information

MDPI

Article

Multilingual Text Summarization for German Texts Using Transformer Models

Tomas Humberto Montiel Alcantara [1], David Krütli [1], Revathi Ravada [1] and Thomas Hanne [2,*]

[1] School of Business, University of Applied Sciences and Arts Northwestern Switzerland, 4600 Olten, Switzerland; revathi.ravada@students.fhnw.ch (R.R.)

[2] Institute for Information Systems, University of Applied Sciences and Arts Northwestern Switzerland, 4600 Olten, Switzerland

* Correspondence: thomas.hanne@fhnw.ch

Abstract: The tremendous increase in documents available on the Web has turned finding the relevant pieces of information into a challenging, tedious, and time-consuming activity. Text summarization is an important natural language processing (NLP) task used to reduce the reading requirements of text. Automatic text summarization is an NLP task that consists of creating a shorter version of a text document which is coherent and maintains the most relevant information of the original text. In recent years, automatic text summarization has received significant attention, as it can be applied to a wide range of applications such as the extraction of highlights from scientific papers or the generation of summaries of news articles. In this research project, we are focused mainly on abstractive text summarization that extracts the most important contents from a text in a rephrased form. The main purpose of this project is to summarize texts in German. Unfortunately, most pretrained models are only available for English. We therefore focused on the German BERT multilingual model and the BART monolingual model for English, with a consideration of translation possibilities. As the source of the experiment setup, took the German Wikipedia article dataset and compared how well the multilingual model performed for German text summarization when compared to using machine-translated text summaries from monolingual English language models. We used the ROUGE-1 metric to analyze the quality of the text summarization.

Keywords: text summarization; natural language processing; language models

check for updates

Citation: Alcantara, T.H.M.; Krütli, D.; Ravada, R.; Hanne, T. Multilingual Text Summarization for German Texts Using Transformer Models. *Information* **2023**, *14*, 303. https://doi.org/10.3390/info14060303

Academic Editor: Katsuhide Fujita

Received: 29 March 2023
Revised: 16 May 2023
Accepted: 22 May 2023
Published: 25 May 2023

1. Introduction

Summarizing texts is becoming increasingly more relevant, due to the massive amount of information on the internet, as it saves time and prevents important information from being forgotten due to an excess of text data [1]. It is pertinent to mention the two most used techniques in text summarization. The extractive summarization focuses on identifying the salient text data to be extracted and clustered, to form a condensed and fluent summary. On the other hand, the abstractive text summarization focuses on compressing an extensive text into a shorter narrative which contains all the important details [2].

This project targets the contribution to the knowledge base of this broad and interesting topic of multilingual text summarization, especially in the multilingual aspect, as most available datasets for summarizations are in English; thus, there is a lack of multilingual data, which affects the performance in other languages. Therefore, we set up experiments with selected German texts, with the end purpose of comparing the quality of the summarized information. In the following Problem Statement, the project's problem is elaborated in order be able to formulate the Thesis Statement and the Research Questions.

1.1. Problem Statement

With the recent advancements in natural language processing (NLP), automated text summarization has gained relevance. Currently, there is a strong focus on the English language. However, the task of text summarization is relevant, irrespective of the language. For instance, companies and people in the DACH region (i.e., Germany, Austria, and Switzerland) would especially benefit from an automatic text summarization for the German language. Most people in this region speak German as their primary language. On the other hand, automatic text summarization could help the companies in this region convey information more easily (i.e., with fewer words).

1.2. Thesis Statement and Research Questions

There are different approaches to overcome this issue. One would be to use a multilingual model that supports the German language. The drawback of this is that these types of models are usually pre-trained on much fewer data than their English counterpart. However, as suggested by Bornea et al. [3], such models may be successful, due to transferring learning capabilities resulting from training data for different languages. Another approach would be to translate the text from German into English and then use an English pre-trained model to summarize the text and translate the resulting summary back into German. The disadvantage of this approach would be translation errors that could occur while processing the information. This research aims to carry out a comparison between the aforementioned approaches.

Thesis Statement: using machine translation to support automated text summarization with monolingual English language models would be a feasible approach when compared to using pre-trained, multilingual language models.

To investigate the thesis statement, the following Research Questions (RQs) are considered:

- RQ1: What is the current body of knowledge regarding automatic text summarization for languages such as German?
- RQ2: What language models can be used for automatic text summarization in German?
- RQ3: How could the language models be used to conduct experiments on automatic text summarization in German?
- RQ4: How should the data be processed?
- RQ5: What is the quality of the automatic text summarization for the particular dataset?

2. Literature Review

The following section shows the literature research that has been carried out in multilingual text summarization. It starts with the available literature on automated text summarization in general, and narrows down to German text summarization. For the literature research, search terms were, for instance, "multilingual text summarization" and "German text summarization" and forward as well as backward reference searching was used.

