Using Pre-trained Language Models for Abstractive DBPEDIA Summarization: A Comparative Study

Hamada M. ZAHERA♣ [a,1] , Fedor VITIUGIN♣ [b] , Mohamed Ahmed SHERIF [a] , Carlos CASTILLO [b,c] and Axel-Cyrille Ngonga NGOMO [a]

[a] *Data Science Group, Department of Computer Science, Paderborn University*
[b] *Web Science and Social Computing Research group, Universitat Pompeu Fabra, Spain*
[c] *ICREA, Catalan Institution for Research and Advanced Studies, Spain*
ORCiD ID: Hamada M. Zahera https://orcid.org/0000-0003-0215-1278, Fedor Vitiugin https://orcid.org/0000-0003-4350-1828, Mohamed Ahmed Sherif https://orcid.org/0000-0002-9927-2203, Carlos Castillo https://orcid.org/0000-0003-4544-0416, Axel-Cyrille Ngonga Ngomo https://orcid.org/0000-0001-7112-3516

Abstract.

Purpose: This study addresses the limitations of current short abstracts of DBPEDIA entities, which often lack a comprehensive overview due to their creating method (i.e., selecting the first two-three sentences from the full DBPEDIA abstracts).

Methodology: We leverage pre-trained language models to generate abstractive summaries of DBPEDIA abstracts in six languages (English, French, German, Italian, Spanish, and Dutch). We performed several experiments to assess the quality of generated summaries by language models. In particular, we evaluated the generated summaries using human judgments and automated metrics (Self-ROUGE and BERTScore). Additionally, we studied the correlation between human judgments and automated metrics in evaluating the generated summaries under different aspects: informativeness, coherence, conciseness, and fluency.

Findings: Pre-trained language models generate summaries more concise and informative than existing short abstracts. Specifically, BART-based models effectively overcome the limitations of DBPEDIA short abstracts, especially for longer ones. Moreover, we show that BERTScore and ROUGE-1 are reliable metrics for assessing the informativeness and coherence of the generated summaries with respect to the full DBPEDIA abstracts. We also find a negative correlation between conciseness and human ratings. Furthermore, fluency evaluation remains challenging without human judgment.

Value: This study has significant implications for various applications in machine learning and natural language processing that rely on DBPEDIA resources. By providing succinct and comprehensive summaries, our approach enhances the quality of DBPEDIA abstracts and contributes to the semantic web community.

Keywords., Abstractive Summarization, Large Language Models, Knowledge Graphs.

♣Equal Contribution

[1]Corresponding Author: Hamada M. Zahera; E-mail: hamada.zahera@uni-paderborn.de

Figure 1. An example of shortened abstract of "Marie Curie" entity in DBPEDIA.

1. Introduction

DBPEDIA is one of the most popular knowledge graphs in the Linked Open Data cloud (LOD) [1]. DBPEDIA has been widely used as a significant resource for accessing and linking knowledge on the web, particularly in the context of the semantic web and linked data. Entity abstracts (`dbo:abstract`) are an essential component of DBPEDIA, as they provide a concise summary of the Wikipedia page for each entity. Moreover, there are two types of DBPEDIA abstracts: 1) *Full abstracts*, which are extracted from first paragraphs of the corresponding WIKIPEDIA article for each entity. 2) *Short abstracts*[2] are automatically created by selecting the first few sentences (i.e., two-three sentences) from the full abstracts [2]. Short abstracts are used to provide users with a comprehensive overview of the most significant information about entities. For example, *Google* employs short abstracts of search concepts in the knowledge panel to offer users a concise summary of the searched entities [3]. However, the method of creating these short abstracts omits other relevant information in the remaining portion of the full abstract. Figure 1 shows an example of the shortened abstract of *"Marie Curie"* entity that is created by truncating[3] the first sentences from its full abstract[4]. This shortened abstract ignores other essential information such as *"The cause of her death was given as aplastic pernicious anaemia, a condition she developed after years of exposure to radiation through her work"*, which is relevant for understanding *Marie Curie's* life and achievements. It is important to note that some short abstracts of DBPEDIA are unavailable in specific languages.

