015), and these extracted sentences are then subsequently compressed and paraphrased by an abstractive model to create a concise summary sentence (see Figure 1).

[1] https://www.pandoraintelligence.com/

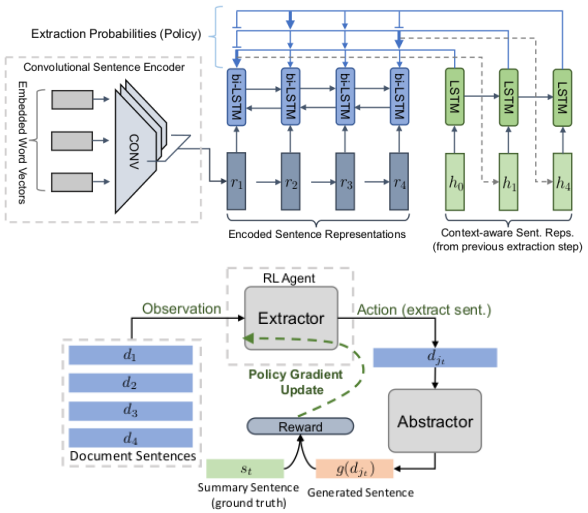

Figure 1: RL model architecture, showing the extractor component (top) and the full architecture (bottom). Images reproduced from Chen and Bansal (2018).

We use this hybrid model on legal data in the current study, arguing that the extractive part of the model can help retain the core facts of the verdicts, while the abstractive part of the model can make the summary shorter and more readable.

The data processing pipeline consists of a number of steps. First, data from Rechtspraak is loaded based on the filtering described in Section 3, and tokenized using Ucto (van Gompel et al., 2012) and Stanford CoreNLP (Manning et al., 2014). Gensim (Rehurek and Sojka, 2010) is used to create word embeddings through Word2Vec. The network itself is trained using the PyTorch framework and CUDA.

The Extractor component (shown in the top of Figure 1) consists of multiple steps. First, word embeddings for all words in a sentence are combined into a sentence representation r_j using a convolutional layer. Then, an encoding step using a bi-directional LSTM layer transforms the sentence representation into a contextual representation h_j using the surrounding sentences. Finally, in a decoding step an LSTM computes the extraction probability of a sentence based on the contextual embedding h_j. The training target for the extraction probability is to minimize the ROUGE distance between the extracted sentence and the reference summary, i.e., the component learns to extract a sentence if there is a similar sentence somewhere in the reference summary.

After a sentence is selected it is processed by the Abstractor component. This component is a sequence-to-sequence model using a bi-directional LSTM as encoder and a unidirectional LSTM as decoder, trained with the objective of transforming the extractive input sentence into the corresponding sentence in the reference summary. The resulting sentence is evaluated by the reinforcement learning component. If a suitable sentence is selected and correctly rewritten, then a reward is generated to reinforce the Extractor component. Conversely, if the similarity between the generated sentence and the reference summary is low, the Extractor component receives negative reward and learns not to extract this sentence.

For training, we used batches of 4 samples and set the checkpoint frequency (number of update steps for checkpoint and validation) on 3000 for the abstractor/extractor network and 300 for RL training. For the abstractor and extractor network we used a batch size of 32. Word2Vec embeddings were trained with a vector size of 128 and a vocabulary of 30,000. Sentence generation was limited to 30 tokens with a beam size of 5. Learning rate for the Adam optimizer is set on 0.001 for maximum likelihood (ML) objectives and 0.0001 for RL training. We set the discount factor for RL on 0.95 and cut the learning rate in half when validation loss stops decreasing, in order to speed up convergence. Gradient clipping is used to prevent exploding gradients and uses a 2-norm of 2.0 for all LSTMs. We use a network of 256 hidden units with one layer. Following the training phase, new summaries are generated for all documents in the test set.