2.1. Automated Text Summarization

Automatic text summarizing provides summaries that incorporate all essential information from the original material and which include crucial sentences. As a result, the information is delivered swiftly, while maintaining the document's original objective. With the rise of the internet and big data, people are becoming overwhelmed by the vast amount of information and documents available on the internet. Many academics are motivated to create a technological solution that can automatically summarize texts [4].

> *"Text summarization approaches can be typically split into two groups: extractive summarization and abstractive summarization. Extractive summarization takes out the important sentences or phrases from the original documents and groups them to produce a text summary without any modification in the original text. Normally the sentences are in the same sequence as in the original text document. Nevertheless, abstractive summarization performs summarization by understanding the original text with the help of linguistic methods to understand and examine the text. The objective of abstractive summarization is to produce a generalized summary, which conveys information in*

> *a precise way that generally requires advanced language generation and compression techniques"*. [5] (p. 1)

Moratanch and Chitrakala [5] suggest that, in comparison to extractive summarizing, abstractive summarization is more efficient, since it pulls information from several texts to build an accurate summary of information. This has grown in prominence, due to its capacity to generate new phrases to convey essential information from text documents. An abstractive summarizer provides the summarized information in a cohesive, grammatically accurate, and easily understandable way. Note that there are also hybrid approaches which combine extractive and abstractive summarization techniques (e.g., by using them in different phases within the overall summarization process), which appear promising [6,7].

Some recent surveys of text summarization techniques are provided by Kanapala, Pal, and Pamula [8], Prudhvi et al. [9], El-Kassas, et al. [10] and Widyassari et al. [4]. Abstractive text summarization techniques have been further discussed, e.g., by Lin and Ng [11], Suleiman and Awajan [12] and Shi et al. [13].

The BERT model is a new language representation model that can be used to perform unsupervised pre-training using a large amount of text [14]. BERT, which stands for Bidirectional Encoder Representations from Transformers, is based on transformers, a deep learning model in which every output element is connected to every input element, and the weightings between them are dynamically calculated based upon their connection. Over the past few years, the BERT model has performed relatively well in natural language processing. Encoder and decoder models such as BERT are used for abstractive text summarization. The BERT framework was pre-trained using text from Wikipedia and can be fine-tuned with question-and-answer datasets. The BERT model utilizes a two-way transformer encoding layer to pre-train deep bidirectional representations of unlabeled text through conditional pre-processing on all layers using left-to-right and right-to-left processing [15].

Language model pretraining has significantly advanced the capabilities of many NLP tasks, ranging from sentiment analysis to question answering, natural language inference, named entity recognition, and textual similarity. Some famous pretrained models include ELMo [16], GPT and various successor models [17] and more recently Bidirectional Encoder Representations from Transformers, BERT [14].

BART is another model which is particularly effective for text generation but also works well for comprehension tasks. It matches the performance of other models such as RoBERTa with comparable training resources and performs very well in a range of abstractive dialogue, question answering, and summarization tasks. The model was developed by a team at Facebook [18].

There are two main strategies to employ pre-trained language models for various tasks: feature-based and fine-tuning. While feature-based approaches such as ELMo [16], use task-specific architectures that include the pre-trained representations as additional features, fine-tuning approaches such as in the Generative Pre-trained Transformer (OpenAI GPT) [17], introduce minimal task-specific parameters, and are trained on the downstream tasks by simply fine-tuning all pretrained parameters. The two approaches share the same objective function during pre-training, where they use unidirectional language models to learn general language representations.

Currently, pre-trained models are limited in capabilities, particularly for fine-tuning approaches. The primary constraint is that standard language models are unidirectional, which limits the available architectures for pre-training. For instance, OpenAI GPT uses a left-to-right architecture, where each token can only attend to preceding tokens in the self-attention layers of the transformer [19]. These restrictions are suboptimal for sentence-level tasks, and can be particularly harmful when employing fine-tuning-based approaches for token-level tasks such as question answering, where it is critical to incorporate context from both directions.

2.2. Multilingual Text Summarization

Most of the models used for text summarization such as BERT have been trained on English text data only, leaving lower-resource languages behind. There are some approaches to overcome this problem.

On the one hand, Machine Translation (MT) can be used to convert one language to another. Multilingual neural machine translation (NMT) is based on training a single model that supports translation from multiple source languages into multiple target languages. Aharoni et al. [20] showed that NMT models can successfully support up to 102 languages for translation to and from English. However, there are some drawbacks with using NMT. One of them is obviously quality issues due to translation errors.

Another approach is to take a multilingual model and use it to perform tasks such as summarization. Soon after the development of BERT by Devlin et al. [14], Google research introduced a multilingual version of BERT (also referred to as mBERT), capable of working with more than 100 languages [21]. The languages used to train the mBERT model were the top 100 languages with the longest Wikipedia entries. This includes the German language. Under the assumption that languages are competing for limited model capacity, however, certain low-resource languages may be under-represented in terms of the neural network model.

The introduction of the improved version of BERT [14] called RoBERTa has had significant impact and increased the relevance of pre-trained models. With GottBERT, a German single-language RoBERTa model was introduced. As a text corpus for the GottBERT model, the OSCAR data set was used [22].