To address these challenges, we leverage pre-trained language models (LLMs) to generate abstractive summaries of DBPEDIA entities. Recently advances in pre-trained language models have led to impressive performance in text summarization tasks, achieving state-of-the-art performance on various benchmark datasets [4–7]. Inspired by this success, we employ two state-of-the-art LLMs in our comparative study: i) BART (short for *Bidirectional Auto-Regressive transformers*) model, which can generate more accurate and coherent summaries by considering the context of a text in both directions (left-to-right and right-to-left) [8], ii) T5 (short for *Text-To-Text Transfer Transformer*) model is based on a transformer architecture with a self-attention mechanism that uses a text-to-text approach, i.e., the T5 model is trained to generate an output text based on an

[2] https://databus.dbpedia.org/dbpedia/text/short-abstracts/
[3] full text of last sentence is "*the first woman to become a professor at the University of Paris*"
[4] https://en.wikipedia.org/wiki/Marie_Curie

input. This allows the T5 model to be used across various tasks (e.g., text summarization, question answering, machine translation). To ensure the accessibility and affordability of our summarization approach, we chose these open-source models (BART and T5) over commercial models (e.g., GPT-3, GPT-4) which require API subscriptions (e.g., OpenAI API) or large computational resources. Moreover, open-source models offer high adaptability and can be readily fine-tuned on domain-specific datasets with minimal effort. Furthermore, previous studies have demonstrated that both BART and T5 can generate summaries of comparable quality to those produced by smaller GPT-3 models [9–11].

We performed several experiments to identify the most suitable pre-trained LLM for generating abstractive summaries of DBPEDIA abstracts in six languages. We used DBPEDIA abstracts in *English, German, French, Italian, Spanish*, and *Dutch* as our evaluation dataset and produced summaries using various LLMs. We then evaluated the quality of the LLMs-generated summaries against the existing shortened abstracts using both human judgments and automated metrics. Furthermore, we investigated the correlation between the automated metrics and human assessments of the summaries' quality. Our evaluation results indicated that LLMs are effective tools for creating informative summaries for DBPEDIA abstracts. However, the choice of LLMs should be adapted to the specific language. We summarize the main contributions of our study as follows:

- To the best of our knowledge, this is the first study to leverage LLMs to generate abstractive summaries of DBPEDIA abstracts compared to the existing method that automatically selects the first few sentences from the full abstracts.
- We compared the performance of different LLMs for generating abstractive summaries in six languages (English, German, French, Italian, Spanish, and Dutch) using human and automated evaluation metrics
- We analyzed the correlation between the automated metrics (BERTScore and self-ROUGE) and the human judgments of the quality of generated summaries.
- We provide a resource of abstractive summaries of all DBPEDIA abstracts (v2022) in English and German.[5]

2. Related Works

LLMs for abstractive summarization. Recent years have witnessed a growing interest in summarizing descriptions of real-world entities in knowledge graphs [12, 13]. This task, known as text summarization, requires selecting the most essential and salient concepts, entities, and relationships from the knowledge graph, and generating a brief and coherent summary of them. Text summarization can generally be divided into two categories: i) *extractive summarization* [14], which involves selecting the most salient and informative sentences from a document to create a summary, and ii) *abstractive summarization* [15], which involves generating a new summary that conveys the main ideas of the original document, potentially using new phrases and sentences that were not present in the original text. Our study focuses on the latter for generating abstractive summaries of DBPEDIA abstracts.

Abstractive summarization is a text-generation process that aims to produce summaries that are fluent and coherent, as well as informative and concise. Previous works

[5] https://zenodo.org/record/7600894

have employed deep neural networks and language generation techniques to achieve this goal, often using a sequence-to-sequence (Seq2Seq) architecture with an attention mechanism or transformers. These methods can generate summaries that are more expressive and natural than extractive summaries, which simply select sentences from the original document. For example, See et al. [16] proposed the pointer-generator network, which combines the ability to generate new words with the ability to copy words from the input text. This hybrid approach allows for the generation of more fluent and accurate summaries as demonstrated by the evaluation results on the *CNN/Daily Mail* dataset, where it outperformed several baselines. Another example is the fine-tuning of pre-trained language models on large-scale summarization datasets, which can lead to substantial improvements in abstractive summarization and generate higher-quality summaries [17]. Pre-trained language models such as T5, BART, and GPT-2 have also achieved outstanding performance in generating high-quality summaries in terms of relevance, fluency, and semantic accuracy [18] Motivated by this success, we propose our approach for employing pre-trained language models to produce abstractive summaries of DBPEDIA abstracts. To the best of our knowledge, this is the first study to apply language models to this task. The existing method for creating summaries of DBPEDIA abstracts (i.e., short abstracts) simply selects the first few sentences from each entity's description.