4.2 BART model

Lewis et al. (2019) introduced BART as an autoencoder for pretraining sequence-to-sequence models for various downstream tasks, such as machine translation, question answering and summarization. The model uses the following tasks for pretraining:
Token masking Similar to BERT, a percentage of tokens in the text are masked at random and the model has to reconstruct the original text.
Sentence permutation The text is split-up in sentences (based on full stops) and then these sentences are shuffled. The model has to reconstruct the text.
Document rotation A new start token is picked at random and the document is rotated such that it starts with this new token. Again, the model has to reconstruct the original text.
Token deletion Tokens are deleted from the text.

The model needs to identify the positions of the deleted tokens.

Text infilling Similar to masking, but here random spans of texts are replaced by a single mask token. The spans mostly have a length of 0 to 9 tokens. Spans of zero length can also be replaced, which is equal to inserting a mask token into the text.

For applying the BART sequence-to-sequence model to the legal dataset, the model was pretrained from scratch using the model configuration described by Lewis et al. (2019) as implemented in the Huggingface library in Python. For pretraining the model and the tokenizer, we used the 'tiny' subset (6B words) of the Dutch part of the mC4 dataset[2] that contains a broad variety of web crawl data. Pretraining was performed on 4 million examples during one epoch with a batch size of 8 (i.e., 500k steps in total). Note that pretraining from scratch was a practical consideration. While a Dutch language model for BART already exists (Liu et al., 2020) this model was too large to be used with our setup, therefore we opted for an additional pretraining step. After pretraining, the model was fine-tuned using 70,140 court verdict documents for 10 epochs with a batch size of 8 (i.e., 88k steps). Then the actual summaries were generated on 9.9k test documents with a minimum length of 40 tokens, a maximum length of 150 tokens, a length penalty of 2.0 and a beam search of size 4. The length constraints were empirically chosen as sensible values for producing useful summaries.

4.3 Human evaluation

For the RL experiments 10 documents were sampled from the dataset and rated on a scale between 1–10 on content and readability (see Table 1), similar to (See et al., 2017; Chen and Bansal, 2018). Two law students were recruited to act as subject matter experts. The participant is asked to read and study a case for 15 minutes, then the generated summary and the reference summary from Rechtspraak are presented (without disclosing the source of the summaries). The participant is asked to provide content and readability ratings as well as a short explanation for their answer. After rating five cases with full reference summaries, another five cases with keyword-only reference summaries were presented. For these five cases the participant is asked whether they prefer the full generated summary or the keyword-only reference summary, again without disclosing the source.

For the BART experiments 40 documents were sampled from the dataset and evaluated by one of the authors. Evaluation was performed on the aspects *informativeness, relevance, fluency, coherence* as defined in Table 1 on a 5-point Likert scale. First, the evaluator read the summary and rated fluency and coherence. Then the full case text was read in order to rate informativeness and relevance of the summary.

4.4 Automatic Evaluation

Results are evaluated using standard ROUGE-1, ROUGE-2 and ROUGE-L F1 measures. The dataset is divided in a random split of 70% (training), 15% (validation) and 15% (test) cases. Hyperparameter tuning is performed on the validation set, while actual evaluation is performed on the test set. For the ML experiments the ROUGE evaluation takes the category of cases and year ranges into account, while for the BART experiments this information was not available.

Furthermore, we evaluate the abstractiveness of the generated summaries, defined as the novel n-gram count of our model compared to the reference summary. This measurement allows us to assess whether our model is actually generating new sentences, as well as whether it writes summaries in a different style compared to the reference summary.

5 Results

5.1 ROUGE evaluation

Table 2 shows the ROUGE scores for both models. ROUGE-1 and ROUGE-L are higher compared to ROUGE-2.

For the RL model specific law categories and dates were available (Table 3). Administrative Law performs best on ROUGE scores, while Private Law performs worst. A possible explanation for this difference is that Administrative Law cases are the largest category in the dataset and the reference summaries for this category relatively long, therefore the model gets a large exposure to this category during training. Regarding time periods, the model seems to perform best on cases between 2001 and 2008, while performing slightly worse on cases from the last decade. This is surprising, because the majority of documents in the dataset belongs to the most recent time period.

