Legal Knowledge and Information Systems
E. Francesconi et al. (Eds.)

doi:10.3233/FAIA220474

Legal Text Summarization Using Argumentative Structures

Bianca STEFFES [a,1], Piotr RATAJ [a,b]
[a] *Saarland University, Saarland Informatics Campus, Saarbrücken, Germany*
[b] *Zentrum für Recht und Digitalisierung, Saarbrücken, Germany*

Abstract. Legal text summarization focuses on the automated creation of summaries for legal texts. We show that the argumentative structure of judgments can improve the selection of guiding principles as a specific kind of summary using judgments of the German Federal Court of Justice as measured by the ROUGE metric. We evaluate our first results and put them in the context of our ongoing work.

Keywords. legal text summarization, German Federal Court of Justice, guiding principles, extractive summarization

1. Introduction

Text summarization algorithms allow automated creation of summaries for arbitrary texts. Especially the legal domain, in which long and complex documents are ubiquitous, might benefit from such algorithms on a large scale. Yet the summarization of legal documents is still confronted with some unsolved problems: the domain specific knowledge and structure seems to hinder simple porting of domain independent summarization algorithms to the legal field (e.g. [1], [2]) and the sheer length of documents and sentences challenges neural models (e.g. [3]). In this work we tackle the summarization of judgments delivered by the German Federal Court of Justice (Bundesgerichtshof, BGH) by automatically selecting guiding principles of the judgments.

1.1. Character of Guiding Principles

Guiding principles (*Leitsätze*) are, roughly speaking, very short formulations of or at least introductions to the main and "important" normative statement(s) that a court finds when deciding a particular case. Such a principle could be: "A will is invalid when written on a computer.". Their function is to quickly brief the reader and provide some orientation regarding the judgment. Although guiding principles are issued by many (German) courts, we only take into account judgments of the BGH in civil matters that contain such guiding principles (which is not the case for all of its judgments). As the BGH is a court of final instance and, thus, judicial review is limited to (important) issues concerning the

[1] Corresponding Author: Bianca Steffes, bianca.steffes@uni-saarland.de.

interpretation of the law (versus the facts) we assume that, in most cases, a large part of the court´s reasoning will be somehow reflected in its guiding principles.

Note that guiding principles, at least the ones that we pick, are issued directly by the court's body that decides the case. This guarantees a high quality of the data. With respect to their content in detail, however, there are no formal rules. In practice, we observe two basic forms in which guiding principles are stated by the court and label them as *topical* and *propositional* respectively: while the former simply introduce the decision's main legal topic (e.g., "On the conditions of a valid will when written on a computer."), the latter contain a normative statement (as is the case with our example from above: "A will is invalid when written on a computer."). We only take into account propositional ones.

Since guiding principles contain the main statement(s) regarding the court´s interpretation of the law they reflect the normative conclusion of its argumentation. On a structural level, this implies that only a specific part of the—rigorously structured—judgment needs to be considered (see further *infra* 4.1.). Thematically, this means that the guiding principles concern a different order than the argumentation justifying them. Furthermore, the guiding principles as well as the argumentation refer to the facts of the case mostly indirectly, allowing us to focus on the normative / legal domain and its language. With this in mind we can think of guiding principles as a specific kind of legal summary of a judgment.

1.2. Related Work

Existing algorithms for text summarization can be roughly grouped into two classes: abstractive and extractive approaches. Abstraction-based algorithms create summaries by paraphrasing the content of a text, while extraction-based methods select sentences from the original document as a summary. In the legal domain, extractive algorithms are most common.

LetSum [4] and *DelSumm* [5] are extractive algorithms that create summaries by, firstly, mapping all sentences to structural parts of a judgment (e.g., facts, reasoning) with the aim of representing each of these parts in their summaries. The sentences which will later constitute the summary are then selected by a tf*idf based ranking and a specific scoring. Both of these categorizations are not applicable to our problem as, for one, they categorize the sentences of a whole judgments while German judgments are published in a similar categorization already, and on the other hand, we intend to work on only one of those categories (reasoning) which they do not provide a structuring for. The scoring of *DelSumm* is highly based on the mentioned structuring, thus, it does not fit our problem.