Evaluating LLM-generated summaries. Evaluating the quality of generated summaries by large language models is a challenging task [19]. One approach is to use manual evaluation, where human experts are asked to grade the summaries based on their understanding and perception of the content [20]. For example, Iskender et al. [21] compared crowdsourcing ratings with expert ratings and automatic metrics such as ROUGE, BLEU, or BERTScore on a German summarization dataset. They found that crowdsourcing can be used as a direct substitute for experts when measuring structure and coherence, but should be considered carefully when judging overall quality, grammaticality, clarity, and summary informativeness. On the other hand, researchers have proposed self-evaluation methods such as BERTScore [22] and Self-ROUGE [23, 24] that compare the quality of generated summaries with respect to the original text. Specifically, the BERTScore metric measures the semantic similarity between a generated summary and its corresponding original text using cosine distance between their contextualized BERT embeddings [22]. For instance, Koroteev [25] demonstrated the use of semantic text-similarity metrics for evaluating the quality of abstractive summaries in Russian. The author argues that semantic text-similarity metrics are a valuable tool for a variety of natural language processing (NLP) tasks, such as machine translation, information retrieval, and text summarization. Due to the lack of gold-standard summaries for DBPEDIA abstracts, we follow the evaluation methods used by previous works [22, 23, 26] that employed BERTScore and Self-ROUGE as well as crowdsourcing evaluation to assess the quality of the generated summaries in our experiments. We provide more details about these evaluation metrics in Section 4.3.

3. Approach

This section explains the preprocessing steps for the input data (DBPEDIA abstracts), followed by the description of the pre-trained models used in our study. Figure 2 depicts the complete pipeline of our approach, which generates abstractive summaries for DBPEDIA abstracts.

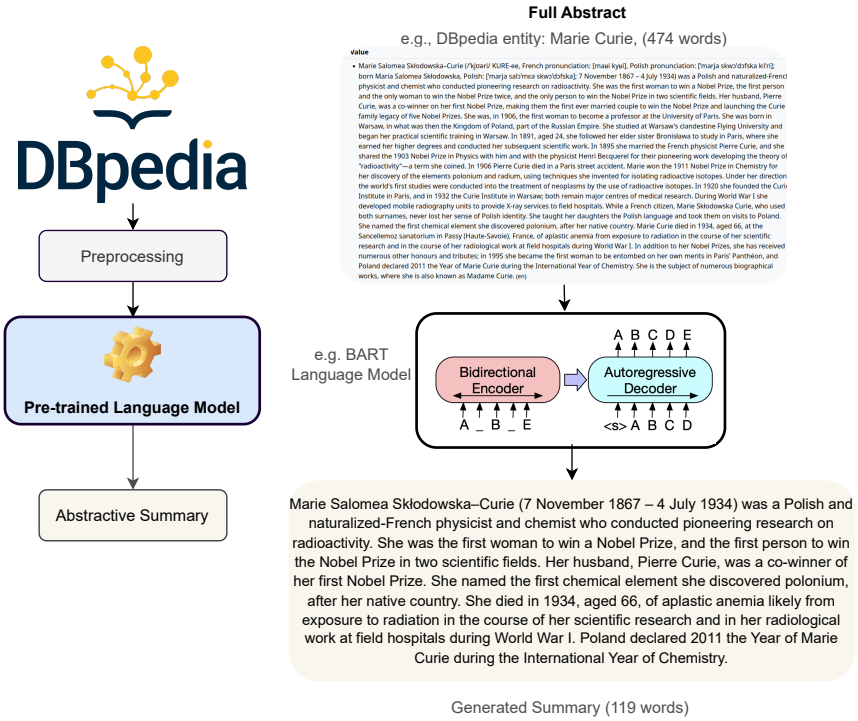


Figure 2. The pipeline of abstractive summarization of DBPEDIA using language models.