[2] https://huggingface.co/datasets/yhavinga/mc4_nl_cleaned

Content	Does the summary contain all important information of the original case description? Does it avoids generating repeated and redundant information?
Readability	Is the summary fluent, grammatical, of suitable length?
Informativeness	How well does the summary capture the key points of the article?
Relevance	Are the details provided by the summary consistent with details in the article?
Fluency	Are the individual sentences of the summary well-written and grammatical?
Coherence	Do phrases and sentences of the summary fit together and make sense collectively?

Table 1: Human evaluation metrics

Model	dataset	ROUGE–1	ROUGE–2	ROUGE–L
RL	Rechtspraak	37.24	16.20	34.07
BART	Rechtspraak	**46.52**	**33.74**	**44.88**
BART	CNN/Daily mail	44.16	21.28	40.90

Table 2: ROUGE scores for the summarization models

5.2 Abstractiveness

Following See et al. (2017), for the RL summaries we compute an abstractiveness score as the ratio of novel n-grams in the generated summary. Figure 2 shows the abstractiveness scores of our model compared to the reference summaries of Rechtspraak. One can see that the RL model generates very different summaries from the reference summaries. For example, 20% or less of 2,3 and 4-grams in our generated summaries are identical to the reference summary. The figure furthermore shows that abstractiveness decreases when more training examples are presented to the model, whereas ROUGE F1 increases. One can argue that as models get more abstractive, ROUGE becomes less suitable to evaluate the quality of a summary.

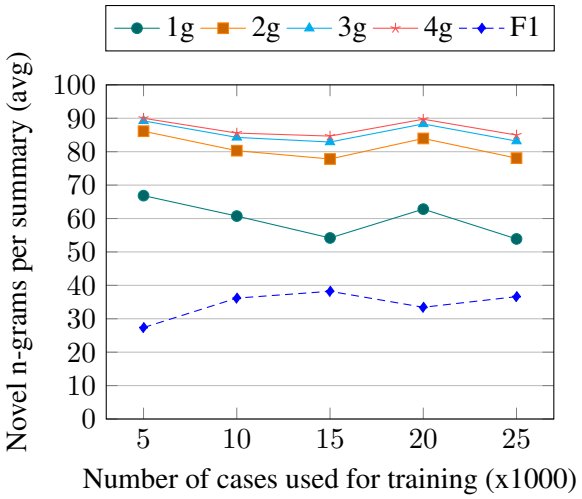

Figure 2: Ratio of novel n-gram counts of summaries compared to the reference summary by number of training cases and development of ROUGE F1

Using a manual check on a sample of the resulting summaries, we observed that the model extracts many sentences from the input document itself. When looking at sentences with similar 4-grams, the model actually used much larger n-grams from the text. However, the model did rewrite and shorten many sentences, thus improving the readability of the text. In addition, redundant information from sentences was removed properly, which made sentences more concise. However, we did note that the model occasionally tends to remove relevant facts and details from sentences, which are needed to understand the case (cf. Figure A1).

5.3 Qualitative evaluation

Exploratory qualitative evaluation by the authors indicated that the model does not introduce many novel sentences. Still, it shows good results for rewriting sentences and removing redundant details from the case, while preserving grammaticality in generated sentences. When important facts (e.g numbers and dates) are present, the model is likely to include these facts in the summary. However, sometimes the model recognized words as not important, which may be caused by the fact that the model has not seen these words often during training. This leads to sentences being cut off too fast. Also, the summary did not always include all elements that are needed in a summary (background, considerations and judgement).

