EurLexSummarization - a new text summarization dataset on EU legislation in 24 languages with GPT evaluation

Valentin Zmiycharov
FMI, Sofia University
"St. Kliment Ohridski"
Sofia, Bulgaria
valentin.zmiycharov@gmail.com

Todor Tsonkov
FMI, Sofia University
"St. Kliment Ohridski"
Sofia, Bulgaria
ttsonkov@gmail.com

Ivan Koychev
FMI, Sofia University
"St. Kliment Ohridski"
Sofia, Bulgaria
koychev@fmi.uni-sofia.bg

Abstract

Legal documents are notorious for their length and complexity, making it challenging to extract crucial information efficiently. In this paper, we introduce a new dataset for legal text summarization, covering 24 languages. We not only present and analyze the dataset but also conduct experiments using various extractive techniques. We provide a comparison between these techniques and summaries generated by the state-of-the-art GPT models. The abstractive GPT approach outperforms the extractive TextRank approach in 8 languages, but produces slightly lower results in the remaining 16 languages. This research aims to advance the field of legal document summarization by addressing the need for accessible and comprehensive information retrieval from lengthy legal texts.

Keywords: Legal texts summarization, Long texts summarization, New dataset.

1 Introduction

The task of automatically summarizing legislative documents poses a formidable challenge, primarily due to the extensive nature of these texts, making them intricate to comprehend and process. This complexity is further compounded when dealing with 24 languages, as there is a relative scarcity of pre-trained models for summarization in comparison to more widely-used languages such as English.

While existing text summarization models have shown promise, they are generally trained on shorter texts, such as social media posts or news articles, which significantly differ in complexity and length from legislative documents. In the forthcoming section, there is an overview of text summarization approaches tailored for longer documents.

This paper presents a new dataset comprising EU legislative documents available in 24 languages. This dataset has undergone cleaning and preprocessing to ensure its utility and accessibility for research and development in the field of legal document summarization.

The contributions of this paper can be summarized as follows:

- Introduction of a multilingual legislative document summarization dataset, cleaned and preprocessed for immediate usage.

- Evaluation of the quality of summaries generated using the GPT model, coupled with a comprehensive comparative analysis against three distinct extractive summarization methods. This research aims to shed light on the efficacy of these techniques in the context of legislative documents and ultimately advance the state of the art in this critical domain.

2 Related work

Text summarization is a fundamental task in natural language processing, with a wide range of applications, from news article summarization to document summarization. In this section, we discuss related work in two key areas: text summarization datasets and the summarization of long and legislative documents.

2.1 Datasets

One critical aspect of text summarization research is the availability of diverse datasets for training and evaluation. The Document Understanding Conference (DUC) and Text Analysis Conference (TAC) have played a pivotal role in advancing text summarization research by providing benchmark datasets. Notable examples include DUC 2003 (National Institute of Standards and Technology, 2003) and TAC 2008 (National Institute of Standards and Technology, 2008), which have spurred innovation in extractive and abstractive summarization tasks.

Moreover, the CNN/Daily Mail dataset introduced by Hermann et al. has been influential in abstractive summarization. This dataset comprises news articles and corresponding human-generated summaries, serving as a valuable resource for training and evaluating abstractive summarization models (Hermann et al., 2015).

In addition to general text summarization datasets, there is a growing interest in domain-specific datasets, particularly in the field of legal text summarization. Legal documents, characterized by their complexity and extensive use of legal terminology, present unique challenges. Recent efforts, such as the creation of the ”Multi-LexSum” dataset (Shen et al., 2022), focus on facilitating summarization specifically for legal texts, thus advancing the state-of-the-art in this domain.

2.2 Summarization of long and legal texts

Summarization of long documents, such as legislative texts, requires specialized techniques. Transformer-based models like BERTSUM (Liu and Lapata, 2019) and PEGASUS (Zhang et al., 2020) have demonstrated state-of-the-art performance in handling lengthy documents. These models leverage the ability to capture context over larger text spans, making them particularly well-suited for summarizing extensive legislative documents.

Efforts to improve legal text summarization also extend to the development of domain-specific pre-trained models. Models like Legal-BERT (Chalkidis et al., 2020), fine-tuned on legal corpora, show promise in accurately summarizing legal documents, offering valuable resources for legal professionals and researchers.