Developers from deepset GmbH released a first German BERT model in 2019 and suggested further improved versions of the models in 2020. The models are pre-trained on the German OSCAR corpus, the Wikipedia dump for German, the OPUS project, and Open Legal Data [23].

With MLSUM, the first large-scale multilingual summarization dataset was introduced. It contains more than 1.5 million article/summary pairs in five different languages, i.e., French, German, Spanish, Russian and Turkish. The data was obtained from online newspapers and enable new research directions for the text summarization community [24].

2.3. ROUGE Metrics

ROUGE is an acronym that stands for Recall-Oriented Understudy for Gisting Evaluation. ROUGE can be used to assess the quality of a summarized text by comparing an automatically generated summary to a set of human-produced reference summaries. This comparison is based on the number of overlapping units known as n-grams, word sequences, and word pairs found in both summaries [25]. Furthermore Eyal et al. [26] emphasize that ROUGE is the most widely used method for evaluating automatic text summarization, with a high correlation with manual evaluation methods. For this reason, we use ROUGE metrics in our experiments, although there are promising alternatives such as an approach based on similarity scores considering contextual embeddings, which appears to provide results better coinciding with human judgments [27]. Another advanced approach, as discussed by Reimers and Gurevych [28], is that of using sentence embeddings to determine the similarity of texts. An evaluation of different approaches for evaluating the quality of text summarization methods is provided by Fabbri et al. [29].

For ROUGE, usually four evaluation methods are distinguished: ROUGE-N, ROUGE-L, ROUGE-W, and ROUGE-S(U) [25,30]. We focus on ROUGE-N, which calculates the overlap in unigrams, bigrams, trigrams, and higher-order n-grams between the generated summary and the reference summary, which is generally carried out by a person. The final score for the candidate summary is calculated using recall, precision, and the F1-score. The measure may adjust for the varied lengths of the candidate and reference summaries by using the F1-score [25].

It is also necessary to explain what an n-gram is; taking as a reference how [31] define it, an n-gram is a group of words or letters with n components that may be sorted.

The following is an example of how to calculate the F1-score from [25], by quantifying first the recall and precision scores from a machine-generated summary and then a human reference summary from the same text.

Machine-generated text:
"Switzerland is an amazing and very lovely country".

Human reference summary:
"Switzerland is an amazing and lovely country".

The ROUGE-$\text{N}_{\text{recall,}}$ counts the number of overlapping n-grams discovered in both the model output and the reference divided by the number of n-grams in the reference. In the above case of 1-g, seven of the seven words in the reference summary overlap.

$$\text{ROUGE} - \text{N}_{\text{recall}} = \frac{\text{Number of } n - \text{grams detected in model and reference}}{\text{Number of } n - \text{grams in reference}} = \frac{7}{7} = 1$$

The ROUGE-$\text{N}_{\text{precision}}$ is estimated in almost the same manner, except that instead of dividing by the reference n-gram count, we divide by the model n-gram count.

$$\text{ROUGE} - \text{N}_{\text{precision}} = \frac{\text{Number of } n - \text{grams detected in model and reference}}{\text{Number of } n - \text{grams in model}} = \frac{7}{8} = 0.875$$

The F1-score provides us with a credible measure of our model's performance, which is dependent not only on the model catching as many words as possible (recall), but also on doing so without producing unnecessary words (precision).

$$\text{ROUGE} - \text{N}_{\text{F1}} = 2 \times \frac{\text{ROUGE} - \text{N}_{\text{precision}} \times \text{ROUGE} - \text{N}_{\text{recall}}}{\text{ROUGE} - \text{N}_{\text{precision}} + \text{ROUGE} - \text{N}_{\text{recall}}} = 2 \times \frac{0.875 \times 1}{0.875 + 1} = 0.93$$

2.4. Research Gap

Experiments with German text summarization have already been conducted. Parida and Motlicek [2] highlighted an implementation for the abstract text summarization task under low resource conditions, which helps to improve the text summarization system in terms of automatic evaluation metrics. Tran and Kruschwitz [32] described a family of approaches to the task of multiclass fake-news classification for English and German. They used fine-tuned transformer architectures and incorporated extractive and abstractive summarization to help deal with long input documents. For the multilingual tasks, they also used automatic machine translation. The results demonstrated that both summarization techniques and automatic machine translation are competitive. There seems to be no further research being carried out on the comparison of results from using multilingual language models for German text summarization and monolingual English language models combined with machine translation. On the other hand, most of the research on text summarization deals with news articles, and it is therefore not so clear how well text summarization works for other types of text.

3. Research Design

Our project employs the DSR methodology as discussed by vom Brocke et al. [33] to conduct an experiment comparing text summarization performed by a multilingual pre-trained model to a monolingual English pre-trained model.

The DSR methodology begins with an awareness phase, which analyzes the current body of knowledge as specified in RQ1. In the suggestion phase, suitable language models for automatic text summarization in German are selected to answer RQ2. The development phase provides the specifications to conduct suitable experiments, including data processing with these models (for answering RQ3 and RQ4). In the evaluation phase, the quality of language models for text summarization is assessed, based on the experiments, to answer RQ5.

