Extractive Text Summarization Using Topic Modeling, Topic Clustering and TextRank

Eva Richter

March 15, 2021

Abstract

In this project, a LDA model, a statistical model that we were taught in the 5th CL assignment, was used. The LDA model with Gibbs Sampling was used to identify and extract the most dominant topics from a set of documents and create an extractive text summarization by clustering the sentences into dominant topics and extracting the summary of each cluster. Consequently, the combination of those summaries form the final extractive summarization for each document. The summary of the clusters was obtained by implementing TextRank, a ranking model, which finds the most relevant sentences in texts and ranks them according to their relevance. Finally the extractive summaries generated by the LDA model were evaluated using ROUGE-N, ROUGE-L and ROUGE-W. The ROUGE results that the model achieved are decent and roughly in line with the results obtained in other papers on automatic text summarization.
The data set used for this project is provided by *SwissText 2019*, containing 100,000 documents together with reference summaries extracted from the German Wikipedia.

1 Introduction

Due to the growing amount of textual material available on the internet, there is a trade-off between the vast, unstructured amounts of data and the time required to read it. In order to reduce the time and effort required to navigate such volumes of text, there is a need to condense the important information in form of summaries [Gambhir and Gupta, 2016].
Apart from that, automatic text summarizations are useful in other contexts as well, like for improving the effectiveness of indexing or facilitating selection processes when researching documents. Furthermore, automatic text summarizations are less biased compared to summaries performed by humans and personalized summaries are convenient in question-answering systems since they provide personalized information [Torres-Moreno, 2014].

2 Automatic Text-Summarization

Automatic text summarization is a sub-discipline of natural language processing, which deals with summarizing source texts while retaining the most important information from the source texts. In the context of different requirements and application possibilities, many approaches of automatic text summarization have been developed. The two general classifications of text summarization are extractive text summarization and abstractive text summarization [Gambhir and Gupta, 2016].

2.1 Extractive Text Summarization

Despite having low coherence, extractive text summarizations are widely used in the field of automatic text summarization since they are less time-consuming and more easy to generate than abstractive text summarizations.
As the name implies, the goal of extractive text summarization is to generate a summary based on a combination of the salient sentences from a source document. In order to achieve this, all sentences in the source document are assigned weights, depending on which the highest ranked sentences are extracted and combined to form a summary [Issam et al., 2020].

2.2 Multi-Document Summarization

Text summarization can be performed either by identifying and summarizing salient information from one document ("single-document summarization") or by considering multiple documents about the topic ("multi-document summarization"). Most state-of-the-art models for single-document summarizations don't achieve good results when applied to longer texts, which is why the multi-document summarization will be applied in this project. Applied to extractive text summarization, multi-document-summarization aims to pick out the most relevant sentences for a topic through a multi-document analysis and combines them to produce a representative summary for all documents relevant to a topic [Issam et al., 2020].

3 LDA Topic Modeling

Topic modeling is a form of text mining and a method of identifying patterns in a corpus, where the topics generated by topic modeling techniques are clusters of similar words. Based on the statistics of words in a document or a set of documents, topic models identify which topics those documents may contain and how the balance of each topic appearing in each document is. Topic models are generally used for different tasks where knowing the topics of documents is crucial, like for example in the context of text summarization in order to create a summary for each document based on the extraction of the most salient sentences.

Latent Dirichlet Distribution (LDA) is a very common example of a topic model, which is applied in order to identify abstract topics that are present in the data set. A Dirichlet distribution refers to a distribution of distributions. Hence LDA is a statistical approach that projects the distribution of documents and topics to the distribution of words and topics. Since each topic is latent and represented by a distribution of words that appear in the data set, each document in the data set is characterized as the distribution of those topics conforming to LDA [Issam et al., 2020].

3.1 Topic Modeling Summarization

Topic modeling text summarization aims to consider a document composed of different topics, which can be classified in the source document by using topic modeling techniques like LDA. These topics are used to create text clusters containing salient sentences from the source document. The clusters are connected to the relevant topics in the source document and the sentences of this document are assigned to the identified topics to achieve better coverage for the summarization of the document. The clusters that the LDA model generates split the document into subdocuments. The subdocuments are then summarized using TextRank to create the extractive summary for the input document [Issam et al., 2020].

4 TextRank

TextRank is a graph-based ranking model for text processing [Mihalcea and Tarau, 2004], which was adapted from PageRank, but is used for ranking sentences instead of web pages. After combining the text in all the documents, the text rank algorithm tokenizes them into sentences and converts these sentences by means of word embeddings in vectorized forms. Based on the vectors, TextRank establishes a similarity matrix, measuring cosine similarity in order to generate a graph of sentences and the summary by selecting the most important sentences [Issam et al., 2020].