The challenge of summarizing complex legal case judgments is addressed by conducting the first systematic comparison of various summarization algorithms (Bhattacharya et al., 2019). Focusing on Indian Supreme Court judgments, the study evaluates both general and legally specialized algorithms, providing assessments against gold standard summaries. The research not only contributes to the advancement of summarization techniques for legal documents but also offers insights from a legal expert’s perspective.

Introducing a novel approach to abstractive summarization of lengthy legal opinions, another method prioritizes the document’s argument structure (Elaraby et al., 2023). By incorporating argument role information, it generates multiple candidate summaries and reranks them based on alignment with the document’s argument structure. Demonstrating superior performance over robust baselines, this approach proves effective in summarizing complex and nuanced legal opinions.

Finally, the challenge of producing abstractive summaries for long texts is an ongoing focus of research. Techniques involving reinforcement learning and advanced decoding mechanisms have enhanced the quality and coherency of abstractive summaries, addressing the unique challenges presented by lengthy documents (Paulus et al., 2017).

In summary, text summarization has made significant strides, especially in the context of legislative documents. Dedicated datasets and advanced models have paved the way for more effective summarization of lengthy and complex texts. Domain-specific challenges in legal text summarization remain a prominent research agenda, with the potential to benefit legal professionals and society at large.

3 Dataset collection and preprocessing

The dataset contains legislative documents and their summaries in 24 languages from the European Union[1]. In addition to producing a new dataset to serve other researchers, these were used for various experiments.

The data is downloaded from the official website of the European Union[2]. It contains legal texts on various topics, including laws, acts and others. Each paper has a summary generated by an expert in the field. All documents and summaries, apart from English, have been translated into up to 23 other languages. Some summaries summarize more than 1 document. At the time of data collection, there are 1,816 summaries and their

[1] https://huggingface.co/datasets/FMISummarization/FMI_Summarization

corresponding full documents in English. A comparison between the data in different languages is made in the following sections.

Many characteristics of the dataset make it valuable to the research field. It contains subject-specific text that is challenging for overtrained generic text models. The texts vary in length, with some documents being extremely long. Having them in different languages provides an opportunity to compare the results and validate whether the experiments work only in a particular case. The small number of documents further complicates the task.

3.1 Download and preprocess the data

For each full document and summary, the full HTML content is downloaded. All documents are crawled by taking the search results page by page and for each result the unique identifiers of the full documents and summaries are stored. In cases where there is no translation of the searched language for the relevant pair of (summary, full document), the information is not saved.

The header (Figure 1) and the list of references from the bottom part (Figure 2) are removed. There are cases where several documents correspond to one link. With them, the specified actions are repeated for each document and the content is concatenated.

Summaries are divided into sections. Figure 3 shows the number of documents with a certain number of sections. The most are documents with 6, 7 and 8 sections. After a detailed analysis of the sections, it turns out that a large number of them do not carry essential information for the summary. 408 different section titles have been identified. Table 1 shows the most popular 10 section names and the number of documents in which they occur for English.

As a summary we consider the text contained in one of the following sections: KEY POINTS, SUMMARY. Some summaries do not contain any of these sections. After manual review, it turned out that they were indeed invalid (example Figure 4). All essential information is contained in these sections, while the rest contain ancillary information that is not key to the summary or is already described in the main section.

Other documents that were removed from the dataset were those for which there was no translation in the respective language. Additionally, there are documents where the summary text is longer than the full document. Following a manual review, additional invalid documents were identified and subsequently removed. A total of 5118 out of 45983 documents (11.1%) were removed.

3.2 Data insights for English texts

The final result is 1816 summaries with their corresponding documents. The average length of full documents is 20716 words and of summaries 509 words. The word count ratio between full papers and summaries averaged 44.4. It is important to mention that the length of the texts is unevenly distributed with the presence of many deviations. For example, there are summaries with over 5,000 words (10 times the average) and full texts with over 800,000 words (40 times the average) (Figure 5).

To better visualize and gain a better idea of the distribution of the data, outliers were removed. For charting purposes only, the following have been removed:

- Complete documents with more than 100,000 words.
- Summaries of more than 2000 words.

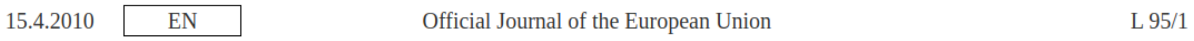
As can be seen in Figure 6 the ratio of word counts between full papers and summaries is not fixed and varies greatly between examples. This makes the task of automatic summarization even more challenging.