In the example shown in Figure A1, the RL model first gives a very short background description of the case and describes one of the considerations. The description of the background of the

	ROUGE-1	ROUGE-2	ROUGE-L
Administrative Law	39.26 (38.82, 39.74)	18.05 (17.49, 18.64)	35.99 (35.54, 36.49)
Private Law	32.83 (32.30, 33.36)	10.72 (10.24, 11.26)	29.46 (28.96, 29.98)
Criminal Law	37.54 (36.94, 38.12)	17.48 (16.76, 18.18)	34.73 (34.13, 35.34)
Tax Law	36.46 (35.63, 37.32)	13.46 (12.59, 14.42)	32.83 (32.03, 33.70)
1970-2000	38.91 (35.10, 43.05)	16.64 (12.35, 21.37)	35.23 (31.39, 39.32)
2001-2008	38.86 (38.19, 39.53)	18.49 (17.61, 19.38)	35.63 (34.93, 36.32)
2009-2018	36.59 (36.23, 36.94)	15.34 (14.92, 15.74)	33.47 (33.10, 33.83)

Table 3: Observations from the RL experiments (ROUGE F_1 with and 95% confidence interval)

case is very short and the consideration is discussed in far too much detail. Also, the judgement of the case is not discussed. In the first sentence, the main subject of the case (Quarantine Facilities for Live Bivalve Molluscs Regulation 2007) is removed, likely because the model has not seen this word before in other documents and thus does not deem it important. Apart from this mistake, the model does a very good job at rewriting the sentence to a more clearer one. In the second sentence the article number and name of the relevant regulation is omitted in the summary. In the third sentence, many unnecessary details are removed. Still, the summary does not include the (important) fact that the defendant was accused of this case and not that they were actually found guilty. The fourth sentence is taken directly from the original case text. In general, the summary goes in too much detail on some parts of the case and fails to give a generalized summary. However, this example does show the power of the model to rewrite sentences into much clearer and shorter ones.

Qualitative exploration of the BART results indicated that some summaries are able to improve on the reference summary significantly in conciseness while retaining all important facts (see Figure A2). Other summaries however seem to go off on the wrong track, and expand on an unimportant detail for several sentences while missing key points.

In future work, postprocessing could help to fix some of the mentioned issues, using a template-based approach where elements from the original text are copied into the generated summary if the model fails to provide specific details (cf. (Xu et al., 2020)).

5.4 Human evaluation

Table 4 describes the results of the human evaluation phase, showing that the participants in the RL experiment found the reference summaries more relevant and more readable compared to the generated summaries, however the variance of the responses was relatively high. For keyword-only reference summaries all participants preferred to use the generated summary. Participants explicitly noted that they disliked a case to be summarised using only keywords, as this approach is much too abstract for the legal sector. In the BART experiments the content rating was split between the aspects of informativeness and relevance, while the readability rating was split between the aspects of fluency and coherence. For all aspects the evaluator rated the reference summary higher than the generated summary, similar to the RL results.

For the RL experiments, participants noted that not all elements needed in the summary (background, considerations and judgement) were always included. For example, some summaries included the facts and the judgements of a case, but failed to explain the considerations. However, the evaluation showed that, to a lesser extent, also reference summaries are found to omit fundamental details of cases. Participants mentioned that the text that was in fact included in the generated summaries was relevant for the case (which could be compared to a high precision and low recall of information content in the summaries). Regarding readability, participants observed that sentences in the generated summary occasionally contained grammatical errors or ended strangely. Also, the order of sentences was criticized, in both generated and reference summaries, as some summaries started with the judgement and ended with background information about the case. For the summaries generated by the BART model similar observations can be made about content aspects, i.e., the topics discussed in the summary are relevant but not all important aspects are always included by the models, which was again also observed for the reference summaries. For the BART model issues

Summary	n	Content		Readability	
		Informativeness	Relevance	Fluency	Coherence
RL generated	10	4.60 ± 2.12		5.55 ± 1.67	
RL reference	10	6.65 ± 1.63		7.00 ± 1.63	
BART generated	40	3.58 ± 1.24	4.03 ± 1.19	4.45 ± 0.90	4.10 ± 1.08
BART reference	40	4.13 ± 1.04	4.80 ± 0.61	4.75 ± 0.67	4.45 ± 0.81

Table 4: Results of the human evaluation experiment (mean and standard deviation). RL summaries are rated on a scale of 1–10, BART summaries are rated on a scale of 1–5.

in fluency and coherence were noticed, however this model suffered less from obvious grammatical errors or cut-off sentences.