5 Problem Statement

In this project a method to perform extractive text summarization is implemented. The topic we covered in class for LDA based topic modeling devises the significant topics, which make up each document in a set of documents. Hence, based on the balance of topics in the documents, the most important sentences are extracted to represent the summaries of the documents.

However, due to its complexity, LDA is a challenging task itself. The problem is to train multiple documents into the LDA model, which requires a lot of time and resources. The extraction of summaries from unseen documents by the model is an equally challenging task, since it assumes that the model learns general topics that are related to the unseen documents during the training of the LDA topic model. Additionally, the data set that was used contains reference summaries for abstractive text summarization. Hence, benchmarking the data set for extractive summarization is challenging in the sense that the reference summaries are not identical, as they often contain paraphrased sentences from the document.

In overall terms, training the topic model, extracting summaries for unseen test documents, and finally comparing the automatic summaries with the reference summaries, which are originally intended for abstractive summarizations, as well as analyzing the results are the problems covered in this project.

6 Methodology

6.1 Data Analysis

The data set on which this project is based is provided by *SwissText 2019* and contains 100,000 documents together with reference summaries for abstract text summarizations extracted from the German Wikipedia.
Since 100 documents are designated as seen test data and 100 more documents as unseen test data, 99 900 documents are labelled as the training corpus on which the Topic Model is trained.
The maximum number of sentences a document from the training data contains is 195 and the minimum number 4. In the unseen test data set, the maximum number of sentences is 115 and the minimum number 20, while a document from the seen test data set contains a maximum of 106 sentences and a minimum of 7 sentences per document.
Regarding the number of words per document, the maximum word count measured for a document in the training data is 1924 and the minimum word count is 79. The maximum number of words appearing in a document from the seen test data is 1751 and the minimum number 133. With respect to the unseen test data, a word appears maximally 1734 and minimally 286 times in a document.

6.2 Data Preprocessing

First, pre-processing techniques are applied to the *SwissText 2019* data set in order to improve the topic modeling results.
Stop words and unnecessary characters are removed. This includes punctuation, numbers, unnecessary spacing, but also words occurring with a frequency higher than 25 times per document, which are mostly function words, and words that occur with a lower frequency than 10 times per document, since they have no importance for the results. In the next step, the data set is tokenized: all the documents are split into sentences and each word in each document is converted to lowercase form. After that, for each document the tokens are lemmatized, i.e converted to their base forms and joined to a list of token sequences for the whole data set. Finally, bigrams are generated from the tokenized sentences in each document, since including frequently appearing phrases as single units improves the results.

6.3 Topic Modeling

After the data is cleaned, it is divided into three parts: 100 documents are labelled as unseen test data, 100 documents are labelled as seen test data and the rest of the 100 000 documents, which make up the *SwissText 2019* corpus, is labelled as training data. The training data is then passed into the LDA model and trained to get the top 20 topics. This model is saved to be used during the extraction of the summary.

6.4 Topic Clustering

The pre-processed document is then passed to extract the dominant topics using the previously trained LDA module. After getting the dominant topics, the dominant topic ID of each sentences is identified and mapped to the dominant topic ID of the document. Doing this, clusters of sentences belonging to each dominant topic are created, which are then passed to the TextRank algorithm. In case only one sentence for a topic is identified, this sentence is used for the summary of the topic. If multiple sentences are identified, TextRank selects the most important sentences, of which the summary will ultimately be composed.

6.5 Document Summarization

Finally, a summary for each dominant topic in the document is generated from each cluster, using the TextRank algorithm. These extracted sentences are merged to devise the summary of the whole document. In case only one sentence is identified for a topic, this sentence will be used to form the summary of the topic, but if multiple sentences are identified, TextRank selects the most important sentences of which the automatic summary will finally be composed.