Derived from the research questions, the independent variable (IV) in this research design is the method used for text summarization. We try to carry out experiments using the following methods:

1. Use machine translation to translate German text into English, summarize the English text with a monolingual model and translate the summary back into German.
2. Use a multilingual model (supporting the German language) to generate a summary of a German text.

We define the primary dependent variable (DV) as the quality of summarization measured by the ROUGE value and by human assessment, as shown in Figure 1.

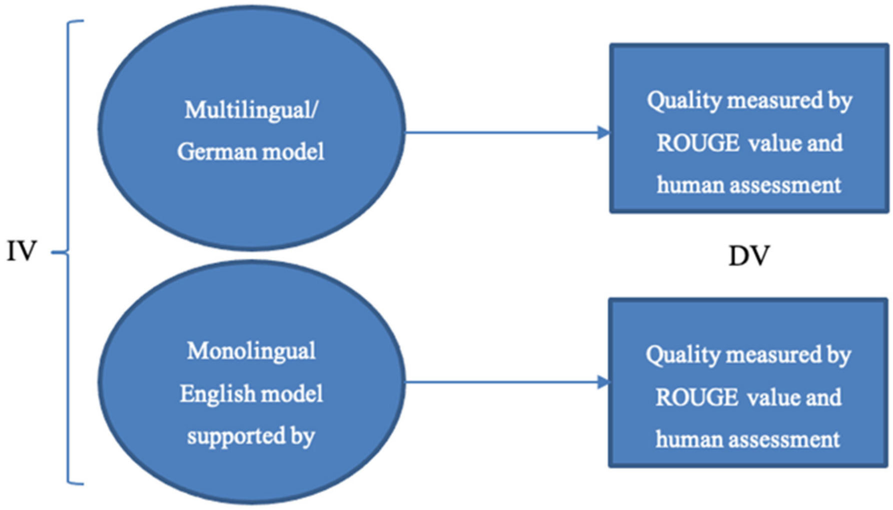

Figure 1. Independent (IV) and dependent (DV) variables.

4. Implementation

To conduct our experiments, we decided to work with the programming language Python. According to the official website [34], Python is an interpreted, high-level, general-purpose programming language. Created by Guido van Rossum and first released in 1991, Python's design philosophy emphasizes code readability, with its notable use of significant whitespace. Its language constructs and its object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects.

There are different libraries and APIs available in Python that can be used for automatic text summarization. We consider the software provided by Hugging Face, Inc. The company provides tools for building applications using machine learning. The Hugging Face hub is a platform that allows users to share machine learning models and datasets [35]. The transformers library is an ongoing effort maintained by the team of engineers and researchers at Hugging Face and supported by a large community of over 400 external contributors. The library is released under the Apache 2.0 license and is available on GitHub. Detailed documentation and tutorials are available on Hugging Face's website [36].

Hugging Face provides pre-trained models for a variety of natural language processing tasks, including text summarization. Some of the models are fine-tuned on specific datasets for the summarization task. The Hugging Face website provides a functionality to filter for relevant models by entering the task to be performed and the language. In our case, the task would be "Summarization" and the languages would be "English" and "German".

For our experiments, two models appear to be of special interest:

1. German BERT2BERT fine-tuned on MLSUM DE for summarization (https://huggingface.co/mrm8488/bert2bert_shared-german-finetuned-summarization, accessed on 16 January 2023): This model is based on the German BERT Model and was fine-tuned on the MLSUM DE dataset for summarization. The German BERT Base Model was trained on German Wikipedia, OpenLegalData, and news articles.

This model when implemented as it is, its showing very little sign of abstractive summarization. But can be considered as a starting point.

2. BART (large-sized model), fine-tuned on CNN Daily Mail (https://huggingface.co/facebook/bart-large-cnn, accessed on 16 January 2023): This model is pre-trained on the English language and fine-tuned on CNN Daily Mail articles. It was introduced by Lewis et al. [18] and matches the performance of RoBERTa, as well as achieving state-of-the-art results on a range of abstractive dialogue, question answering, and summarization tasks.

We were not able to identify a feasible multilingual language model such as mBERT that was already fine-tuned on text summarization. However, there is a German BERT model available.

Hugging Face provides an API for its pre-trained models. The API can be accessed by signing up for an API key on the Hugging Face's website. Once an API key is created, it can be used to make requests to the API and access the pre-trained models.

It is worth noting that Hugging Face's Transformer API is a service with a free and paid plan. The free plan allows you to use models that are smaller in size and perform a limited number of requests, whereas the paid plans give you access to larger models and more requests.

A typical API request for Hugging Face's Transformer API would involve making an HTTP request to a specific endpoint, using the API key for authentication. The request would include the input data, such as a text that needs to be summarized or a language that needs to be translated, and any other necessary parameters, such as the model to use or the specific operation to perform. The API would then process the request and return the results in the form of a JSON response.

