





RulingBR: A Summarization Dataset for Legal Texts

Diego de Vargas Feijó^(✉)  and Viviane Pereira Moreira 

Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil
{dvfeijo,viviane}@inf.ufrgs.br
<http://www.inf.ufrgs.br/>

Abstract. Text summarization consists in generating a shorter version of an input document, which captures its main ideas. Despite the recent developments in this area, most of the existing techniques have been tested mostly in English and Chinese, due in part to the low availability of datasets in other languages. In addition, experiments have been run mostly on collections of news articles, which could lead to some bias in the research. In this paper, we address both these limitations by creating a dataset for the summarization of legal texts in Portuguese. The dataset, called RulingBR, contains about 10K rulings from the Brazilian Federal Supreme Court. We describe how the dataset was assembled and we also report on the results of standard summarization methods which may serve as a baseline for future works.

Keywords: Summarization · Dataset · Legal · Law

1 Introduction

Text summarization is an important task in Natural Language Processing. It consists in generating a shorter version of the text given as input, capturing its main ideas. In the last few years, summarization has undergone significant developments. Notably, many of the new techniques being applied rely on deep learning strategies to go beyond the previously established state-of-the-art results [19, 22, 25, 26]. Despite the recent boom in this area, the majority of works have been using English and Chinese datasets due in part to the low availability of resources in other languages.

Another limitation of the current research is that it focuses on news articles, for which the task consists in generating the headline or a very short summary. For example, models trained on the DUC-2004 task can only generate summaries of up to 75 characters [14, 19], and the input consists of only one or two sentences.

News articles usually begin with a *teaser* sentence used as a catch for the reader, which sums up the contents of the full article. So, the task of guessing the title can generally obtain good results by simply extracting the first few words of the article. The excessive focus on this type of text introduces bias in the techniques being developed. For example, Google's Textsum model [24] for

summarization uses just the first two paragraphs of the article. Another possible approach is to weight the sentences in descending order from the start, in favor of the first few sentences [23].

We believe there is a need for datasets with different text styles and longer summaries with contents taken from several parts of the input. This would allow a more realistic setting and potential for employing summarization in a wider set of applications.

In this paper, we report on the creation of RulingBR – a dataset for the summarization of legal texts in Portuguese containing over 10K decisions from the Brazilian Federal Supreme Court. Our contribution aims at addressing two limitations of the research in summarization: (i) the low availability of resources for languages other than English and Chinese, and (ii) the excessive focus on summarizing news articles. We have assembled a language resource in Portuguese to enable the development of methods for this language. The second contribution is to do with the style of the texts, which will contribute to a greater variety in the research on summarization.

2 Related Work

There are a few datasets that have been used for evaluating summarization techniques on generic domains. The following are available in Brazilian Portuguese.

TeMário [17] is composed of 100 news articles. Each text contains a pair of reference summaries: one was made by a human and the other was automatically generated.

Summ-it [6] is an annotated corpus that contains anaphoric coreferences. These are newspaper articles annotated from the Brazilian *Folha de São Paulo* newspaper.

CSTNews [1] is another annotated corpus. It is composed of 50 text collections and each collection has about four documents. It uses texts from the following Brazilian news sources *Folha de São Paulo*, *Estadão*, *O Globo*, *Jornal do Brasil*, and *Gazeta do Povo*.

The most widely used datasets are available in English and are described below.

The Annotated Gigaword [18] is the largest static corpus of English news documents available [15]. It contains over 10 million documents from seven news sources, annotated with syntactic and discourse structure. It was not specifically built to be a summarization dataset, but it has been used for this purpose by simulating that the headline would be a summary of the article.

CNN/Daily Mail was purposely designed for summarization as each article comes paired with a short set of summarized bullet points that represent the highlights of the text. It is frequently used for question answering [5] and is composed of about 300 thousand articles.

Opinosis [8] contains customer reviews about a product they bought. Each product description has five reviews. This is a small dataset containing only 51 articles.