3.3 Comparing data on different languages

Using the same methodology, data was downloaded in all 24 available languages. The number of documents in different languages varies, with English expectedly the most (1816) and Irish the fewest (511). The word count ratio between full documents and summaries is fairly constant across all languages with slight variations due to the specifics of the language and the documents being translated.

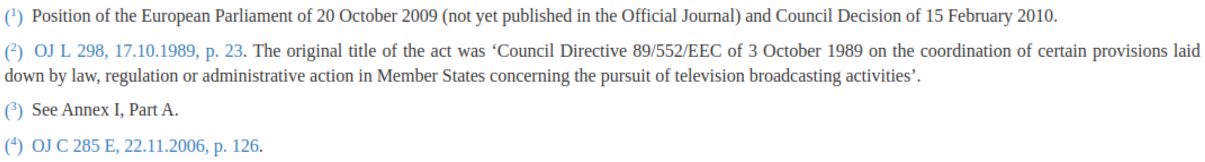
For each language, the key sections were identified, originating from the names in the language and analyzing the translations in the respective languages and the number of documents with the sections.

[2] https://eur-lex.europa.eu/homepage.html



15.4.2010 EN Official Journal of the European Union L 95/1

Figure 1: Header of the document.



(1) Position of the European Parliament of 20 October 2009 (not yet published in the Official Journal) and Council Decision of 15 February 2010.

(2) OJ L 298, 17.10.1989, p. 23. The original title of the act was 'Council Directive 89/552/EEC of 3 October 1989 on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the pursuit of television broadcasting activities'.

(3) See Annex I, Part A.

(4) OJ C 285 E, 22.11.2006, p. 126.

Figure 2: References of the document.

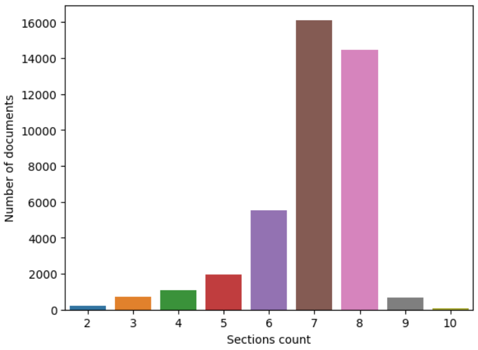


Figure 3: Number of sections in the summaries in the dataset. Most summaries have 6,7 or 8 sections. Minimum number of sections is 2 and maximum is 10.

4 Experimented approaches

4.1 Baseline

We assumed that the most important information is at the beginning of the document/section. Therefore, our baseline approach collects the first k consecutive sentences from the full texts. The number k of sentences is decided based on the number of words in the original summary. We continue adding sentences until we reach the number of words. Abstractive approaches like Pegasus (Zhang et al., 2020) also use the first part of the text. The results from the experiment reported below provide evidence that this assumption is good enough for a baseline.

4.2 TF-IDF Summarizer

We have also tested basic extractive summarization (Malik, 2019) based entirely on TF-IDF. The first step of the algorithm is to split the full text into a list of sentences. After that all special characters and stop words are removed. Stop words for various languages were sourced from the Stopwords ISO collection[3]. Then all sentences are tokenized. Next, the weighted frequency of occurrences of all words must be calculated. The weighted frequency of each word can be found by dividing its frequency by the frequency of the most occurring word. After that, the words in the original sentences are replaced by their respective weighted frequency. The weighted frequency for the words removed during preprocessing is zero. For each sentence, the sum of weighted frequencies is calculated. Only sentences with more than three words are evaluated to avoid the ones that do not contain enough information. Finally, the sentences are sorted in descending order by the sum of the weighted frequencies. The summary contains the sentences at the beginning of the ordered list. The number of sentences to be selected is based on the ratio between the number of sentences in the training dataset. The algorithm does not require training and is entirely based on the content of the full document.