For the evaluation of the quality of the automatically generated text summaries, the Python library "rouge" can be used. The get_scores method of this library returns three metrics, ROUGE-N using a unigram (ROUGE-1) and a bigram (ROUGE-2), as well as ROUGE-L. For each of these metrics, we receive the F1 score f, precision p, and recall r.

Various services can be used for text translation. DeepL is a neural machine translation service developed by DeepL GmbH, a German-based company. It uses artificial intelligence to translate text from one language to another and is known for producing translations of high quality. The service is available online and can be accessed through the DeepL website, or via an API for integration into other applications. In August 2021, DeepL released a Python client library for the DeepL API. It makes it easier for developers working with Python to build applications with DeepL.

To avoid bias in sampling, the data is selected randomly using the sample() function from the Python library "random". The sample size is set to 50, which means that each record in the source dataset has a probability of 0.05% of being selected. The sample size must be limited, due to the restrictions of the DeepL API Free account.

The ROUGE-1 F1 score, which is described in further detail in Section 2.3, is used as the evaluation measure. This measure takes precision, as well as recall, into account. The Python program iterates over each element selected in the sample, computes the summaries, and calculates the ROUGE metric, based on the reference summary.

5. Experiments and Evaluation Results

This section provides an overview of the setup of our experiments. It describes what data is going to be processed and what methods we use to process the data for further evaluation. Please note that the used methods BERT2BERT and BART were used as off-the-shelf versions provided by Hugging Face, i.e., without any further hyperparameter tuning.

5.1. Dataset

Due to the large number of well-proposed benchmark datasets that have a large capacity for summarized articles in both extractive and abstractive techniques, English is the gold standard for text summarizing. Therefore, it was quite challenging to find a German dataset with a rich corpus. However, we found and utilized in this experiment a dataset from German Wikipedia. This is one of the few datasets accessible for text summarization systems in German. The dataset contains 100,000 Wikipedia articles that were automatically extracted and further curated by removing pictures, titles, tables, and references. Each of these data items is equipped with a reference summary.

5.2. Text Summarization Quality (ROUGE)

In our experiment, we use the N-measure of ROUGE to evaluate the summarized text from the model. The ROUGE score is a sort of assessment measure commonly employed in automated summarization systems. These metrics compare a systematically produced summary to a human-made summary. Specifically in this experiment, we use the ROUGE-1 scores, which are based on the overlap of unigrams between the two summaries. The measure was introduced in Section 2.3.

5.3. Evaluation Results

The result is stored in a comma-separated value (CSV) file on the filesystem of the computer running the Python script. The structure of the file is shown in Table 1. Some example results from our CSV file are shown in Table 2.

Table 1. Structure of the CSV files.

Attribute	Description
Source Text	Contains the text of the full German Wikipedia article.
Reference Summary	Contains the human-produced summary of the German Wikipedia article.
Computed Summary BERT	Contains the summary produced by the German BERT2BERT fine-tuned model.
F1 Score BERT	Contains the ROUGE-1 F1 value of the computed summary.
Computed Summary BART	Contains the German translated summary produced by the BART (large-sized) fine-tuned model.
F1 Score BART	Contains the ROUGE-1 F1 value of the computed summary.

Table 2. Example Extract from a CSV file.

#	Reference Summary	Computed Summary German BERT	F1 Score BERT	Computed Summary BART	F1 Score BART
1	Das Ehrenmal für die Seckbacher Gefallenen der Weltkriege steht innerhalb der Grünanlagen des Lohrparks auf dem Lohrberg in dem zu Frankfurt am Main gehörenden Stadtteil Seckbach.	Ein Krieger–Ehrenmal die Seckbacher, die der deutschen Einigungskriege 1864, 1866/71/71 gefallen sind, steht vor dem Kirchhof der Marienkirche.	0.1168	Vor dem Kirchhof der Marienkirche steht ein Kriegerdenkmal für die in den deutschen Einigungskriegen 1864, 1866 und 1870/71 gefallenen Seckbacher. Das Denkmal ist Teil der Anlage des 1924 angelegten Lohrer Parks und wird von einem 5 Meter hohen Kreuz dominiert.	0.2055
...	...	...	...	...	...
50	Klaus Draeger ist ein deutscher Ingenieur und Manager. Er war als Mitglied des Vorstands der BMW AG für die Bereiche Einkauf und Lieferantennetzwerk zuständig.	Draeger legte 1975 das Abitur ab und studierte danach von 1975 bis 1985 Maschinenbau an der Karlsruhe. Von 1982 bis 1985 war er Chef der BMW AG.	0.3143	Draeger legte 1975 sein Abitur am Alexander-von-Humboldt-Gymnasium Konstanz ab. Von 1975 bis 1981 studierte er Maschinenbau an der Universität Karlsruhe. Am 1. September 1985 trat er als Trainee in die BMW AG ein und war später in verschiedenen Führungspositionen tätig. Im Jahr 2006 wurde Draeger in den Vorstand des Unternehmens berufen.	0.4663

Based on the resulting CSV, the average ROUGE-1 F1 score can be calculated for both the German BERT and the BART model. For the German BERT model, we get an average score of 0.1820, and for the BART model an average score of 0.1971. Hence, the BART model performs slightly better than the German BERT model. Further statistics on the ROUGE-1 F1 score are provided in Table 3.