DUC¹ stands for Document Understanding Conference. It has run a specific summarization track since 2001. In 2008, DUC became a summarization track inside the Text Analysis Conferences (TAC). These datasets contain human-produced per-document and multiple document summaries.

RulingBR differs from these related datasets because in the legal domain, documents are generally lengthier and their structure is very different from the structure of news articles. As a consequence, the assumption that the most important ideas will be in the first few sentences is not valid.

3 A Summarization Dataset with Legal Documents

For the purpose of text summarization in the legal domain, we searched for a source with a large number of publicly available documents. Thus, we chose to use the *Supremo Tribunal Federal* (STF) as our source. The STF is the highest court in Brazil and has the final word interpreting the country’s Federal Constitution. All of its decisions must be published online and are available in its internet portal².

3.1 Structure of the Documents

The full decision document, called (*inteiro teor*), is composed of four parts, namely: “Ementa”, “Acórdão”, “Relatório”, and “Voto”, which we now describe.

- The *Ementa* is a brief summary of the main topics discussed in each case and how the judges decided. We will be using the *Ementa* as the *reference summary* that automatic methods should aim to produce. In our corpus, the size of the *Ementa* was typically around 7% of the size of the full content.
- The *Acórdão* is a brief description of how each judge has decided and what the final decision was. This section represents around 2% of the full content.
- The *Relatório*, meaning report, is a compilation of the main arguments and events that happened during the trial. In general, this section accounts for about 22% of the full content.
- The last section, called *Voto*, may contain one vote, in case that the other judges agree with the first judge, or individual votes for each judge, otherwise. Because the votes need to address all the points raised by the petitioners, this tends to be the largest section covering around 69% of the full content.

The *Ementa* is useful for lawyers and other legal professionals when they are searching for decisions about a given topic. A good text should not be long, generally less than one page, making it a good summary of the full decision.

¹ <https://duc.nist.gov/>.

² <http://www.stf.jus.br/>.

```
{
  "ementa": "Embargos de declaração em recurso extraordinário com agravo. 2. Decisão monocrática. (...) 5. Agravo regimental a que se nega provimento.",
  "acordao": "Vistos, relatados e discutidos estes autos, acordam os ministros do Supremo Tribunal Federal, em Segunda Turma, (...), por unanimidade, converter os embargos de declaração em agravo regimental e, a este, negar provimento, nos termos do voto do Relator.",
  "relatorio": "(...) Trata-se de embargos de declaração opostos contra decisão que negou provimento a recurso, ao fundamento de que a natureza da matéria versada nos autos reveste-se de índole infraconstitucional. Aponta-se violação direta à Constituição Federal, em especial, aos artigos (...).",
  "voto": "(...) Tendo em vista o princípio da economia processual, recebo os embargos de declaração como agravo regimental e, desde logo, passo a apreciá-lo. (...)"}

```

Fig. 1. Example of a document already divided into sections in JSON format.

3.2 Data Collection

In order to obtain the documents, the Scrapy [21] library was used to browse the search pages and to download the documents. Only a few documents from the years 2010 and 2011 could be successfully parsed. Thus most documents are dated from 2012 to 2018.

The raw text we obtained contains some undesired pieces of texts such as headings, footers, page numbers, *etc.* We used regular expressions to identify the starting and ending points of each section of interest and remove unwanted text. Finally, the text of the sections was dumped as a JSON object, one object per line.

In Fig. 1, we show an extract from a short document in the final JSON format. The ellipsis indicates the omission of content to save space.

The final file has about 173 MB and contains 10,623 decisions and can be downloaded from <https://github.com/diego-feijo/rulingbr>. There are around 26 million tokens in the entire dataset.

We investigated whether there is a correlation between the length (in tokens) of the *Ementa* section and all other sections combined (full document). This correlation would be important for us to determine the desired summary size when using the automatic summarizers. The calculated correlation coefficient was 0.39, which is considered weak and is reflected by a large dispersion.

4 Evaluating Summarization Systems on RulingBR

In this Section, we present results of out-of-the-box extractive summarization strategies on RulingBR dataset. The goal is to provide baseline results for future summarization techniques.