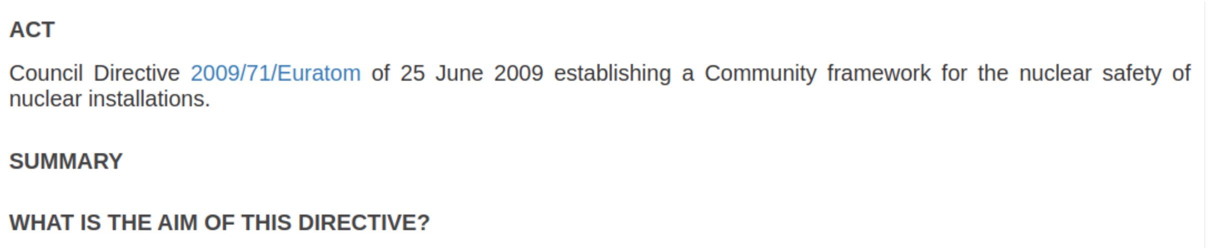
4.3 TextRank

For each sentence, we produce a vector of embeddings using Word2Vec (Mikolov et al., 2013). Word2Vec is an algorithm for generating a fixed-length distributed vector representation of all words in a huge corpus. The efficiency of Word2Vec is due to two reasons — one is the use of fixed-size vectors, which means that the size of the vector does not depend on the number of unique words in the corpus. Second, incorporating semantic information into vector representations. Word2Vec vectors are very effective at grouping similar words together. The algorithm can make strong judgments based on the position of the word in the corpus. For example, ”handsome” and ”nice” are similar, and therefore their vector representation will be very similar. The resulting vectors allow us to represent each sentence as a set of vectors for each word in them. To obtain vectors of the same size for

[3] https://github.com/stopwords-iso/stopwords-iso

Section name	Number of documents
KEY POINTS	1726
SUMMARY OF:	1608
BACKGROUND:	1601
RELATED DOCUMENTS	1172
MAIN DOCUMENT	1087
KEY TERMS	704
FROM WHEN DOES THE REGULATION APPLY?	509
WHAT IS THE AIM OF THE REGULATION?	465
MAIN DOCUMENTS	377
SUMMARY	341

Table 1: Section names and number of documents in which they appear (10 most popular).



ACT

Council Directive 2009/71/Euratom of 25 June 2009 establishing a Community framework for the nuclear safety of nuclear installations.

SUMMARY

WHAT IS THE AIM OF THIS DIRECTIVE?

Figure 4: Summary with missing text.

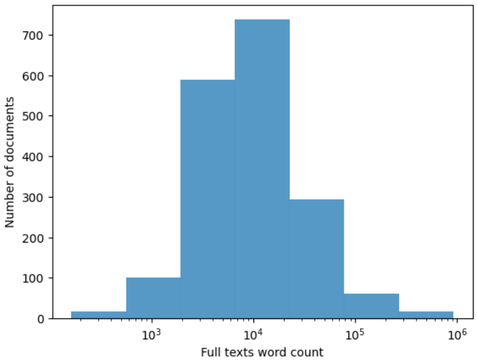


Figure 5: Number of words in full documents plotted on a logarithmic scale (x-axis). Examples with above 100000 words skew the average word count statistics.

all sentences, we zero-padded all but the longest sentences.

Once we have vectors that represent each sentence, we prepare a similarity matrix, which is square with the number of columns and rows equal to the number of sentences in the text. On it we apply the PageRank algorithm, which gives us a score for each sentence.

Once we have the scores, similar to the previous algorithm, we sort the sentences by the results obtained in descending order and concatenate the sentences in order until we reach the length of the original summary.

4.4 GPT summarization

We used the latest advancements in the machine learning field to generate summaries. We used OpenAI's GPT API[4].

We experimented with different prompts. The system prompt we ended up with is "*You are a lawyer.*" The aim is to order the agent to speak as if it is a lawyer. This matches the profile of the people who created the summaries. The document prompt is "*You will correctly answer the questions about the following text:*" concatenated with the full text. The prompt for the final task for the GPT API is "*Summarize the text without introductory words.*". The reason we added the ending "*without introductory words*" is the following: The GPT API has maximum token limits, which are not enough for most of the texts. Therefore, we had to break the full text into chunks. Without this addition to the prompt, all the generated texts started with introductory words like "The text outlines" or "The most important parts of the text are". These are repeated for each chunk and such introductions are not present in the real summaries. Prompts are translated into the 24 different languages using an automatic translator.

