Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6651–6655
Marseille, 11–16 May 2020

Summarization Corpora of Wikipedia Articles

Dominik Frefel
University of Applied Sciences Northwestern Switzerland
Institute of Data Science, Bahnhofstrasse 6, 5210 Windisch
dominik.frefel@fhnw.ch

Abstract

In this paper we propose a process to extract summarization corpora from Wikipedia articles. Applied to the German language we create a corpus of 240 000 texts. We use ROUGE scores for the extraction and evaluation of our corpus. For this we provide a ROUGE metric implementation adapted to the German language. The extracted corpus is used to train three abstractive summarization models which we compare to different baselines. The resulting summaries sound natural and cover the input text very well. The corpus can be downloaded at https://github.com/domfr/GeWiki.

Keywords: Machine Learning, Natural Language Processing, Summarization, Corpus

1. Introduction

Advances in deep learning have led to the development of a variety of sequence-to-sequence models (Sutskever et al., 2014; Bahdanau et al., 2014) to solve natural language tasks. Rush et al. (2015) were the first to apply these models to abstractive summarization. More recently, the Transformer model has reached state of the art performance in various natural language processing tasks (Vaswani et al., 2017) and has successfully been used for abstractive summarization (Gehrmann et al., 2018; Hoang et al., 2019). These models require a large number of text samples for the training and evaluation.

Large summarization corpora are only sparsely available and most research is focused on the English language. We explore a process to extract summarization data for other languages. Giannakopoulos et al. (2015) use Wikipedia featured articles to evaluate on multiple languages, but their data is not sufficient to train an abstractive model. Based on their idea, we propose a method to extract summarization data from Wikipedia articles and apply it to the German language.

This paper is structured as follows: Section 2 describes our method of corpus construction. In section 3 we present our German corpus and in section 4 our experiments and results. Section 5 contains conclusions.

2. Corpus construction

2.1. Source

Featured Wikipedia articles are required to have a summarizing introduction of the main body (Wikipedia-Featured-articles, 2019). There are only about 2000 featured articles written in German, which is not enough to train deep learning models. However, numerous ordinary articles have a summarizing introduction as well. Our goal is to identify texts with introductions that can be used as a summary of the rest of the article. We hypothesize that useful texts can be selected using only a few metrics, namely: Compression ratio, summary length and ROUGE scores ($\alpha = 0$)[1] of the introduction.

2.2. German Rouge Metric

We define a ROUGE implementation[2] adapted to the German language following the work of Lin (2004). It applies a number of preprocessing steps:

- Split the given text into sentences
- Tokenize each sentence
- Remove stop words
- If possible, replace tokens with their Germanet baseform (Hamp and Feldweg, 1997) (Henrich and Hinrichs, 2010)
- Split compound words
- Stem each token using the German Snowball stemmer (Porter, 1980)
- Replace all uppercase with lowercase letters
- Replace the corresponding umlaut with ”ae”, ”ue”, and ”oe”
- Replace all ”ß” with ”ss”

The preprocessing steps are similar to the English ROUGE script (pyrouge, 2019). The main difference is the splitting of compound words. Word compounding is common in German and can prevent partial matches. For example, if a generated summary contains the word ”car” and the reference the words ”police car” we measure ROUGE-1 $= 67$ in English. The German translation (”Polizei” and ”Polizeiauto”) has a score of ROUGE-1 $= 0$. With word splitting (”Polizeiauto” to ”Polizei” and ”Auto”), the summary has the same score in both languages.