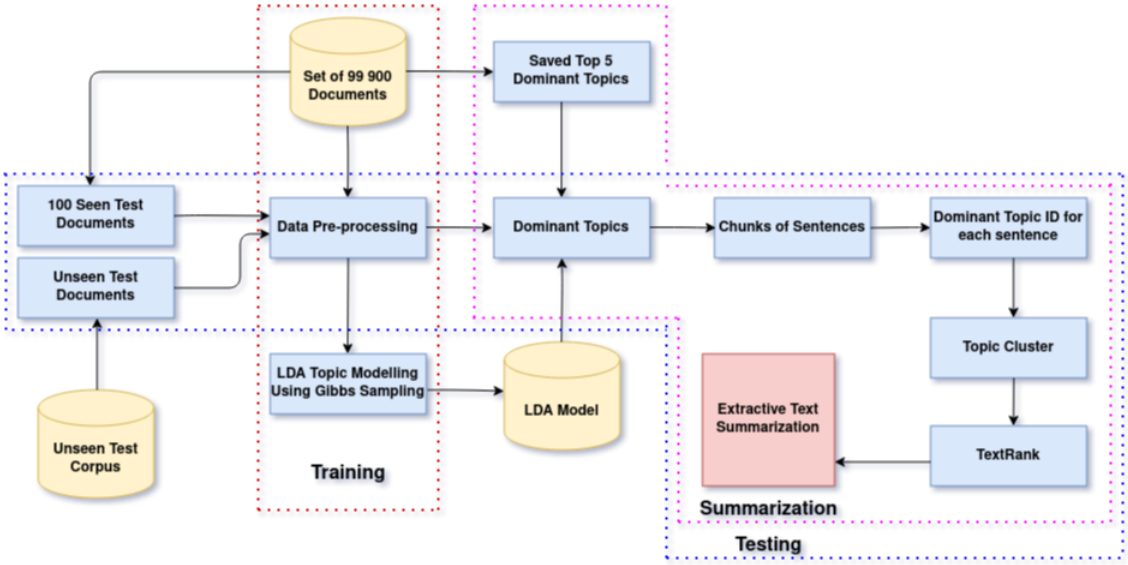

Figure 1: System Architecture

7 ROUGE

The most prominent evaluation metric for automatic text summarizations is ROUGE (Recall-Oriented Understudy for Gisting Evaluation). It compares the automatically produced summaries against reference summaries available in the source data set, which are normally produced by humans.

The ROUGE metric set contains five evaluation metrics, three of which are used for the following evaluation of the extractive text summarizations. ROUGE-N (ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4) measures the quality of automatic summaries in terms of the overlap of N-grams between the system and reference summaries. ROUGE-L measures the longest co-occurring n-grams between the reference summary and the automatic summary and ROUGE-W evaluates the weighted longest common subsequence between the summaries to be evaluated. The general ROUGE equation can be represented as follows:

$$ROUGE - N = \frac{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)} \quad (1)$$

However, it should be mentioned that despite being the most common evaluation metric, the ROUGE scores are in many cases not a good indicator for the summaries' quality, since they assign higher scores to summaries whose wording resembles the reference summaries than to summaries that have been rephrased to a higher degree [Issam et al., 2020].

8 Result Analysis

	Precision	Recall	F-Score
ROUGE-1	16.87	31.44	**19.96**
ROUGE-2	3.49	7.55	4.30
ROUGE-3	0.39	2.45	1.23
ROUGE-4	0.36	1.09	0.50
ROUGE-L	18.01	31.45	**21.37**
ROUGE-W	10.29	11.27	9.30

Table 1: average ROUGE scores on seen test data

	Precision	Recall	F-Score
ROUGE-1	21.34	24.75	**20.99**
ROUGE-2	4.11	4.47	3.94
ROUGE-3	1.29	1.46	1.25
ROUGE-4	0.47	0.51	0.46
ROUGE-L	20.90	24.55	**21.14**
ROUGE-W	11.70	6.99	7.68

Table 2: average ROUGE scores on unseen test data

The seen and unseen set of documents and their summaries were evaluated using ROUGE-N, ROUGE-L and ROUGE-W. The main purpose of analysing the results for unseen test documents is to analyse the multi-document summary process, where whole documents are modelled into LDA. Even though the LDA model has not seen the test data before, it was able to extract topics from it, which resulted in an average F1-score of approximately 20% as shown in Table 2.

The best results on the seen test data are provided by ROUGE-L, which shows an F1-score of 21.37 and ROUGE-1 with an F1-score of 19.96. The precision score of 16.87% of the ROUGE-N metric indicates that 16.87% of the N-grams in the automatically produced summary are also present in the base summary, while the recall of 31.44 shows that the automatically generated summary contains 31.44% of the N-grams which are also identified in the base summary. ROUGE-W and especially ROUGE-2, -3, -4 give significantly worse results, the worst of which come from ROUGE-4, indicating an F1-score of only 0.46.
The results for the unseen test data are similar. Indicating an F1-score of 21.14, ROUGE-L again provides the best result, followed by ROUGE-1 with an F-score of 20.99. While the F1-score for ROUGE-1 for the unseen test data is slightly higher than on the seen test data, the F1-score for ROUGE-L on the seen test data is slightly higher than for the unseen test data. As can be expected, the scores of ROUGE-W and ROUGE-2, -3, -4 for the unseen test data are similarly poor as for the seen data.