Due to constraints on time and resources in this research project the number of participants was small, leading to large confidence intervals and only a small number of data points. For future work, it would be advised to have a much larger group of subjects, which would also allow to test different versions of our model by changing filters or hyperparameters. Also, it can be interesting to use experienced legal professionals in this type of research, instead of evaluation by law students (RL model) or the paper authors (BART model).

In this evaluation we have seen that there are some issues relating to relevance and readability, such as grammatical errors or missing content. For improving readability, a parser could be implemented in the decoding function that can give a signal when a sentence is cut off too early, giving this sentence a lower score in the beam search algorithm. Also, post-processing can fix some problems regarding nouns, as the models did not always use these correctly when generating sentences.

For improving relevance, an implementation of a neural network that can identify the three core elements needed in a summary can prove useful. Alternatively, a clustering algorithm can be used to find diverse topics in the text, and then identify the most important sentences in these clusters.

6 Conclusion & Discussion

In this work, a dataset containing over 400K Dutch court verdicts was used to train a hybrid reinforcement learning-based model, as well as a transformer-based BART model. We evaluated generated summaries based on ROUGE, abstractiveness, and through a human evaluation experiment using legal experts. Our experiments report an F1 score of 46.52 (ROUGE-1), 33.74 (ROUGE-2) and 44.88 (ROUGE-L) for the BART model, comparable to state-of-the-art results achieved on the CNN/Daily Mail dataset.

The models did not introduce many novel n-grams, but showed good performance in rewriting and shortening sentences. The evaluation also showed the potential to improve the model, following observations that the model may cut sentences off too early and does not always include all elements (background, considerations and judgement) in the summary. Furthermore, while important facts were generally included and the rewriting process showed adequate results, still unnecessary case details are found in the generated summaries.

Considering the level of abstractiveness the models showed the capability of rewriting long and redundant sentences found in legal text to much shorter ones. Quantitatively it was shown that the model generates a large number of novel n-grams compared to the reference summaries from the dataset. Due to the inverse relationship of the abstractiveness and ROUGE score of a document, a good performance in producing novel n-grams actually reduces the score on the summary quality evaluation measured with ROUGE, which was confirmed by the analysis of the evaluation results. Therefore, we argue that ROUGE scores are not fully representative as a metric for abstractive summarization. While alternative methods are being developed (Zhang et al., 2019; Yuan et al., 2021) ROUGE is likely to remain an influential evaluation approach, however these results should be interpreted carefully when comparing models and approaches.

To complement ROUGE scores, a human evaluation study was conducted to evaluate both generated and reference summaries on readability and content. Especially for the RL model the results show a large difference in relevance between reference summaries (6.7/10) and generated summaries (4.6/10), and a slightly smaller difference in readability (7.0/10 vs 5.6/10). However, the participants

in the RL experiments noted that the generated summaries did contain key information about the case and preferred it to using a reference summary consisting of only keywords. For the BART experiments the difference across the four dimensions informativeness, relevance, fluency and readability were perceived to be smaller but still the reference summaries were preferred for all dimensions.

It has been argued in the literature that an abstractive summary may be less accurate and can lead to misinterpretations of a judge's intent (Yousfi-Monod et al., 2010). Furthermore, as argued by Jain et al. (2021), there are many citations (to e.g. previous cases or articles of law) which cannot be ignored. However, with more data being available, improved hardware and matured algorithms, the accuracy of abstractive models is increasing. Furthermore, citations can often be extracted from legal texts using basic regular expressions and the relevant legal articles or precedent cases can be provided as metadata, which can then be presented in combination with the abstractive natural language summary. Furthermore, we argue that even less accurate summaries can be useful as a tool for quickly searching through huge databases of cases. Furthermore, there is also the possibility to combine abstractive models with more domain-specific constraints, such as citing law articles and structuring the summary into facts, arguments and decision.

