Andrej Miščič
Dragomelj 78c, 1230 Domžale, Slovenija
Study programme: Data Science, Computer and information science, MAG
Enrollment number: 63160228

Committee for Student Affairs
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko
Večna pot 113, 1000 Ljubljana

The master's thesis topic proposal
Candidate: Andrej Miščič

I, Andrej Miščič, a student of the 2nd cycle study programme at the Faculty of computer and information science, am submitting a thesis topic proposal to be considered by the Committee for Student Affairs with the following title:

Slovenian: **Avtomatsko povzemanje pravnih besedil**
English: **Automatic summarization of legal documents**

This topic was already approved last year: ***YES***

I declare that the mentors listed below have approved the submission of the thesis topic proposal described in the remainder of this document.

I would like to write the thesis in English because it is required by my study programme.

I propose the following mentor:

Name and surname, title: Assist. Prof. dr. Slavko Žitnik
Institution: Faculty of Computer and Information Science, University of Ljubljana
E-mail: slavko.zitnik@fri.uni-lj.si

Ljubljana, 4. december 2021.

Proposal of the masters thesis topic

1 The narrow field of the thesis topic

English: natural language processing

2 Key-words

English: automatic summarization, extractive, abstractive, legal documents

3 Detailed thesis proposal

Past approvements of the proposed thesis topic:
The proposed thesis has already been approved last year.

3.1 Introduction and problem formulation

In recent times we have been a witness to drastic increases in the amounts of available text data. This data belongs to various domains and sources, news, books, scientific papers, etc.; and presents an invaluable source of knowledge and information. However, due to the sheer bulk of this data it is many times infeasible to extract relevant information in reasonable time. Automatic summarization is an active subfield of natural language processing that attempts to deal with this problem. Its goal is to compress longer text documents into shorter summaries while preserving as much salient information as possible. Automatic summarization has been applied in many domains, such as newspaper articles [1] and scientific publications [2].

In this work we focus on applying automatic summarization techniques on legal text documents, i.e., legislation, judgements or legal articles. Legal documents present a challenge in modern natural language processing due to the usage of domain-specific vocabulary, long, complex sentences and the overall nested structure they commonly follow [3]. We are cooperating with Lexpera d.o.o., a company that provides a legal information portal. A system that is capable of producing summaries of long legal documents can therefore serve as a tool for the users to quickly skim over documents to gain information. Furthermore, such system may allow users to find relevant search results faster.

3.2 Related work

The approaches to automatic summarization can be generally divided into extraction-based and abstraction-based ones. Extractive summarization reformulates the problem to selecting the most important subset of original document's sentences. On the other hand, abstractive summarization attempts to build a representation of the original document and use it to construct a summary in a paraphrasing manner, i.e. using other words. We use both types of approaches in our work.

Some of the early successes of extractive approaches were achieved by constructing a graph of vertices representing sentences and edges representing some notion of sentence similarity. Approaches, such as LexRank [4] and TextRank [5], use PageRank algorithm to select the top ranking sentences. They can work in an unsupervised manner, but are outperformed by more recent work. Neural approaches treat extractive summarization as a binary sentence classification, i.e. whether a sentence belongs to the summary [6][7]. Cheng and Lapata [6] propose an encoder-extractor framework consisting of two RNNs where the former encodes the sentences and the latter acts as a binary classifier. Neural extractive approaches achieved a large performance gain when combined with pre-trained language models, such as BERT [8]. Liu et al. [9] build an extractive model by adding transformer layers on top of a pre-trained BERT and jointly training the model for binary sentence classification. The approach is, however, limited by BERT's fixed input size.

Abstractive summarization is in its essence a sequence-to-sequence problem, therefore encode-decoder architectures can be used. Nallapati et al. [10] use RNNs for both encoder, decoder and utilize the attention mechanism allowing decoder to put focus on useful parts of the encoded input. This approach, however, struggles with out-of-vocabulary words and tends to repeat itself. Coverage mechanism, first introduced for machine translation [11], is used by See et al. [12] to mitigate repetition. Nevertheless, RNNs based approaches inherently perform worse as the length of input increases. More recently, in the transformers era of NLP, a pre-train and fine-tune approach has gained popularity. To that end, Zhang et al. [13] propose Pegasus, a transformer encoder-decoder that is pre-trained to generate masked important sentences in the input. This pre-trained model is then further fine-tuned on actual summarization datasets and achieves state-of-the-art on several important benchmarks.