As mentioned, the results on the seen and unseen test data do not show big discrepancies. Since the LDA model also modelled the unseen data nicely, the ROUGE scores on the unseen data are similarly good as the ROUGE scores on the seen data, in some cases even better. Despite minimal differences, ROUGE-L and ROUGE-1 provide the best F1-scores, whereas ROUGE-W and ROUGE-2, -3, -4 deliver poor scores. The good results of ROUGE-L are due to the fact that the longest common N-grams in a sequence are automatically included without having to predefine

an N-gram length. Furthermore, it appears that regarding ROUGE-N, the F1-score decreases as the N-gram parameter in ROUGE increases. This is because it is likely for a single word to be identical in the automatically generated summary and in the reference summary. The longer the considered sequence of words is, the less likely it becomes that all these words overlap in the automatic summary and the reference summary, since the same content can be expressed by synonyms or paraphrases, which ROUGE does not take into account.
In addition, the reference summaries are actually generated for abstractive summarizations. Hence, the result could be presented better with actual extractive sumamries. However, due to the lack of available data sets for the German language, the comparison was opted with abstractive reference summaries. Nonetheless, it can be concluded that the system performs nicely and produces F1-scores similar as in [Issam et al., 2020].

A test document from the *SwissText 2019* data set is given below.

Nach dem Abitur an der Gisela-Oberrealschule in Munchen-Schwabing studierte Hofmann, dessen Vater Regierungsvizeprasident in Ansbach war, Rechts- und Staatswissenschaften an der Ludwig-Maximilians-Universitat Munchen sowie der Friedrich-Alexander-Universitat Erlangen-Nurnberg. Er war danach von 1958 bis 1961 als Regierungsassessor und zuletzt als Regierungsrat in Augsburg und Schwabmunchen innerhalb der bayerischen Innenverwaltung tatig. Wahrend seines Studiums in Erlangen wurde er Mitglied der Studentenverbindung "Corps Bavaria Erlangen". 1961 wurde Hofmann Personlicher Referent von Walter Scheel, der kurz zuvor zum ersten Bundesminister fur wirtschaftliche Zusammenarbeit ernannt worden war. Nach vierjahriger Tatigkeit wurde er 1965 in diesem Bundesministerium Referatsleiter fur Sudostasien und Ostasien, ehe er vom 1. Januar 1969 an Leiter des dortigen Referats fur internationale Fragen der Entwicklungspolitik war. Nachdem Walter Scheel Bundesaussenminister der sozialliberalen Koalition geworden war, ernannte er Hofmann im Oktober 1969 zum Leiter des Ministerburos des Auswartigen Amtes. Im September 1972 folgte die Beforderung zum Ministerialdirigenten und Ernennung zum Leiter des Leitungsstabes des Auswartigen Amtes. Im Anschluss daran wurde Hofmann im Juli 1973 vom Bundesvorstand der FDP zum Bundesgeschaftsfuhrer der FDP berufen und ubernahm dieses Amt am 1. September 1973 als Nachfolger von Joachim Stancke. Wahrend dieser Zeit war er zugleich auch Geschaftsfuhrer Inland der Friedrich-Naumann-Stiftung. Hofmann ist Mitglied der Jury des Verbandes Liberaler Akademiker zur Vergabe des Arno-Esch-Preises. 1977 kehrte er ins Auswartige Amt zuruck und wurde als Nachfolger von Werner Ahrens zum Botschafter in Danemark ernannt und behielt dieses Amt bis zu seiner Ablosung durch Rudolf Jestaedt 1981. Wahrend dieser Zeit war er unter anderem 1980 auch Stellvertreter des Parlamentarischen Staatssekretars im Bundesministerium fur Jugend, Familie und Gesundheit, Fred Zander, als Leiter der Delegation beim Frauenkongress der Vereinten Nationen in Kopenhagen. Danach wurde Hofmann 1981 Nachfolger von Helmut Redies als Botschafter in Venezuela und bekleidete diese Funktion bis zu seiner Ablosung durch Hans Werner Loeck 1985. Danach erhielt er seine Akkreditierung als Botschafter in Norwegen, wo er Nachfolger des im Amt verstorbenen Johannes Balser wurde. Als solcher hielt Hofmann auch Vortrage vor der Deutsch-Norwegischen Gesellschaft zu bilateralen diplomatischen Beziehungen. Nach siebenjahriger Verwendung in Norwegen folgte ihm 1992 Helmut Wegner, wahrend er wiederum Nachfolger von Reinhold Schenk als Botschafter in Schweden wurde. Das Amt des Botschafters in Schweden ubte Hofmann bis zu seiner Versetzung in den Ruhestand 1997 aus.