This study fills the following gaps in current research on (legal) text summarization. First, very few research on legal summarization has made use of an abstractive summarization model. The authors are aware of two approaches only, of which the first shows comparatively low evaluation scores (Bhattacharya et al., 2019), and the second is based on US Congressional Bills (Zhang et al., 2018), which, while they can be considered legal documents, are rather different from the case verdicts and decisions we consider. For example, Bills – essentially numbered lists of laws and statutes – are much more structured than verdicts, and the language used in Bills is much more generic because it does not pertain to individual cases like verdicts.

Second, like (Zhang et al., 2020), our work shows that unsupervised neural models originally developed for news articles can be successfully used on legal documents, which differ significantly from news articles both in terms of length and in terms of internal structure and distribution of relevant content elements. Furthermore, no previous research has applied an abstractive summarization model on a dataset of legal documents in Dutch, showing that our unsupervised language models are robust considering the legal language of the documents presented to the model.

For both models, long texts are still difficult to process due to technical limitations on input representation. A case verdict document can easily surpass such length constraints and will be truncated (e.g., to 1024 words) as a result prior to summarization. With respect to future research, models designed to process longer text (Beltagy et al., 2020; Yang et al., 2020) therefore seems promising. Also, even though pre-trained language models are known for their ability to generalize across domains, the model of (Zhang et al., 2018) used to obtain the high levels of performance on Congressional Bills shows a relatively average performance on the CNN/Daily Mail news dataset, which might support the hypothesis that the document structure (rather than the model itself) is the predominant factor for the summary evaluation scores. Applying the current two methods on the Congressional Bills would provide more insight into the reasons behind the performance differences.

Currently, the RL model uses static Word2Vec word embeddings created on the fly on the Rechtspraak dataset. In contrast, the BART contextual embeddings were pretrained on the general-purpose C4 dataset. While the BART model already outperforms the RL model by a significant margin for both ROUGE scores and human ratings, it would be interesting to investigate whether pretraining BART on domain-specific data (i.e., Dutch legal text) would result in an additional performance increase. A practical problem however is data availability: the C4 subset currently used contains 6B words of crawled web pages, which is difficult to match with Dutch legal text.

Other future work includes a more detailed analysis of summaries generated by the BART model. We have observed that the overall quality of the BART summaries is higher compared to the summaries generated by the RL model, in terms of grammaticality and topicality. It would be interesting to compare the detailed observations made for the RL model, such as the abstractiveness, relation to law categories and time frames, and missing legal aspects in the summaries, to establish whether the BART model supports these observations as well.

References

Iz Beltagy, Matthew Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Paheli Bhattacharya, Kaustubh Hiware, Subham Rajgaria, Nilay Pochhi, Kripabandhu Ghosh, and Saptarshi Ghosh. 2019. A comparative study of summarization algorithms applied to legal case judgments. In *European Conference on Information Retrieval*, pages 413–428. Springer.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Ilias Chalkidis and Dimitrios Kampas. 2019. Deep learning in law: early adaptation and legal word embeddings trained on large corpora. *Artificial Intelligence and Law*, 27:171–198.

Yen-Chun Chen and Mohit Bansal. 2018. Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting. *arXiv:1805.11080 [cs]*. ArXiv: 1805.11080.

Vladimir Eidelman. 2019. Billsum: A corpus for automatic summarization of us legislation. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 48–56.

Atefeh Farzindar and Guy Lapalme. 2004. Legal Text Summarization by Exploration of the Thematic Structure and Argumentative Roles. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 27–34, Barcelona, Spain. Association for Computational Linguistics.

Filippo Galgani, Paul Compton, and Achim Hoffmann. 2012. Combining Different Summarization Techniques for Legal Text. In *Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data*, HYBRID '12, pages 115–123, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ben Hachey and Claire Grover. 2006. Extractive summarisation of legal texts. *Artificial Intelligence and Law*, 14(4):305–345.