The price of GPT API usage is based on the

[4] https://platform.openai.com/docs/guides/gpt

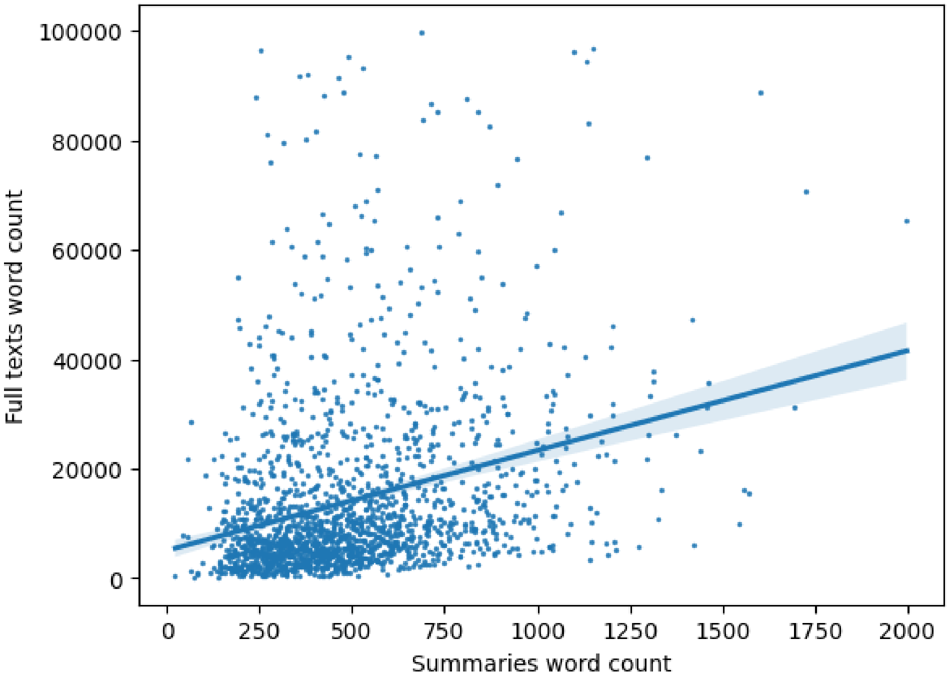


Figure 6: Each point represents the word count data of a summary and the corresponding word count of the full paper after removing outliers. The line and the space around it shows a linear regression model that best approximates the word count ratio between summaries and full papers.

number of tokens used. At the time of writing the paper the price for the model is $0.0015/1K tokens for Input and $0.002/1K tokens for output. Calculating all summaries for all languages would have cost $600 with this model and $14000 for the gpt-4 model. For price considerations we used the gpt-3.5-turbo model and generated summaries for 10% of the texts for all languages which cost us around $60 with additional costs of around $40 for various experiments.

The next challenge we faced was the ratio between the full text and the summary. As mentioned before we are aiming at generating summaries with the same length as the original summaries. On the other hand for GPT output we cannot precise the exact output length, only the maximum output length. When we call GPT API we have a maximum context of 4K tokens, which includes the input and the output. When we subtract the number of tokens from the prompt we are left with less than 4K for input text and output combined. What we tried first is to determine the ratio between the original full text and the summary. Let's say it is 9/1. We would then split the full text into chunks that are about 90% of the tokens allowed and leave 10% for the output. The problem we faced with this strategy is that in many cases the maximum tokens allowed for the output is so small that the model cannot finish what it is trying to answer.

After experimenting with different maximum token counts we decided to fix the desired ratio between summary and full text to be 1/5 - roughly 600 tokens for summary, 3000 for full text. We multiply the length of the summary by 5. We additionally multiply this value by 1.5 because we cannot specify the exact number of tokens, but only max tokens, which means that we are aiming our input of full texts to be 7.5 times longer than the summaries we are trying to generate. If the original full text is longer than that we added a preprocessing step to reduce its content. We calculate the text rank scores for all sentences in the full text and remove the sentences with the lowest score until this ratio is achieved. This way our approach is combining extractive and abstractive methodologies for longer texts to achieve its goals. If the ratio is lower than 7.5 no preprocessing is done.

After the preprocessing step is executed, the text is split into chunks. Each chunk is a list of full sentences that do not reach the maximum allowed limit for tokens. Each chunk is summarized and results are concatenated

5 Experiments and results

In this section, our experiments are designed to comprehensively assess and compare the listed text summarization approaches across all 24 languages using the legislation dataset.

5.1 Experiments Design

The most widely used metric for the evaluation of text summarization is rouge (Recall-Oriented Understudy for Gisting Evaluation). Rouge is a set of metrics used for evaluating automatic summarization and machine translation software. The metrics compare an automatically produced summary to a human-produced summary. Rouge-N refers to the overlap of n-gram between the system and reference summaries. Rouge-L refers to Longest Common Subsequence (LCS) based statistics. The longest common subsequence problem considers sentence-level structure similarity naturally and identifies the longest co-occurring in sequence n-grams automatically. In particular rouge-1, rouge-2, and rouge-L F1 scores were used in the conducted experiments.