Table 3. Statistical figures for the ROUGE-1 F1 scores of the sample data.

Measure	German BERT	BART
Average	0.1820	0.1971
Median	0.1649	0.1952
Standard Deviation	0.0889	0.0607
Variance	0.0079	0.0036
Minimum	0.0000	0.1010
Maximum	0.4444	0.3181

It should also be noted that the text of the translated summaries produced using the BART model is usually longer than the ones produced by the German BERT model. On average, the summaries of the translated BART model have a length of 54 words, while the German BERT model produces summaries with a length of 25 words. The reference summaries have an average length of 44 words.

Based on a human assessment of the computed summaries, some anomalies can be identified. Sometimes the computed summaries of the German BERT model simply consist of copies of the leading sentences of the source articles. Table 4 shows an example of this.

Table 4. Anomaly in computed summary.

#	Source Article	Reference Summary	Computed Summary German BERT
8	Nach dem Schulbesuch studierte Berger zuerst an der Cornell University, schloss dieses Studium 1967 mit einem Bachelor of Arts ab und wurde auch Mitglied der Ehemaligenvereinigung “Quill and Dagger”. Ein anschliessendes postgraduales Studium der Rechtswissenschaften an der Law School der Harvard University beendete er 1971 mit einem Juris Doctor cum laude. Im Anschluss wurde er Redenschreiber von George McGovern während dessen Kandidatur für die Demokratische Partei bei der Praäsidentschaftswahl 1972 und lernte während dieser Zeit auch Bill Clinton kennen. [. . .]	Samuel “Sandy” R. Berger war ein US-amerikanischer Rechtsanwalt, der zwischen 1997 und 2001 während der zweiten Amtszeit von US-Präsident Bill Clinton auch Nationaler Sicherheitsberater der USA war und der 2005 von einem Gericht wegen Entwendung geheimer Dokumente verurteilt wurde.	Nach dem Schulbesuch studierte Berger zuerst an der Cornell University, schloss dieses Studium 1967 mit einem Bachelor of Arts ab und wurde auch Mitglied der Ehemaligenvereinigung “Quill and Dagger“.

Furthermore, the summaries produced with the German BERT model sometimes end abruptly or do not cover the topic discussed in the source articles. One example is the following summary created by the BERT model based on a Wikipedia article that discusses a new car design for stock car auto racings: “Der neue VW–Motor ist der beste Golfer der Welt: Der neue Motor ist auch für die Kunden attraktiver”. The summary does not make any sense at all: a golf player cannot be the best car engine in the world. Moreover, the car manufacturer Volkswagen (VW) is not mentioned in the source article at all. In general, the translated summaries produced by the BART model seem to be much more consistent and comprehensible.

6. Discussion

In this section, we look at the consequences of employing BERT and BART for summarizing text, and further evaluate the evaluation metric ROUGE. The section finishes with thoughts on future study.

6.1. BERT and BART Summarization

We faced some limitations in finding a multilingual model at Hugging Face that had been fine-tuned in summarization tasks that were case sensitive, which could have altered the meaning of multiple terms as well as the ROUGE assessment findings. As a result of this limitation, we determined that it was preferable to compare two monolingual models, the German monolingual BERT and English monolingual BART.

According to our results, the monolingual English model BART outperforms the monolingual German model BERT in abstractive German text summarizing, despite the fact that it requires an additional step to translate the content from German to English in order to summarize the text using BART and then translate it back to German, which may be seen as a handicap. However, there is a chance that DeepL improves the accuracy of the text while translating the English summary from BART into German. It is also worth noting that the English monolingual BART was pre-trained with far more and with higher-quality data than the German monolingual BERT, which makes it more challenging for BERT to create a high-quality text summary, particularly regarding abstractive summarization.

6.2. ROUGE Metric Evaluation

According to our experiment, ROUGE is a suitable assessment metric; however, it has several limitations. ROUGE, in particular, cannot account for various words with the same meaning, since it assesses syntactical matching rather than semantics. As a result, if two sequences had the same meaning but used different words to convey it, they may be awarded a low ROUGE score. It has also been remarked that ROUGE metrics have particular limitations for agglutinative languages [37], which is, however, not the case for the German language.

This reveals a flaw in the ROUGE criteria presently used to characterize the state of the art for summarizing, which rely completely on reference summaries to assess the quality of generated summaries.

7. Conclusions

The aim of our research project was to identify how the abstractive text summarization works for the German language using multilingual models such as BERT, to compare the resulting summaries with those from other models such as BART, and then use this to assess the quality of a summarized text. There seems to be no further research being carried out on the comparison of results from using multilingual language models for German text summarization and monolingual English language models combined with machine translation. On the other hand, most of the research on text summarization deals with news articles, and it is therefore not so clear how well text summarization works for other types of text.