4.1 Experimental Setup

In order to establish some baselines using this corpus, we have run a few automatic summarization experiments using two common libraries.

The first library used was Gensim [20]. This is a software framework for Natural Language Processing that implements some popular algorithms for topical inference and has a TextRank implementation for summarization. This library implements a variation of the TextRank [13] algorithm.

The second library used was the Sumy package [3]. It has a large variety of algorithms implemented using a common interface which makes it easier to run and compare the results.

The choice of summarization algorithms was motivated by the fact that they could be applied directly to the text without requiring additional information such as part-of-speech tags or headlines.

The algorithms used in this experiment were the following.

TextRank uses a graph-based ranking model for text processing. This algorithm applies unsupervised methods for keyword and sentence extraction and is based on ideas borrowed from HITS [10] and PageRank [16]. Both Gensim and Sumy implement variations of the TextRank algorithm. The Gensim implementation was improved [2] replacing the cosine by the Okapi-BM25 similarity function.

Luhn [12], which uses statistical information derived from word frequency and distribution to compute a relative measure of significance, first for words and then for sentences. The set of sentences with the highest scores are extracted to make up the summary.

LexRank [7] is a graph-based method to compute a relative measure of importance which is based on the concept of eigenvector centrality in a graph representation of the sentences. In this model, a connectivity matrix based on intra-sentence cosine similarity is used as the adjacency matrix for the graph representation of sentences.

4.2 Evaluation Metrics

The most commonly used metric for evaluating summarization algorithms is Rouge [11], which stands for Recall-Oriented Understudy for Gisting Evaluation. Its goal is to provide a measure of quality of an automatically generated summaries in comparison against a reference summary produced by humans.

The Rouge metric checks for overlapping text segments between the automatically generated summary and the reference summary. Different levels of granularity can be used. Rouge-1 counts the occurrences unigrams that appear in the automatically generated and the reference summaries. Rouge-2, counts how many bigrams were found (in the same order). Rouge-L stands for the longest common sub-sequence between the automatically generated and the reference summaries.

4.3 Experimental Procedure

Although both libraries used in the experiments support stopword removal, stemming, and tokenization, we opted to apply it beforehand as preprocessing steps

to make sure that the same operations were applied in all settings. For most of the stages, we have used the Natural Language Toolkit (NLTK) [4], which is a widely used library for processing of natural language documents. It contains functions and trained models in many languages. We used this library for filtering, stemming, and tokenization.

Stopword Removal – In order to try to make a fair analysis of the content produced in the summaries, stopwords should be removed since their presence could artificially inflate the quality metrics (since the reference summaries would certainly contain many such words). We have used the Portuguese stop-list provided with NLTK. Also, we have filtered any token with fewer than two characters. This was done because these tokens have low discrimination power, and, as we are generating a summary, we expect that the words should contain relevant semantic meaning.

Stemming – This technique conflates the variant forms of words into a single stem. We used the NLTK implementation of the RSLP-Stemmer [9].

Tokenization – This is the task of separating the text into chunks. It is used for dividing the text into sentences and then into words. Recognizing the start and end of sentences is crucial for the extractive summarization algorithms because they will compute the score of each sentence and output the highest scored sentences. The tokenizer must identify situations such as when sentences were not being finished by a period (*e.g.* Hurry up!) or when a period was being used for an abbreviation (*e.g.* Mr. John) rather than to indicate the end of a sentence. Again, we used the NLTK implementation of the Punkt tokenizer trained for the Portuguese language.

Standardization – The documents in the corpus significantly vary in length due to the several subjects that are covered by the decisions. In order to try to generalize a pattern, some outliers needed to be dropped. Using a token (word) as measuring unit, we calculated the mean and the standard deviation for the summaries (99.53 ± 91.17) and for the full contents (1397.44 ± 2101.73). In order to reduce the dispersion, we removed outliers. Input documents with fewer than 300 words or more than the mean plus 3 times the standard deviation were treated as outliers. In a similar fashion, summaries with fewer than 19 words or more than the mean plus 3 times the standard deviation were also removed. With this standardization, we removed 616 decisions, which represent 5.80% of the total. Full contents mean became 1200.65, with a standard deviation of 893.86; Summary mean became 91.79, with a standard deviation of 62.92. The frequency distribution after the cleaning can be seen in the histograms of Fig. 2.