2.3. Extraction and Processing

As illustrated in Figure 1 we observe that most Wikipedia featured articles have scores of ROUGE-1 $\geq$ 60 and ROUGE-2 $\geq$ 15. Furthermore, most have a compression ratio[3] of $\geq$0.025. To keep the word sequence lengths of the corpus in a manageable range, introductions are limited to a length of 25 to 150 words. We define that a text must fulfill these four restrictions to be included in the corpus. From

[1] With $\alpha = 0$ ROUGE measures the recall scores

[2] Our ROUGE implementation is available on Github: https://github.com/domfr/GeRouge

[3] Compressed text length divided by the uncompressed length

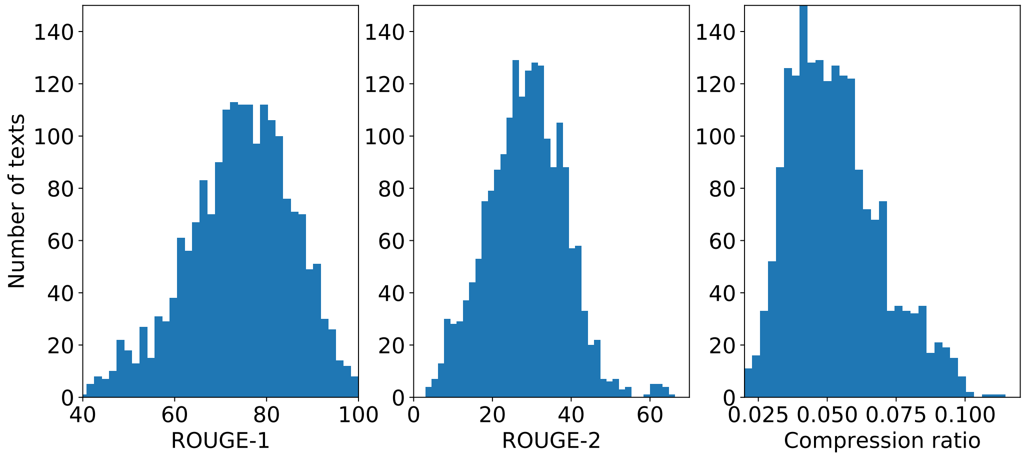

Figure 1: Metrics of German Wikipedia featured articles

the approximately 2 million German Wikipedia articles, we generate a corpus of 240 000 texts[4].

Wikipedia articles contain several items that are not useful for our purpose. In order to cleanup the text corpus, the following items are removed:

- Titles
- Table of contents
- Text inside of parentheses or brackets
- Hyperlinks (without removing the link label)
- Source annotations and labels
- Tables, images or any other media
- Enumerations

3. Corpus Statistics

The corpus is randomly split into training, evaluation and test sets. For each set the computed sentence and word statistics are reported in Table 1. Excluding the number of articles, all values represent the average per set.

Statistic	Train	Eval	Test
# articles	220 000	10 000	10 000
# sentences / text	26.3	26.3	26.7
# sentences / summary	2.1	2.1	2.1
# words / text	583.0	581.8	594.8
# words / summary	37.4	37.1	37.1
Compression ratio	19.0%	19.1%	19.3%

Table 1: Corpus statistics

The 240 000 articles are hierarchical categorized (Wikipedia-German-Categories, 2019). Refer to Table 2 for an overview of the most represented categories. Note that the proportions do not sum up to 100%. A single article usually belongs to more than one category.

The categories were not regarded during the creation of the training, evaluation and test set. However, the distribution of frequent categories are similar. Table 3 provides an overview of the most represented categories in each set.

Category	# articles	Proportion
People	65 776	27.4%
Science and technology	61 401	25.6%
Nations and states	56 151	23.4%
Art and culture	40 847	17.0%
Society	24 499	10.2%
Politics	17 714	7.4%
History	13 677	5.7%
Sport	11 942	5.0%

Table 2: Most represented categories for the whole corpus

Category	Train	Eval	Test
People	27.5%	26.8%	26.6%
Science and technology	25.6%	25.0%	25.9%
Nations and states	23.4%	23.8%	23.1%
Art and culture	17.0%	17.1%	17.3%
Society	10.1%	11.0%	10.4%
Politics	7.4%	7.4%	7.3%
History	5.7%	5.7%	5.6%
Sport	5.0%	4.9%	4.8%