The *MMR algorithm* [6] selects the sentences for the summary based on how predictive they are for the outcome of the case, e.g, which party wins. It uses an iterative selection process to pre-select particularly predictive sentences and creates the final summary based on this subset using Maximum Marginal Relevance. As guiding principles do not, as such, give an indication to the outcome of a case, such a selection would not be helpful.

A simpler approach is implemented by *CaseSummarizer* [7] which selects sentences mostly based on tf*idf, the information on whether a sentence is at the beginning of a paragraph and the occurrences of dates, and known entities in the sentences. Using neural networks for summarization tasks allows extractive summaries, as shown, e.g., by the Chinese *GIST* [8] using different ensemble models, as well as abstractive results by, e.g.,

fine-tuning pre-trained language models like *BERT2BERT* or *BART* [3]. Unfortunately, such approaches are in need of a high amount of data which is hardly available in the German legal field.

Compared to existing work, our approach allows us to work with the limited judgment data available in Germany and operates on a completely different level of structuring of judicial decisions than previous work. Thus we contribute to the current research by showing preliminary results on text summarization of judgments passed by the BGH. To our knowledge, little work on summarization tasks has been done on German judgments so far. As previous work may only be partially applicable, we show that the integration of the argumentative structure achieves significantly higher ROUGE-scores than our baseline on judgments of the BGH.

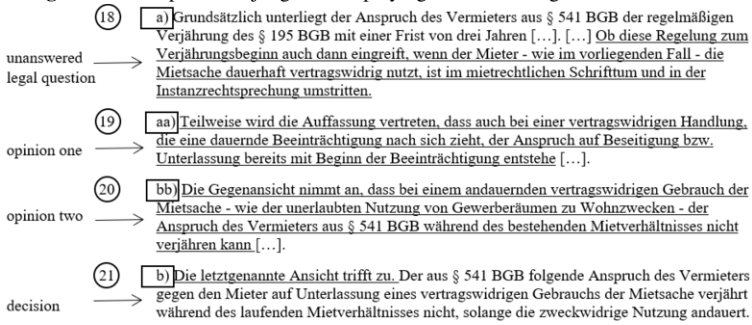
2. Working with the Argumentative Structure

To present our results we will first give an introduction to how Section II of the legal grounds of a judgment (*Entscheidungsgründe*) of the BGH is structured. Then, we will explain the data that we worked on, present our approach, and, finally, evaluate our preliminary results.

2.1. Argumentative Structure

Legal documents and especially judgments are highly structured texts. The reasoning concerning the legal grounds of a decision basically follows a treelike structure: The relevant points of law are each addressed and then discussed in more detail, one point after the other. Deeper levels contain further elaboration on the respective legal aspect, discuss sub-questions or give differing opinions of lower courts and literature. An example can be found in Figure 1.[2]

Figure 1. Excerpt from a judgment displaying the treelike argumentative structure



The example already shows a variety of different structuring elements: The numbers in circles at the left hand side depict the consecutive numbering of paragraphs (*Randnummern*) of the judgment. The corresponding paragraphs contain logical units of the text. The listing in the rectangles at the beginning of the paragraphs indicate the argumentative

[2]The judgment can be found at juris.bundesgerichtshof.de/cgi-bin/rechtsprechung/document.py?Gericht=bgh&nr=91902 and an English image version at legalinf.de/jurix22.

structuring level; elements of the listing might contain several paragraphs. Elements of the same structuring level are of equal abstraction level and oftentimes continue the line of reasoning. In case of our example above, an—from the BGH's perspective—open legal question is stated at the end of paragraph 18. The following paragraphs 19 and 20 (of the lower structuring level) present different opinions on this question as derived from lower court decisions. In paragraph 21 we return to the higher structuring level again and the BGH decrees which of these opinions is, according to them, correct. This example already shows how the reasoning and the argumentative structure can give an indication for finding the guiding principles: The last sentence in paragraph 18 gives the explicit concluding decision of the court and is the guiding principle of this particular judgment.