4.4 Model Parameters

It is important to notice that the evaluation scores could be affected by the size of the generated summaries. That happens because a longer summary would probably have a greater recall and, as consequence, a higher Rouge score.

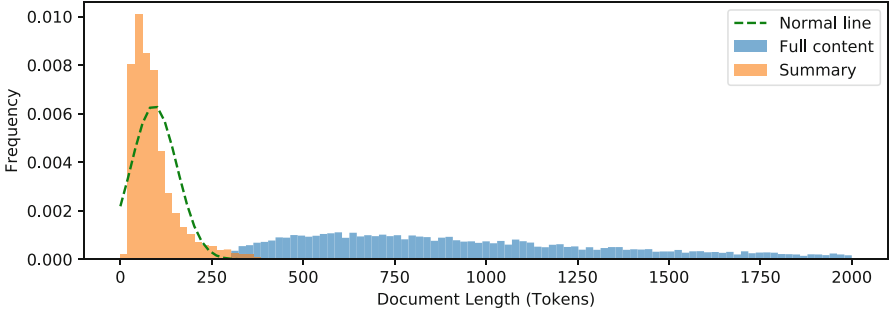


Fig. 2. Frequency distribution of the length of the summaries and the full content of the documents.

The libraries that generate automatic summaries receive as parameter the size of the desired output. As we discussed earlier, there is no strong correlation between the length of the document and the length of the summary. As shown in the histograms of Fig. 2, the size of the reference summaries can vary roughly between 30 and 150 tokens. So, setting the desired summary size to a fixed value will introduce an error as the size will be different from the size of reference summary. Nevertheless, we had to stick to a fixed size.

In these libraries, the output is entire sentences, so the total of words can be much smaller or larger than the desired output size. For example, the Gensim library receives the number of desired words, it computes the best sentences and will append them to the output until the difference between the desired output and the generated output is minimized. The Sumy library receives only the number of desired sentences, so the output may have a size completely different from the size of the reference summary (either much larger or much smaller).

In our dataset, sentence length can vary a lot. It is possible to find one-word sentences and sentences with a few hundred words. So, it is fairer to run our experiments with different size parameters. This way the results are not negatively impacted by an arbitrary choice of size.

4.5 Results

A higher Rouge score reflects a higher similarity between the automatically generated summary and the reference summary. Our goal when running this evaluation is to establish how standard extractive algorithms perform on this dataset. We have no intent in comparing those algorithms, as this would require evaluations under many different contexts and parameters.

As Table 1 shows, Rouge F-Score and Recall increase when the summary is longer. So, for a fair comparison, we used the scores of the runs in which the absolute differences between the length of the generated summary and its reference is minimized. Figure 3 shows the scores for Gensim using the desired output of 80 words, Luhn’s algorithm with a fixed output of one sentence, LexRank and TextRank algorithms with a fixed output of two sentences.

In our experiments, the Gensim library generally performed slightly better. But, the highest Rouge-1 F-Score was obtained using the LexRank asking for four sentences in the Sumy library.

Table 1. Results of the summarization using different lengths of outputs. Best results per metric are shown in bold.