U. Hahn and I. Mani. 2000. The challenges of automatic summarization. *Computer*, 33(11):29–36.

Deepali Jain, Malaya Dutta Borah, and Anupam Biswas. 2021. Summarization of legal documents: Where are we now and the way forward. *Computer Science Review*, 40:100388.

Ambedkar Kanapala, Sukomal Pal, and Rajendra Pamula. 2017. Text summarization from legal documents: a survey. *Artificial Intelligence Review*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Chao-Lin Liu and Kuan-Chun Chen. 2019. Extracting the gist of Chinese judgements of the Supreme Court. In *Proceedings of ICAIL '19*. ACM.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *arXiv*.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

Marie-Francine Moens and Rik de Busser. 2002. First steps in building a model for the retrieval of court decisions. *International Journal of Human-Computer Studies*, 57(5):429–446.

Oliver Ray, Amy Conroy, and Rozano Imansyah. 2020. Summarisation with majority opinion. In S. Villata et al., editor, *Legal Knowledge and Information Systems: JURIX 2020*, pages 247–250. IOS Press.

Radim Rehurek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50. ELRA.

M. Saravanan, B. Ravindran, and S. Raman. 2006. Improving Legal Document Summarization Using Graphical Models. In *Proceedings of the 2006 Conference on Legal Knowledge and Information Systems: JURIX 2006: The Nineteenth Annual Conference*, pages 51–60, Amsterdam, The Netherlands, The Netherlands. IOS Press.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. *arXiv:1704.04368 [cs]*. ArXiv: 1704.04368.

Caroline Uyttendaele, Marie-Francine Moens, and Jos Dumortier. 1998. SALOMON: Automatic abstracting of legal cases for effective access to court decisions. *Artificial Intelligence and Law*, 6:59–79.

Maarten van Gompel, Ko van der Sloot, and Antal van den Bosch. 2012. Ucto: Unicode Tokeniser. Technical Report 12-05, ILK.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer Networks. In C. Cortes, N. D.

Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2692–2700. Curran Associates, Inc.

Huihui Xu, Jaromír Šavelka, and Kevin Ashley. 2020. Using argument mining for legal text summarization. In S. Villata et al., editor, *Legal Knowledge and Information Systems: JURIX 2020*, pages 184–193. IOS Press.

Liu Yang, Mingyang Zhang, Cheng Li, Mike Bendersky, and Marc Najork. 2020. Beyond 512 tokens: Siamese multi-depth transformer-based hierarchical encoder for long-form document matching. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20)*, pages 1725–1734. ACM.

Mehdi Yousfi-Monod, Atefeh Farzindar, and Guy Lapalme. 2010. Supervised Machine Learning for Summarizing Legal Documents. In *Advances in Artificial Intelligence*, Lecture Notes in Computer Science, pages 51–62. Springer.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating generated text as text generation. In *35th Conference on Neural Information Processing Systems (NeurIPS 2021)*. Curran Associates, Inc.

Jianmin Zhang, Jiwei Tan, and Xiaojun Wan. 2018. Towards a Neural Network Approach to Abstractive Multi-Document Summarization. *arXiv:1804.09010 [cs]*. ArXiv: 1804.09010.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT.

Linwu Zhong, Ziyi Zhong, Zinian Zhao, Siyuan Wang, Kevin D. Ashley, and Matthias Grabmair. 2019. Automatic summarization of legal decisions using iterative masking of predictive sentences. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, ICAIL '19, page 163–172. Association for Computing Machinery.