In recent research, little emphasis has been put into summarizing legal documents, mostly due to their length, but also due to the lack of standard datasets. However, recently Kornilova et al. introduced BillSum [3], a summarization dataset comprised of US Congressional Bills. Moreover, they evaluate several extractive approaches, showing that some of these methods also work on long legal texts.

3.3 Expected contributions

Our main contributions are twofold. We will develop an extractive and an abstractive neural network based summarization system for legal documents. Furthermore, we will provide insights into constructing summarization corpora, in terms of how document length, size of the train set and annotations affect performance. This will prove useful for our collaborators at Lexpera, to utilize developed summarization methods on new data and for different languages (Slovene, Croatian, Turkish, English).
In our work we will implement extractive and abstractive summarization methods. LexRank and sequence-to-sequence RNNs will serve as our respective baselines, that we will try to improve upon with more recent methods. We will evaluate developed models on legal corpora, with which we aim to achieve further contribution of providing an overview of these methods' adaptivity to legal domain and thereby expanding the limited literature on the subject of legal summarization.

3.4 Methodology

We will approach our work with the following steps:

1. Review and acquisition of data sources. We will work closely with Lexpera to review their collection of legal documents and find relevant datasets. Furthermore, to make our work relevant in the more broader research field, we will utilize some standard legal summarization datasets, such as recently proposed BillSum [3].

2. Review of methods/models from the literature, both relevant work in the field of legal summarization and more broader field of automatic summarization.

3. Implementation of relevant models and adaption to legal domain. For recent methods, such as Pegasus [13], there exist large pre-trained general models, that we will fine-tune on our datasets. All our implementations will be in Python with PyTorch as the preferred deep learning framework.

4. Evaluation and analysis. We will evaluate our trained models using the ROUGE metric [14] that is commonly used for summarization. We will analyze the effects of document length and training set size on performance to provide Lexpera with information on how to approach constructing a summarization corpora.

We will structure our work in a cyclic manner, meaning that after evaluating the methods and analyzing their strengths and weaknesses, we will go back and reassess both the data and methods with the aim to improve and better understand them.

3.5 References

[1] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, P. Blunsom, Teaching machines to read and comprehend, in: Advances in Neural Information Processing Systems, 2015, pp. 1693–1701.

[2] A. Cohan, F. Dernoncourt, D. S. Kim, T. Bui, S. Kim, W. Chang, N. Goharian, A discourse-aware attention model for abstractive summarization of long documents, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics, 2018, pp. 615–621.

[3] A. Kornilova, V. Eidelman, Billsum: A corpus for automatic summarization of US legislation, in: Proceedings of the 2nd Workshop on New Frontiers in Summarization, 2019, pp. 48–56.

[4] G. Erkan, D. R. Radev, Lexrank: Graph-based lexical centrality as salience in text summarization, Journal of Artificial Intelligence Research 22 (2004) 457–479.

[5] R. Mihalcea, P. Tarau, Textrank: Bringing order into text, in: Proceedings of the 2004 conference on Empirical Methods in Natural Language Processing, 2004, pp. 404–411.

[6] J. Cheng, M. Lapata, Neural summarization by extracting sentences and words, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016, pp. 484–494.

[7] R. Nallapati, F. Zhai, B. Zhou, SummaRuNNer: a recurrent neural network based sequence model for extractive summarization of documents, in: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, 2017, pp. 3075–3081.

[8] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, pp. 4171–4186.

[9] Y. Liu, M. Lapata, Text summarization with pretrained encoders, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019, pp. 3721–3731.

[10] R. Nallapati, B. Zhou, C. Gulcehre, B. Xiang, et al., Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond, in: Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning, 2016, pp. 280–290.

[11] Z. Tu, Z. Lu, Y. Liu, X. Liu, H. Li, Modeling coverage for neural machine translation, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016, pp. 76–85.

[12] A. See, P. J. Liu, C. D. Manning, Get to the point: Summarization with pointer-generator networks, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017, pp. 1073–1083.

[13] J. Zhang, Y. Zhao, M. Saleh, P. Liu, Pegasus: Pre-training with extracted gap-sentences for abstractive summarization, in: International Conference on Machine Learning, PMLR, 2020, pp. 11328–11339.

[14] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text summarization branches out, 2004, pp. 74–81.