Table 3: Categories per set

Category	Train	Eval	test
Male	87.4%	87.8%	87.5%
Female	12.6%	12.2%	12.5%
German	39.1%	40.4%	39.7%
American	15.6%	16.0%	15.5%
Austrian	4.7%	4.6%	4.5%
British	4.0%	4.3%	4.3%
French	2.8%	2.5%	2.4%
Swiss	2.6%	2.6%	2.2%

Table 4: Gender and nationality distribution of articles about people

More than a quarter of all articles are about a person, it is the most represented category. Of these articles, only about 12% describe a female person. They are mostly about German speaking people, followed by English and French. Refer to Table 4 for an overview of the distribution of gender and nationality.

[4] The summarization corpus is available on Github: https://github.com/domfr/GeWiki

	ROUGE-1	ROUGE-2	ROUGE-L
Random-3	18.12	5.52	14.17
Lead-3	21.32	7.05	16.81
Textrank (Řehůřek and Sojka, 2010)	23.56	8.00	17.56
Pointer-Generator (adapted from Klein et al. (2017))	40.80	23.64	33.39
Pointer-Transformer (Frefel, 2020)	45.91	26.80	37.16
Pointer-Transformer with extraction (Frefel, 2020)	**51.84**	**33.56**	**41.80**

Table 5: ROUGE scores

4. Experiments

4.1. Abstractive summarization

The corpus is evaluated on three different abstraction models. The first model is a Pointer-Generator network. We use the implementation of Klein et al. (2017) which is based on the work of See et al. (2017). The second model (Pointer-Transformer) is a Transformer network (Vaswani et al., 2017). We extend it with a copy generator mechanism similar to See et al. (2017) to allow for copying of tokens instead of generating (Frefel, 2020). The third model (Pointer-Transformer with extraction) utilizes sentence extraction prior to abstractive summarization (Frefel, 2020).

4.2. Baselines

The abstractive models are compared to several baselines. The Random-3 baseline randomly extracts three sentences from a given input text and combines them as a summary. Similarly, Lead-3 extracts the first three sentences of the input text. We further use the extractive Textrank algorithm (Mihalcea and Tarau, 2004) as implemented by Řehůřek and Sojka (2010) to pick the three most useful sentences as summary.

4.3. Discussion

The ROUGE scores ($\alpha = 0.5$)[5] of the models and baselines are given in Table 5. We use the ROUGE implementation as described in Section 2.2.. The abstractive models score considerably higher than the baselines. This indicates that the corpus can be used to successfully train abstractive summarization models.
Three example summaries of the most common categories are included in the Appendix. We analyzed the summaries manually and noticed that Wikipedia articles exhibit patterns that our models learned to replicate. The summary about Dieter Ehret (Example 7.1.) starts with: "Dieter Ehret ist ein baden-württembergischer Politiker." The sentence follows a simple pattern that is repeatedly found throughout the corpus: "[First name] [last name] is a [location] [occupation]". We speculate that these patterns help our models to reach high scores. This is an issue of using Wikipedia as data source. The summaries are nevertheless fairly abstractive and do not contain long copied sentences, which is a frequent problem of summarization algorithms. Apart from that, we did not find frequent problems or errors that result from our corpus. The biggest weakness of the models is that summaries can contain incorrect information. This happens in Examples 7.2. and 7.3. for some of our models. However, despite of occasional errors the summaries are useful, cover the corresponding article well and sound natural.

5. Conclusion

In this paper we introduce a new German summarization corpus built from Wikipedia articles and we propose a language independent process that can be used to create summarization corpora for other languages. We also present a German ROUGE implementation and use it to evaluate our models. We show that our corpus can be used to train abstractive models which produce useful summaries.

6. Acknowledgements

I would like to thank Prof. Dr. Manfred Vogel for his support, ideas and invaluable advice.