2.2. Data

We investigate whether the argumentative structure of a judgment allows us to select the sentences for the guiding principles. Therefore, we inspected existing judgments of the BGH and gathered 100 judgments with extractive guiding principles already formulated by the court.[3] For all of these 100 judgments we determined the exact positions of the respective guiding principles with respect to the argumentative structuring. To avoid overfitting, we used another set of 100 judgments as an unknown test set for the validation of our results. The judgments in this test set did not contain extractive guiding principles but abstractive ones. The reason we chose such a dataset as a test set is that most existing judgments contain abstractive guiding principles. Note that we measure our resulting summaries with the ROUGE metric. As the judgments used for our analysis contain the exact sentences of the guiding principles, a perfect summarization algorithm may reach a ROUGE-score of 1. This is impossible as regards the remaining test set as it does not contain sentences that are syntactically identical to the guiding principles. Therefore, it is only natural that the ROUGE-scores of the results on these judgments are lower than on the other 100 cases.

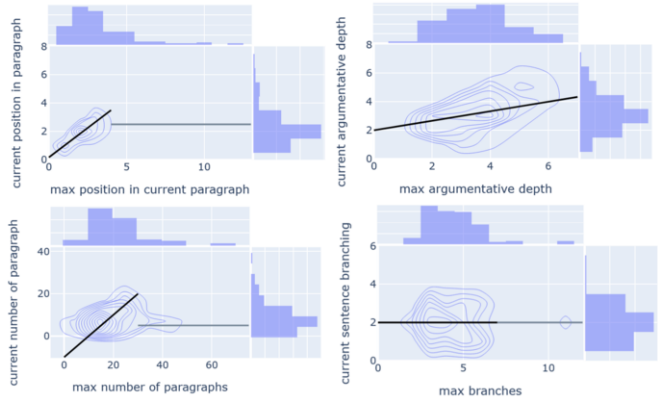
2.3. Ranking Sentences based on Aspects of the Argumentative Structure

In regards to the argumentative structuring as a treelike structure, we focused on the position of a sentence in the paragraph, the argumentative depth of the sentence in the tree, the number of the paragraph of the sentence (counted from the beginning of Section II of the legal grounds), and the number of the branch with respect to one parent node. To get further insight into the impact of these features on the guiding principles, we considered the extractive guiding principles and their corresponding values of these features in relation to their maximum values in these judgments. An illustration can be found in Figure 2 (density plots) which shows, e.g., that the guiding principles in our judgments were mostly found in a branch number close to 2 (density is highest, lower right image). The histograms give an indication of the data distribution of the features (e.g. branch number, at the right) and the maximum of that feature in the judgment (e.g. maximum branch number in the judgment, at the top).

Based on these insights we derived linear functions to approximate the optimal values for a sentence to be chosen as a guiding principle (lines in the images). Except for the feature *depth in the argumentative structure*, we had a long tail in the distributions of the

[3]The cases were accessed at www.rechtsprechung-im-internet.de.

Figure 2. Parameter distributions and approximation functions for the 100 extractive guiding principles



maximum values. Therefore, we decided to use different partially defined approximation functions (gray lines) for the tails as the knowledge derived from this data is much less reliable. In case of the branch numbers, this resulted in the same function for both parts.

For our actual ranking of sentences, we first pre-processed the judgments by removing stopwords and sentences without at least one verb in present tense (as guiding principles are always written in present tense), as well as lemmatizing and case normalizing the words. Similar to *CaseSummarizer* [7], we then calculated a ranking for each sentence as the sum of tf*idf values of its words and normalizing the score by the length of the sentence ($rank_{tfidf}$). Then, we calculated the final ranking by the following formula: $rank_{final} = rank_{tfidf} + \sigma * (p_d * f_d(d, max_d) + p_p * f_p(p, max_p) + p_n * f_n(n, max_n) + p_b * f_b(b, max_b))$ with σ the standard deviation of $rank_{tfidf}$ and p_d, p_p, p_n and p_b tunable parameters for the argumentative depth (d), position in paragraph (p), number of paragraph (n) and the number of branching (b). The functions f_d, f_p, f_n and f_b calculate the shortest distance of the current value (e.g., of the argumentative depth of a sentence) to the derived functions of the most likely values. The final selection of the ranked sentences is done by first selecting the sentence with the highest score and then adding as many sentences from the top of the ranking until the selection has a length of approx. 2.47% of the original judgment, which was the average length of guiding principles compared to the original judgment in a set of 5000 judgments.