After processing the text, the LDA topic model is run on the test document to find the topics from this test document. The model generates a number of text clusters, which corresponds to the number of topics, identified in the text. The clusters created around the identified topics contain important text from the source document, which is relevant to the particular topics, that are identified. The clusters for this text generated by the model are shown below.

15: "Danach erhielt er seine Akkreditierung als Botschafter in Norwegen, wo er Nachfolger des im Amt verstorbenen Johannes Balser wurde"; "Nach siebenjahriger Verwendung in Norwegen folgte ihm 1992 Helmut Wegner, wahrend er wiederum Nachfolger von Reinhold Schenk als Botschafter in Schweden wurde"; "Das Amt des Botschafters in Schweden ubte Hofmann bis zu seiner Versetzung in den Ruhestand 1997 aus"; "Hofmann ist Mitglied der Jury des Verbandes Liberaler Akademiker zur Vergabe des Arno-Esch-Preises"; "Wahrend seines Studiums in Erlangen wurde er Mitglied der Studentenverbindung "Corps Bavaria Erlangen"

6: "Nach vierjahriger Tatigkeit wurde er 1965 in diesem Bundesministerium Referatsleiter fur Sudostasien und Ostasien, ehe er vom"; "Das Amt des Botschafters in Schweden ubte Hofmann bis zu seiner Versetzung in den Ruhestand 1997 aus"; "Wahrend dieser Zeit war er zugleich auch Geschaftsfuhrer Inland der Friedrich-Naumann-Stiftung"; "1961 wurde Hofmann Personlicher Referent von Walter Scheel, der kurz zuvor zum ersten Bundesminister fur wirtschaftliche Zusammenarbeit ernannt worden war"; "ls solcher hielt Hofmann auch Vortrage vor der Deutsch-Norwegischen Gesellschaft zu bilateralen diplomatischen Beziehungen"

2: "Hofmann ist Mitglied der Jury des Verbandes Liberaler Akademiker zur Vergabe des Arno-Esch-Preises"

0: "Danach erhielt er seine Akkreditierung als Botschafter in Norwegen, wo er Nachfolger des im Amt verstorbenen Johannes Balser wurde"

In the final step, a summary for each cluster is created, which are combined to form the extractive text summarization of a test document, on which the model is run. In this example the generated summary consists of the two sentences shown below.

Hofmann ist Mitglied der Jury des Verbandes Liberaler Akademiker zur Vergabe des Arno-Esch-Preises. 1961 wurde Hofmann Personlicher Referent von Walter Scheel, der kurz zuvor zum ersten Bundesminister fur wirtschaftliche Zusammenarbeit ernannt worden war.

Even though the reference summary shown below is a little shorter than the summary extracted by the model and the generated summary doesn't contain all of the salient topics from the reference summary, most of the salient topics are included in both.

Harald Hofmann ist ein deutscher Jurist und ehemaliger Diplomat, der unter anderem Bundesgeschaftsfuhrer der FDP und Botschafter in Danemark, Venezuela, Norwegen und Schweden war.

9 Conclusion

Extractive text summarizations have gained great popularity as they provide a way to extract the most important information from texts and link it into a summary, greatly reducing the time spent navigating large amounts of text. Hence, in this project the implementation of a statistical model, i.e. LDA for topic modelling, was used successfully to extract significant information in the form of dominant topics representing sentences.
The results for the evaluated summaries were good for ROUGE-1 and ROUGE-L, poor for ROUGE-2, -3, -4 and ROUGE-W. Overall, the results were decent and roughly in line with the results obtained in other papers on automatic text summarization. Including extractive instead of abstractive reference summaries would have improved the results, but couldn't be realized due to the lack of such German data sets.
Although extractive summaries have distinct limitations and are less complex and coherent than abstract summaries, their importance is increasing as they provide a simple and robust method of summarising (large) amounts of text.

The implementation of such an extractive summarization method has broadened my understanding of the LDA model and given me an insight on Natural Language Processing pipelines. Conse-

quently, an important domain of Natural Language Understanding was implemented and analysed in this project.

References

Mahak Gambhir and Vishal Gupta. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, 47:1–66, 2016.

K.A.R. Issam, S. Patel, and C.N. Subalalitha. Topic modeling based extractive text summarization. *International Journal of Innovative Technology and Exploring Engineering*, 9/6:1710–1719, 2020.

R. Mihalcea and P. Tarau. Textrank: Bringing order into texts. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, 2004.

J.-M. Torres-Moreno. *Automatic Text Summarization*. Wiley-ISTE, 1st edition, 2014.