Algorithm	Length	Abs Dif	Rouge-1			Rouge-2			Rouge-L		
			F	P	R	F	P	R	F	P	R
Gensim	60	467,138	0.27	0.30	0.11	0.11	0.14	0.27	0.14	0.20	0.16
Gensim	80	448,162	0.29	0.29	0.14	0.12	0.14	0.32	0.15	0.19	0.19
Gensim	100	497,309	0.30	0.28	0.16	0.13	0.13	0.37	0.16	0.17	0.22
Gensim	120	587,463	0.30	0.27	0.18	0.13	0.12	0.40	0.15	0.16	0.25
Luhn	1	475,039	0.23	0.29	0.09	0.10	0.14	0.23	0.13	0.21	0.14
LexRank	1	616,140	0.21	0.35	0.07	0.09	0.16	0.18	0.11	0.27	0.11
TextRank	1	544,613	0.22	0.33	0.08	0.10	0.17	0.20	0.12	0.26	0.13
Luhn	2	503,009	0.27	0.27	0.14	0.12	0.13	0.32	0.15	0.17	0.20
LexRank	2	498,650	0.27	0.32	0.11	0.11	0.14	0.28	0.15	0.21	0.17
TextRank	2	543,740	0.24	0.31	0.12	0.11	0.15	0.27	0.14	0.21	0.18
Luhn	3	729,632	0.29	0.26	0.17	0.13	0.12	0.40	0.15	0.15	0.25
LexRank	3	513,033	0.29	0.30	0.15	0.12	0.13	0.35	0.16^a	0.19	0.21
TextRank	3	673,067	0.26	0.29	0.15	0.12	0.14	0.33	0.14	0.18	0.22
Luhn	4	1,025,284	0.30	0.25	0.20	0.13	0.11	0.45	0.14	0.13	0.28
LexRank	4	609,894	0.31	0.29	0.17	0.13	0.12	0.40	0.16^a	0.17	0.25
TextRank	4	859,861	0.27	0.27	0.17	0.12	0.13	0.38	0.14	0.16	0.25

^a Both had exactly the same score.

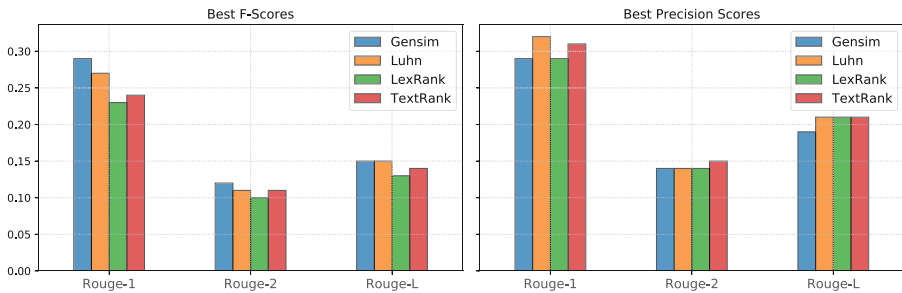


Fig. 3. F-Score and precision for the different summarization algorithms.

5 Conclusion

In this paper, we presented the RulingBR dataset, a corpus that can be used for natural language summarization. It differs from the existing corpora because it covers the legal domain and it is in Portuguese. We have analyzed different aspects of the dataset such as its organization, the size of each section, and how it can be used for the summarization task. We ran an experiment using different algorithms and libraries to establish baseline summarization results.

Despite the fact that the *Ementa* is a useful summary for legal professionals, it is not clear that the traditional general approaches for summarization could be directly applied to the legal domain producing texts that cover the same topics that a human would select.

The desired summary should contain the main topics discussed in the text. Perhaps, the desired output summary could be improved by appending these main topics, named entities, and compound terms. Also, we observed that the summary is composed of the final part of the *Acórdão*, the topics taken from the *Relatório*, and several ideas discussed in the *Voto*.

As a future work, we intend to test Neural models for summarization in order to identify the relevant aspects of the document and generate a summary in the style produced by a human.