3.1. Preprocessing

We note that advanced language models such as BART and T5 are pre-trained on large-scale text corpora and can handle variations in capitalization, stopwords, and word forms [27]. Thus, we do not need to preprocess the text with lowercase, stopword removal, and stemming or lemmatization before applying these models for text summarization. However, we need to format the input text according to the specific requirements of the language models [8, 28]

- *Tokenization:* Tokenization is the process of breaking down text into smaller units, called tokens, that can be characters, subwords, or words. Language models require input text to be tokenized using their own tokenizers, which handle punctuation and special characters appropriately as well as maintain compatibility with the model's preprocessing requirements.
- *Truncating and Padding*: To ensure a uniform length of input sequences for language models, input text that is longer or shorter than a predefined maximum length needs to be padded or truncated. The padding process involves appending special tokens, such as $\langle pad \rangle$, to the end of shorter sequences, while truncation requires removing excess tokens from longer sequences.
- *Formatting*: Language models require specific input formatting to distinguish between different tasks. For a text summarization task, a task prompt (e.g., "summarize") should be used to indicate the desired output.

- *Handling Special Tokens*: Language models use a set of unique tokens, like $\langle eos \rangle$, $\langle bos \rangle$, $\langle unk \rangle$, and $\langle pad \rangle$ to indicate the start/end of a sentence, unknown words, and padding, respectively. It is essential to incorporate these tokens into the input text during preprocessing to ensure proper functioning.
- *Post-processing*: After generating summaries, it may be necessary to conduct post-processing steps to improve the readability and coherence of the output. These steps may include removing redundant or irrelevant tokens, reassembling the sentence structure, and applying appropriate capitalization and punctuation.

3.2. Pre-trained Language Models for Abstractive DBPEDIA *Summarization*

With the advent of pre-trained language models, the field of NLP has been revolutionized, resulting in significant improvements in various tasks, including abstractive summarization [29]. BART [8] and T5 [28] are among the state-of-the-art models for abstractive text summarization. We summarize each mode as follows:

- BART model is a denoising autoencoder that employs a bidirectional encoder and a left-to-right decoder. This model is pre-trained on a large-scale corpus by reconstructing the original text after being corrupted by various noise functions, such as token masking and sentence permutation. This pre-training strategy enables BART to learn a rich latent space representation of the input text, which is useful for generating coherent and contextually relevant summaries. Moreover, BART has exhibited strong performance in abstractive summarization tasks, outperforming previous state-of-the-art models on the benchmark summarization CNN/Daily Mail and XSum datasets [30].
- T5 model is another powerful language model based on the transformer architecture. It is designed with a unified text-to-text framework, which allows fine-tuning on different NLP tasks by simply converting them into text-to-text problems. Additionally, T5's pre-training objective, which involves reconstructing corrupted input text, enables it to learn rich representations that can be leveraged for generating abstractive summaries [31].

4. Evaluation

We conducted our experiments to answer the following research questions:

Q_1: Which LLM is suitable for generating summaries of DBPEDIA abstracts in which language, based on human evaluation and automated similarity metrics?

Q_2: What is the correlation between human ratings and automated metrics in evaluating the informativeness, coherence, conciseness, and fluency of the generated summaries?

4.1. Evaluation Dataset

Our goal is to evaluate the performance of pre-trained large language models in summarizing DBPEDIA abstracts. For this purpose, we created a dataset of 600 DBPEDIA abstracts in six languages (English, German, French, Spanish, Dutch, and Italian), with

100 abstracts randomly selected for each language. We selected the target languages based on the availability of *Short abstracts* dataset except for Japanese due to its special tokenization process. Table 1 provides a statistical overview including *the number of abstracts* in each language and *the average number of sentences*.

4.2. Models

We employed four different models in our study: three variants of the BART model ($\text{BART}_{\text{large-50}}$, $\text{BART}_{\text{large-CNN}}$, and $\text{BART}_{\text{weak-sup}}$) and the pre-trained T5_{large} model. We provide a brief description of each baseline as follows:

- $\text{BART}_{\text{large-50}}$ is a multilingual model with 139*M* parameters, 12 layers, and a hidden size of 768 and supported 50 languages [32].
- $\text{BART}_{\text{large-CNN}}$ is a large-scale variant of BART model with 400*M* parameters, 12 encoder, and decoder layers. Furthermore, the model was fine-tuned on a collection of news articles and their golden-standard summaries from *CNN/DailyMail* dataset [33].
- $\text{BART}_{\text{weak-sup}}$ is a weakly-supervised BART model [34], which is fine-tuned via incorporating rich external knowledge from CONCEPTNET [35].
- T5_{LARGE} [28] is a pre-trained text-to-text transformer model that can generate text for different NLP tasks. It has 770*M* parameters and is trained on a large corpus of web texts using a masked language modelling objective.