7. Appendix

7.1. Example 1 (People)

Reference summary Dieter Ehret ist ein baden-württembergischer Politiker, ehemaliges Mitglied der FDP und war Mitglied des Landtages von Baden-Württemberg.

Pointer-Generator Dieter Ehret ist ein deutscher Politiker. Er war von 2006 bis 2011 Abgeordneter zum Österreichischen Nationalrat.

Pointer-Transformer Dieter Ehret ist ein deutscher Politiker. Er war von 2006 bis 2011 Mitglied des Landtags von Baden-Württemberg.

Pointer-Transformer with extraction Dieter Ehret ist ein deutscher Politiker. Er war von 2006 bis 2011 Abgeordneter im baden-württembergischen Landtag.

Text Nach dem Abitur 1978 in Kenzingen studierte Dieter Ehret Bauingenieurwesen in Karlsruhe und Architektur in Wiesbaden. Seit 1984 ist er selbständiger Bauingenieur in Weisweil. Dieter Ehret war seit 1991 Mitglied der FDP. ... Vom 13. April 2006 bis zum Mai 2011 gehörte Ehret dem Landtag von Baden-Württemberg an. ... Im Februar 2010 wurde bekannt, dass Ehret für die Landtagswahlen 2011 nicht mehr kandidieren würde. Als Grund hierfür nannte er die Energiepolitik der schwarz-gelben Regierung. Er hält die vorgesehene Laufzeitverlängerung von Atomkraftwerken für inakzeptabel. Im Herbst 2011 erklärte Ehret schliesslich - ebenso wie zuvor bereits Schött - seinen Austritt aus der Partei, blieb aber parteiloses Mitglied der FDP-Fraktion im Kreistag. ...

[5] With $\alpha = 0.5$ ROUGE measures the F_1 scores

7.2. Example 2 (Science and technology)

Reference summary Die Schwarz-Weide oder Schwarzwerdende Weide ist eine Art aus der Gattung der Weiden in der Familie der Weidengewächse. Das natürliche Verbreitungsgebiet liegt in Europa und reicht bis ins westliche Sibirien. Sie wird sehr selten verwendet.

Pointer-Generator Die Schwarz-Weide ist ein Strauch Baum aus der Gattung der Weiden mit meist 6 bis 10 Zentimeter langen Blattspreiten. Das natürliche Verbreitungsgebiet der Art liegt in China.

Pointer-Transformer Die Schwarz-Weide ist eine Pflanzenart aus der Gattung der Weiden. Sie kommt im nordöstlichen China vor.

Pointer-Transformer with extraction Die Schwarz-Weide ist eine Pflanzenart aus der Gattung der Weiden in der Familie der Weidengewächse. Das natürliche Verbreitungsgebiet reicht von Nordeuropa bis nach Osteuropa und Südeuropa.

Text Die Schwarz-Weide ist ein dicht verzweigter Strauch oder selten Baum, der Wuchshöhen von 2 bis 5 Metern erreicht. Die dunkelbraune bis schwarzbraune Rinde der Zweige ist anfangs kurz samtig behaart, seltener mehr oder weniger kahl. ... Das natürliche Verbreitungsgebiet reicht von Nordeuropa über Mitteleuropa bis nach Osteuropa und Südeuropa. In Asien gibt es Bestände in Westsibirien. Die Schwarz-Weide wächst in Auen und an Gewässerufern auf frischen bis feuchten, schwach sauren bis alkalischen Untergrund aus Sand, Kies oder Schotter an sonnigen bis lichtschattigen, kühlen bis kalten Standorten. Das Verbreitungsgebiet wird der Winterhärtezone 4 zugeordnet mit mittleren jährlichen Minimaltemperaturen von -34,4 bis -28,9C. ... Die Schwarz-Weide wird nur sehr selten verwendet.