References

1. Aleixo, P., Pardo, T.A.S.: CSTNews: um cópús de textos jornalísticos anotados segundo a teoria discursiva multidocumento CST (cross-document structure theory) (2008)
2. Barrios, F., López, F., Argerich, L., Wachenchauser, R.: Variations of the similarity function of TextRank for automated summarization. arXiv preprint [arXiv:1602.03606](https://arxiv.org/abs/1602.03606) (2016)
3. Belica, M.: Sumy: module for automatic summarization of text documents and HTML pages, April 2018. <https://github.com/miso-belica/sumy>
4. Bird, S.: NLTK: the natural language toolkit. In: Proceedings of the COLING/ACL on Interactive Presentation Sessions, COLING-ACL 2006, pp. 69–72. Association for Computational Linguistics, Stroudsburg (2006)
5. Chen, D., Bolton, J., Manning, C.D.: A thorough examination of the CNN/Daily mail reading comprehension task. CoRR abs/1606.02858 (2016). <http://arxiv.org/abs/1606.02858>
6. Collovini, S., Carbonel, T.I., Fuchs, J., Coelho, J.C., Rino, L., Vieira, R.: Summ-it: Um corpus anotado com informações discursivas visando à sumarização automática. In: V Workshop em Tecnologia da Informação e da Linguagem Humana, Congresso da SBC, pp. 1605–1614 (2007)
7. Erkan, G., Radev, D.R.: LexRank: graph-based lexical centrality as salience in text summarization. J. Artif. Intell. Res. **22**, 457–479 (2004)
8. Ganesan, K., Zhai, C., Han, J.: Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. In: Proceedings of the 23rd International Conference on Computational Linguistics, pp. 340–348. Association for Computational Linguistics (2010)

9. Huyck, C., Orenco, V.: A stemming algorithm for the Portuguese language. In: International Symposium on String Processing and Information Retrieval, SPIRE, p. 0186, November 2001
10. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. *J. ACM* **46**(5), 604–632 (1999)
11. Lin, C.Y.: Rouge: a package for automatic evaluation of summaries. *Text Summarization Branches Out* (2004)
12. Luhn, H.P.: The automatic creation of literature abstracts. *IBM J. Res. Dev.* **2**(2), 159–165 (1958)
13. Mihalcea, R., Tarau, P.: TextRank: bringing order into texts. In: Proceedings of EMNLP 2004 and the 2004 Conference on Empirical Methods in Natural Language Processing, July 2004
14. Nallapati, R., Xiang, B., Zhou, B.: Sequence-to-sequence RNNs for text summarization. *CoRR abs/1602.06023* (2016). <http://arxiv.org/abs/1602.06023>
15. Napoles, C., Gormley, M., Van Durme, B.: Annotated gigaword. In: Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-Scale Knowledge Extraction, AKBC-WEKEX 2012, pp. 95–100. Association for Computational Linguistics, Stroudsburg (2012)
16. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: bringing order to the web. In: Proceedings of the 7th International World Wide Web Conference, Brisbane, Australia, pp. 161–172 (1998). citeseer.nj.nec.com/page98pagerank.html
17. Pardo, T.A.S., Rino, L.H.M.: *Temário: Um corpus para sumarização automática de textos*. Universidade de São Carlos, Relatório Técnico, São Carlos (2003)
18. Parker, R., Graff, D., Kong, J., Chen, K., Maeda, K.: English gigaword fifth edition, linguistic data consortium. Google Scholar (2011)
19. Paulus, R., Xiong, C., Socher, R.: A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304* (2017)
20. Řehůřek, R., Sojka, P.: Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pp. 45–50. ELRA, Valletta, May 2010
21. ScrapingHub: Scrapy - a fast and powerful scraping and web crawling framework (2018). <https://scrapy.org>
22. See, A., Liu, P.J., Manning, C.D.: Get to the point: summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368* (2017)
23. Xiao: PyTeaser: Summarizes news articles, April 2018. <https://github.com/xiaoxu193/PyTeaser>
24. Xin Pan, P.L.: Models: models and examples built with TensorFlow, April 2018. <https://github.com/tensorflow/models>
25. Yin, W., Pei, Y.: Optimizing sentence modeling and selection for document summarization. In: Proceedings of the 24th International Conference on Artificial Intelligence, IJCAI 2015, pp. 1383–1389. AAAI Press (2015)
26. Zhang, X., Lapata, M.: Sentence simplification with deep reinforcement learning. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, 9–11 September 2017, pp. 584–594 (2017)