For this reason we have focused on the German BERT2BERT model and the BART model. The German BERT2BERT model is based on the German BERT Model and was fine-tuned on the MLSUM DE dataset for summarization. The German BERT Base Model was trained on German Wikipedia, OpenLegalData and news articles. The BART model was pre-trained on the English language and fine-tuned on CNN Daily Mail articles. BART is another model particularly effective for text generation, but also works well for comprehension tasks. Most of the models used for text summarization have been trained predominantly on English text data.

On the one hand, machine translation (MT) can be used to convert one language into another. Multilingual neural machine translation (NMT) enables the training of a single model that supports translation from multiple source languages into multiple targets.

However, there are some drawbacks with using NMT. One could be quality issues, due to translation errors. Another approach is to take a multilingual model and perform tasks such as summarization with it. Soon after the development of BERT, Google research introduced a multilingual version of BERT (also referred to as mBERT), capable of working with more than a hundred languages. The experimental results show that the monolingual BART model would be a better approach when compared to the German BERT model for abstractive text summarization using a large dataset such as German Wikipedia. With the help of the ROUGE-1 metric and a human assessment, we have found that the BART model, in combination with the translation service, outperforms the German BERT model.

As one of the limitations of our study is the considered dataset, we would suggest further studies with a larger sample size and possibly a wider range of text types. In fact, we plan to utilize results from this research in the context of multilingual recommender technologies (such as for technology recommender systems based on web crawling and summarization), as explored for a pure English language use case in [38]. However, due to the limited availability of suitable data, the current paper focused on using Wikipedia articles for the experiments. It would also be interesting to better explore text summarization capabilities for other non-English languages.

There are some further exciting potentials for future research, based on the findings of this study. Due to several weaknesses discovered in the ROUGE measure, it would be highly significant to construct assessment criteria that encompass more than simply the reference description, providing a fascinating avenue for future research.

Emphasizing the significance of high-quality pre-training data for increasingly complicated language creation tasks such as abstractive summarization, we propose that future summarizing research in German concentrate on the creation of a better pre-trained German BERT model in order to enhance outcomes, particularly for abstractive summarization.

Author Contributions: Methodology, software, validation, investigation, writing—original draft preparation, T.H.M.A., D.K. and R.R.; conceptualization, supervision, writing—review and editing, T.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: We have used a dataset from German Wikipedia made available via https://www.swisstext.org/2019/shared-task/german-text-summarization-challenge.html.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Patel, A.; Siddiqui, T.J.; Tiwary, U.S. A language independent approach to multilingual text summarization. In Proceedings of the Conference RIAO 2007, Pittsburgh, PA, USA, 30 May–1 June 2007.
2. Parida, S.; Motlicek, P. Abstract Text Summarization: A Low Resource Challenge. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 5993–5997. [CrossRef]
3. Bornea, M.; Pan, L.; Rosenthal, S.; Florian, R.; Sil, A. Multilingual transfer learning for QA using translation as data augmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 2–9 February 2021; Volume 35, pp. 12583–12591.
4. Widyassari, A.P.; Rustad, S.; Shidik, G.F.; Noersasongko, E.; Syukur, A.; Affandy, A.; Setiadi, D.R.I.M. Review of automatic text summarization techniques & methods. *J. King Saud Univ.—Comput. Inf. Sci.* **2022**, *34*, 1029–1046. [CrossRef]
5. Moratanch, N.; Chitrakala, S. A survey on abstractive text summarization. In Proceedings of the 2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT), Nagercoil, India, 18–19 March 2016; pp. 1–7. [CrossRef]
6. Wang, S.; Zhao, X.; Li, B.; Ge, B.; Tang, D. Integrating extractive and abstractive models for long text summarization. In Proceedings of the 2017 IEEE International Congress on Big Data (BigData Congress), Honolulu, HI, USA, 25–30 June 2017; IEEE: Toulouse, France; pp. 305–312.
7. Mahajani, A.; Pandya, V.; Maria, I.; Sharma, D. A comprehensive survey on extractive and abstractive techniques for text summarization. In *Ambient Communications and Computer Systems: RACCCS-2018*; Springer: Singapore, 2019; pp. 339–351.
8. Kanapala, A.; Pal, S.; Pamula, R. Text summarization from legal documents: A survey. *Artif. Intell. Rev.* **2019**, *51*, 371–402. [CrossRef]