Appendix: summary examples

Case (ECLI:NL:CBB:2013:212)
. . . . Process By letter of 18 June 2012, the appellant addressed a request to the defendant to take enforcement action against [A] B.V. and [B] B.V. for (alleged) violation of the provisions of the Quarantine Facilities for Live Bivalve Molluscs Regulation 2007. By letter dated 19 June 2012, the defendant responded. By letter of 27 June 2012, the appellant repeated the request to take enforcement action. By letter of 9 July 2012, the defendant responded again. By letter of 3 August 2012, the defendant responded to the appellant's subsequent letters of 24 and 27 July 2012. . . . Considerations . . . 3.1 Is it not disputed that the appellant has requested the defendant to initiate disciplinary proceedings for (alleged) breach of the prohibition in Article 2(1) of the Regulation on quarantine facilities for live bivalve molluscs 2007. Appellant subsequently served notice on Defendant for failing to decide to initiate disciplinary proceedings, she claims, and then brought an appeal under Article 6:12 of the General Administrative Law Act (Awb). Which first of all raises the question of whether an appeal under Article 6:12 of the General Administrative Law Act was open to the appellant in this case. The Board answers this question in the negative. Article 15 of the Corporate Disciplinary Procedure (Organisation) Act 2004 stipulates that the chairperson of the defendant, once he or she has established that there has been an (alleged) offence, shall refer the matter to the disciplinary court within a reasonable period of time by means of a written statement. . . . .
Reference summary
Refers to appeals against failure to take a decision on time. Such proceedings are not open against the chairman of the Dutch Fish Product Board not initiating disciplinary proceedings. Making such a request is not an application within the meaning of Article 1:3 of the General Administrative Law Act. The corporate organisation disciplinary law has its own provisions.
Generated summary
Request for enforcement action. Not in dispute is that appellant has requested the defendant to initiate disciplinary proceedings for (alleged) breach of the prohibition in Article 2, first. Defendant failed to decide on the filing of a disciplinary case. Which raises the question of whether recourse to Article 6:12 of the General Administrative Law Act was open to appellants in this case.

Figure A1: Example from the dataset showing the generated summary of the RL model. The example is translated from Dutch into English for illustration purposes. The colored (marked) sentences correspond to the sentences that were extracted by the model.

Case (ECLI:NL:CBB:CRVB:2012:BW3333)
[...] Considerations [...] The job expert has selected functions which he considers the appellant capable of performing and has established the loss of earning capacity to less than 35%. As decided on January 7th, 2010, it is established that appellant starting October 20, 2009 is not entitled to benefits regarding the Law Work and Income. The insurance physician has concluded that the impairments are established correctly. He has studies the letter by the general practitioner dated December 31, 2009. According to the insurance physician there are no grounds to assume further impairments. [...] The court has judged that the decision was based on a sufficiently appropriate medical foundation. There is no reason to doubt the conclusion by the insurance physicians that appellant is capable of performing simple, routine tasks. [...] For the appeal appellant has entered information by the general practitioner and treating psychologist. [...] The insurance physician has stated that this information is not a reason to assume further impairments. The Council follows this conclusion. [...] Establishing sickness- or handicap-based impairments to performing labor based on the medical condition of the appellant should in principle be considered the expertise of the insurance physician. [...] Appellant is rightfully considered to be able to perform the selected tasks. [...] It did not become apparent that the workload exceeds the capabilities of the appellant.
Reference summary
Refusal to grant "WIA Law" benefits. The contested decision is based on a sufficiently appropriate medical foundation. The information from the general practitioner and psychologist do not substantiate that the appellant is more impaired than the insurance physician has established. Establishing sickness- or handicap-based impairments to performing labor based on the medical condition of the appellant should in principle be considered the expertise of the insurance physician. Appellant is rightfully considered to be able to perform the selected tasks. It did not become apparent that the workload exceeds the capabilities of the appellant.
Generated summary
Refusal to grant WIA benefits. Less than 35% incapacitated. Sufficient medical and employment-related foundation. No reason to doubt the conclusion by the insurance physicians that appellant is capable of performing simple, routine tasks.

Figure A2: Example from the dataset showing the generated summary of the BART model for an appeal case. The example is translated from Dutch into English for illustration purposes. The colored (marked) sentences correspond to sentences in the generated summary. It can be observed that the reference summary is almost completely extractive, while the BART summary contains both abstractive and extractive sentences.