We started generating summaries that have the same length as the original summary. This way the precision and recall are the same. This way we avoid the problem of changing the F-score due to generating larger or smaller summaries and focus entirely on the relevance of the sentences and not on the length of the summary. For all experimented approaches we selected only sentences with at least 3 words.

5.2 Analysis of Results

Table 2 shows the results of the different approaches. The TF-IDF approach improves the baseline for all languages. TextRank and GPT outperform TF-IDF for all languages. For 8 of the languages, GPT outperforms TextRank, while for the others TextRank is the best.

Lang	Baseline	TF-IDF	TextRank	GPT
BG	0.284	0.297	0.318	**0.323**
CS	0.248	0.269	0.285	**0.287**
DA	0.338	0.350	0.379	**0.389**
DE	0.326	0.338	**0.368**	0.364
EL	0.279	0.286	0.309	**0.316**
EN	0.364	0.384	**0.416**	0.405
ES	0.381	0.389	**0.414**	0.408
ET	0.194	0.209	**0.232**	0.216
FI	0.252	0.252	**0.286**	0.269
FR	0.375	0.385	**0.416**	0.396
GA	0.335	0.328	**0.348**	0.324
HR	0.235	0.243	**0.268**	0.268
HU	0.290	0.290	**0.320**	0.318
IT	0.378	0.375	**0.412**	0.399
LT	0.235	0.242	**0.262**	0.252
LV	0.238	0.238	**0.265**	0.261
MT	0.229	0.232	**0.265**	0.242
NL	0.302	0.308	0.336	**0.339**
PL	0.254	0.260	**0.279**	0.271
PT	0.383	0.391	**0.423**	0.388
RO	0.359	0.377	**0.402**	0.400
SK	0.237	0.251	0.269	**0.270**
SL	0.249	0.253	0.284	**0.288**
SV	0.330	0.327	0.370	**0.374**

Table 2: Rouge 1 F1 scores for all 24 languages for all experiment types.

6 Conclusion

The paper introduces a new multilingual dataset of European legislative laws, encompassing 24 languages. This dataset, characterized by significant variations in document length and a limited number of documents per language, presents a valuable resource for the field of automatic summarization.

We conducted extensive experiments, employing three extractive summarization approaches, and introduced a novel two-step methodology that harnesses the capabilities of GPT models for summarization. Notably, our two-step approach outperforms existing methods in some languages, demonstrating its potential as an effective summarization technique. However, in certain cases, TextRank surpasses it in performance.

In addition to our summarization findings, we offer valuable insights into the practical considerations of using GPT, including detailed information on associated costs. This information is essential for researchers and practitioners looking to leverage state-of-the-art models in real-world applications.

By addressing the complex task of summarizing European legislative laws in diverse languages, our work contributes to the advancement of the field, offering a valuable resource and novel techniques for future research and applications in automatic summarization and multilingual natural language processing.

7 Acknowledgements

This research is partially supported by Project UNITe BG05M2OP001-1.001-0004 funded by the OP "Science and Education for Smart Growth" and also co-funded by the EU through the ESI Funds.

References

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 *Advances in Information Retrieval*, pages 413–428, Cham. Springer International Publishing.

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.

Mohamed Elaraby, Yang Zhong, and Diane Litman. 2023. Towards argument-aware abstractive summarization of long legal opinions with summary reranking. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7601–7612, Toronto, Canada. Association for Computational Linguistics.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders.

Usman Malik. 2019. Text summarization with nltk in python. https://stackabuse.com/text-summarization-with-nltk-in-python/.

Tomas Mikolov, Kai Chen, G.s Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Proceedings of Workshop at ICLR*, 2013.

National Institute of Standards and Technology. 2003. DUC 2003 Dataset. https://duc.nist.gov/duc2003/tasks.html.

National Institute of Standards and Technology. 2008. TAC 2008 Dataset. https://tac.nist.gov/2008/summarization/update.summ.08.guidelines.html.

Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization.

Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. 2022. Multilexsum: Real-world summaries of civil rights lawsuits at multiple granularities.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. Pegasus: pre-training with extracted gap-sentences for abstractive summarization.