2.4. Evaluation

For our evaluation, we compared our results to a random selection of sentences, a ranking using only tf*idf values and *CasesSummarizer* (Table 1). We distinguished between the results concerning the judgments containing abstractive and extractive guiding principles and optimized the parameters of our approach separately for each of these datasets.

As expected, the ROUGE values on the abstractive judgments were significantly lower than in the extractive judgments and the optimized formula for the extractive version is by no means optimal for the abstractive judgments and vice versa. Compared to a random ranking and a simple tf*df based ranking we could significantly increase the ROUGE score of the results. Especially in comparison to *CaseSummarizer* we constantly achieved high ROUGE-L scores, which indicates that we are (more) successfully able to identify sentences close to the meaning of the original guiding principles.

Table 1. Evaluation results using ROUGE metric

	extractive judgments		abstractive judgments	
	ROUGE-1	ROUGE-L	ROUGE-1	ROUGE-L
random ranking	0.2259	0.1950	0.1403	0.1751
tf*idf ranking	0.2874	0.3280	0.1208	0.1775
CaseSummarizer	0.3886	0.2520	0.2310	0.1890
our work (optimized on extractive dataset)	0.4134	0.4141	0.1696	0.2163
our work (optimized on abstractive dataset)	0.3596	0.3643	0.1881	0.2232

3. Conclusion and Future Work

We found that our approach of using the argumentative structure of the judgments seems to be indeed a fruitful starting point for creating guiding principles in the case of the BGH. Integrating the argumentative structure in a ranking for extractive summarization achieves higher results than our baseline and performs especially well in the ROUGE-L metric.

In our ongoing work we intend to further compare our approach to other existing algorithms and make use of semantics to determine whether they relate to the argumentative structure. Furthermore, we plan to extend our analysis to abstractive guiding principles and their relation to the argumentative structure. Other aspects, like highly recurrent terms, might also increase the ROUGE-score of selections.

References

[1] Deroy A, Bhattacharya P, Ghosh K, Ghosh S. An Analytical Study of Algorithmic and Expert Summaries of Legal Cases. In: Legal Knowledge and Information Systems; 2021. p. 90-9.

[2] Bhattacharya P, Hiware K, Rajgaria S, Pochhi N, Ghosh K, Ghosh S. A Comparative Study of Summarization Algorithms Applied to Legal Case Judgments. In: Advances in Information Retrieval. Cham: Springer International Publishing; 2019. p. 413-28.

[3] Yoon J, Junaid M, Ali S, Lee J. Abstractive Summarization of Korean Legal Cases using Pre-trained Language Models. In: 2022 16th International Conference on Ubiquitous Information Management and Communication (IMCOM); 2022. p. 1-7.

[4] Farzindar A, Lapalme G. LetSum, an automatic Legal Text Summarizing system. In: Legal Knowledge and Information Systems, Jurix 2004: The Seventeenth Annual Conference; 2004. p. 11-8.

[5] Bhattacharya P, Poddar S, Rudra K, Ghosh K, Ghosh S. Incorporating Domain Knowledge for Extractive Summarization of Legal Case Documents. In: Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law. ICAIL '21. New York, NY, USA: ACM; 2021. p. 22–31. Available from: https://doi.org/10.1145/3462757.3466092.

[6] Zhong L, Zhong Z, Zhao Z, Wang S, Ashley KD, Grabmair M. 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. New York, NY, USA: ACM; 2019. p. 163–172. Available from: https://doi.org/10.1145/3322640.3326728.

[7] Polsley S, Jhunjhunwala P, Huang R. CaseSummarizer: A System for Automated Summarization of Legal Texts. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations. Osaka, Japan: The COLING 2016 Organizing Committee; 2016. p. 258-62. Available from: https://aclanthology.org/C16-2054.

[8] Liu CL, Chen KC. Extracting the Gist of Chinese Judgments of the Supreme Court. In: Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law. ICAIL '19. New York, NY, USA: ACM; 2019. p. 73–82. Available from: https://doi.org/10.1145/3322640.3326715.