9. Prudhvi, K.; Bharath Chowdary, A.; Subba Rami Reddy, P.; Lakshmi Prasanna, P. Text summarization using natural language processing. In *Intelligent System Design—Proceedings of Intelligent System Design: INDIA 2019*; Springer: Singapore, 2020; pp. 535–547.
10. El-Kassas, W.S.; Salama, C.R.; Rafea, A.A.; Mohamed, H.K. Automatic text summarization: A comprehensive survey. *Expert Syst. Appl.* **2021**, *165*, 113679. [CrossRef]
11. Lin, H.; Ng, V. Abstractive summarization: A survey of the state of the art. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 9815–9822.
12. Suleiman, D.; Awajan, A. Deep learning based abstractive text summarization: Approaches, datasets, evaluation measures, and challenges. *Math. Probl. Eng.* **2020**, *2020*, 9365340. [CrossRef]
13. Shi, T.; Keneshloo, Y.; Ramakrishnan, N.; Reddy, C.K. Neural abstractive text summarization with sequence-to-sequence models. *ACM Trans. Data Sci.* **2021**, *2*, 1–37. [CrossRef]
14. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2019**, arXiv:1810.04805. [CrossRef]
15. Liu, B. *Sentiment Analysis and Opinion Mining*; Springer: Cham, Switzerland, 2022.
16. Peters, M.E.; Neumann, M.; Zettlemoyer, L.; Yih, W. Dissecting Contextual Word Embeddings: Architecture and Representation. *arXiv* **2018**, arXiv:1808.08949. [CrossRef]
17. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving language understanding by generative pre-training. Preprint. 2018. Available online: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf (accessed on 28 March 2023).
18. Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; Zettlemoyer, L. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *arXiv* **2019**, arXiv:1910.13461. [CrossRef]
19. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is All you Need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*. Available online: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html (accessed on 16 January 2023).
20. Aharoni, R.; Johnson, M.; Firat, O. Massively Multilingual Neural Machine Translation. *arXiv* **2019**, arXiv:1903.00089. [CrossRef]
21. GitHub, 2022. Available online: https://github.com/google-research/bert/ (accessed on 16 January 2023).
22. Scheible, R.; Thomczyk, F.; Tippmann, P.; Jaravine, V.; Boeker, M. GottBERT: A pure German Language Model. *arXiv* **2020**, arXiv:2012.02110. [CrossRef]
23. Chan, B.; Schweter, S.; Möller, T. German's Next Language Model. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain, 8–13 December 2020; pp. 6788–6796. [CrossRef]
24. Scialom, T.; Dray, P.-A.; Lamprier, S.; Piwowarski, B.; Staiano, J. MLSUM: The Multilingual Summarization Corpus. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Virtual, 16–20 November 2020; pp. 8051–8067. [CrossRef]
25. Lin, C.-Y. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2004; pp. 74–81. Available online: https://aclanthology.org/W04-1013 (accessed on 16 January 2023).
26. Eyal, M.; Baumel, T.; Elhadad, M. Question Answering as an Automatic Evaluation Metric for News Article Summarization. *arXiv* **2019**, arXiv:1906.00318. [CrossRef]
27. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.Q.; Artzi, Y. BERTScore: Evaluating text generation with BERT. *arXiv* **2019**, arXiv:1904.09675.
28. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *arXiv* **2019**, arXiv:1908.10084.
29. Fabbri, A.R.; Kryściński, W.; McCann, B.; Xiong, C.; Socher, R.; Radev, D. Summeval: Re-evaluating summarization evaluation. *Trans. Assoc. Comput. Linguist.* **2021**, *9*, 391–409. [CrossRef]
30. Park, J.; Song, C.; Han, J. A study of evaluation metrics and datasets for video captioning. In Proceedings of the 2017 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS), Okinawa, Japan, 24–26 November 2017; pp. 172–175. [CrossRef]
31. Giannakopoulos, G.; Karkaletsis, V.; Vouros, G.; Stamatopoulos, P. Summarization system evaluation revisited: N-gram graphs. *ACM Trans. Speech Lang. Process.* **2008**, *5*, 5. [CrossRef]
32. Tran, H.N.; Kruschwitz, U. Ur-iw-hnt at CheckThat! 2022: Cross-lingual Text Summarization for Fake News Detection. In Proceedings of the CLEF 2022: Conference and Labs of the Evaluation Forum, Bologna, Italy, 5–8 September 2022; p. 9.
33. Vom Brocke, J.; Hevner, A.; Maedche, A. (Eds.) Design Science Research. In *Cases*; Springer International Publishing: Cham, Switzerland, 2020. [CrossRef]
34. Pythonorg General Python, F.A.Q. Python Documentation. 2023. Available online: https://docs.python.org/3/faq/general.html (accessed on 16 January 2023).
35. Hugging Face. In Wikipedia. 2022. Available online: https://en.wikipedia.org/w/index.php?title=Hugging_Face&oldid=1118178422 (accessed on 16 January 2023).
36. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv* **2020**, arXiv:1910.03771. [CrossRef]

37. Fikri, F.B.; Oflazer, K.; Yanikoglu, B. Semantic similarity based evaluation for abstractive news summarization. In Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021), Bangkok, Thailand, 5–6 August 2021; pp. 24–33.
38. Campos Macias, N.; Düggelin, W.; Ruf, Y.; Hanne, T. Building a Technology Recommender System Using Web Crawling and Natural Language Processing Technology. *Algorithms* **2022**, *15*, 272. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.