7.3. Example 3 (Nations and states)

Reference summary Der Landkreis Sankt Ingbert war ein Landkreis im Saarland. Zum Kreis gehörten 19 Bürgermeisterämter mit 27 Gemeinden, von denen 7 zu Ämtern zusammengefasst waren. Dazu zählten zwei Städte: St. Ingbert und Blieskastel.

Pointer-Generator Der Landkreis St. Ingbert war ein Landkreis im Bezirk Speyer in der Preussischen Provinz Kaliningrad.

Pointer-Transformer Der Landkreis St. Ingbert war ein Landkreis im Regierungsbezirk Saarbrücken in der Preussischen Rheinprovinz. Er umfasste am 1. Januar 1957 die Städte Zweibrücken und Gemeinden des Saarlandes.

Pointer-Transformer with extraction Der Landkreis St. Ingbert war von 1902 bis 1974 ein Landkreis in der Bundesrepublik Deutschland. Sein Gebiet gehört heute zum Saarpfalz-Kreis.

Text Der Landkreis grenzte 1973 im Uhrzeigersinn im Westen beginnend an die Landkreise Saarbrücken, Ottweiler und Homburg. Im Süden grenzte er an das französische Département Moselle. Der Landkreis St. Ingbert wurde am 1. Oktober 1902 aus Teilen des bayerischen Bezirksamtes Zweibrücken, nämlich aus dem Amtsgericht St. Ingbert und denjenigen Gemeinden des Amtsgerichts Blieskastel, welche den Distrikt gleichen Namens bildeten, zusammengesetzt. ... Am 1. Januar 1974 wurde der Landkreis im Zuge der Gebiets- und Verwaltungsreform zusammen mit dem Landkreis Homburg und dem Ort Rentrisch, jedoch ohne die Orte Ensheim, Eschringen und Schnappach, die an den Regionalverband Saarbrücken fielen, in den Saar-Pfalz-Kreis eingegliedert.

8. Bibliographical References

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Frefel, D. (2020). Abstractive summarization with extraction. submitted.

Gehrmann, S., Deng, Y., and Rush, A. M. (2018). Bottom-up abstractive summarization. *CoRR*, abs/1808.10792.

Giannakopoulos, G., Kubina, J., Conroy, J., Steinberger, J., Favre, B., Kabadjov, M., Kruschwitz, U., and Poesio, M. (2015). MultiLing 2015: Multilingual summarization of single and multi-documents, on-line fora, and call-center conversations. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 270–274, Prague, Czech Republic, September. Association for Computational Linguistics.

Hamp, B. and Feldweg, H. (1997). Germanet - a lexical-semantic net for german. In *In Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15.

Henrich, V. and Hinrichs, E. (2010). Gernedit - the germanet editing tool. pages 19–24, 01.

Hoang, A., Bosselut, A., Çelikyilmaz, A., and Choi, Y. (2019). Efficient adaptation of pretrained transformers for abstractive summarization. *CoRR*, abs/1906.00138.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. (2017). Opennmt: Open-source toolkit for neural machine translation. *CoRR*, abs/1701.02810.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

Mihalcea, R. and Tarau, P. (2004). TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain, July. Association for Computational Linguistics.

Porter, M. (1980). An algorithm for suffix stripping, program 14 (3).

pyrouge. (2019). https://pypi.org/project/pyrouge/0.1.3/.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA. http://is.muni.cz/publication/884893/en.

Rush, A. M., Chopra, S., and Weston, J. (2015). A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages

379–389, Lisbon, Portugal, September. Association for Computational Linguistics.

See, A., Liu, P. J., and Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. *CoRR*, abs/1704.04368.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *CoRR*, abs/1706.03762.

Wikipedia-Featured-articles. (2019). https://en.wikipedia.org/wiki/wikipedia:featured_articles.

Wikipedia-German-Categories. (2019). https://de.wikipedia.org/wiki/hilfe:kategorien.