4.3. Evaluation Metrics

Automated Evaluation. To evaluate the quality of LLMs-generated summaries with respect to the full DBPEDIA abstracts, we employ the following metrics:

- *Self-ROUGE* is a self-evaluation metric that measures the similarity between the generated summaries and the original text by computing their *n*-gram overlaps [36]. Due to the lack of gold-standard summaries for DBPEDIA abstracts, we employ Self-ROUGE to extract *n*-grams tokens from both the generated summaries and the full DBPEDIA abstracts and calculate the Precision, Recall, and F_1 scores based on the *n*-grams overlaps (ROUGE metric). Following previous works [23, 26, 37], we selected the top-3 sentences with the highest ROUGE scores (i.e., the ROUGE scores of each sentence when using the rest of the sentences as the reference summary) as the reference text (*silver-standard summaries*) in a greedy manner.
- *BERTScore* [22] measures the similarity between the generated text and the reference text using contextualized embeddings from the pre-trained BERT model. In our study, we employ the full DBPEDIA abstract as a reference text, since there are no golden summaries for the DBPEDIA abstracts. Moreover, we obtain the embedding vector for each token in LLMs-generated summaries ($x = x_1, x_2, \cdots, x_{|x|}$) and DBPEDIA full abstracts ($y = y_1, y_2, \cdots, y_{|y|}$) from the pre-trained BERT model. Each token $x_i \in x$ is aligned to the most similar token in $y_i \in y$ and vice-versa. To achieve this, we compute the pairwise cosine similarity between each token in the generated summary ($x_i \in x$) and each in its corresponding in the full abstract ($y_i \in y$). The cosine similarity is defined as $cos(x_i, y_j) = \frac{x_i^T \cdot y_j}{||x|| \cdot ||y||}$. In LLMs, the embeddings are typically normalized to a unit vector, i.e. $||x||$ and $||y||$ are 1, therefore this

Table 1. The statistics of evaluation dataset

	English	**Spanish**	**German**	**French**	**Italian**	**Dutch**
Number of abstracts	100	100	100	100	100	100
Average number of sentences	6.5	4.98	5.6	3.4	3.17	6.3

computation is simplified to $x_i^T \cdot y_j$. Furthermore, Precision (P), Recall (R), and F_1 scores are computed based on BERTscores as follows:

$$P_{BERT} = \frac{1}{|x|} \sum_{x_i \in x} \max_{y_j \in y} x_i^T \cdot y_j \tag{1}$$

$$R_{BERT} = \frac{1}{|y|} \sum_{y_j \in y} \max_{x_i \in x} x_i^T \cdot y_j \tag{2}$$

$$F_{1_{BERT}} = 2 \times \frac{R_{BERT} \cdot P_{BERT}}{R_{BERT} + P_{BERT}} \tag{3}$$

Human Evaluation. In the absence of reference summaries, crowdsourcing services have become an effective alternative to easily and quickly recruit users (i.e., crowdworkers) in performing manual evaluations of DBPEDIA abstractive summarization. We used the *SurgeHQ*[6] crowdsourcing platform to conduct our experiments, as illustrated in Figure 3. We bounded the evaluation of generated summaries to crowdworkers who are fluent in the target languages. The evaluation procedure contained two main tasks: i) crowdworkers were instructed to select the most appropriate summary that best summarized the full DBPEDIA abstracts. In particular, they compared the summaries LLMs-generated summaries to the shortened DBPEDIA abstracts and ii) they rated each summary, including the shortened ones, using a 4-point *Likert scale*, according to the following criteria:

- *Informativeness* measures how well a generated summary captures the essential information in the source text. A summary is informative if it accurately represents the main ideas and critical points of the original content.
- *Coherence* relates to the logical flow and organization of the summary, ensuring that the ideas and concepts are clearly presented and connected. A summary is coherent if it is easily comprehensible and maintains a well-structured narrative that conveys the main points without confusion.
- *Conciseness* assesses how effectively a summary expresses the essential information from the source text in a clear and succinct manner, without unnecessary repetition or wordiness. A summary is concise if it conveys the key points using the minimum number of words possible, making it an efficient representation of the original content.

[6] https://www.surgehq.ai/

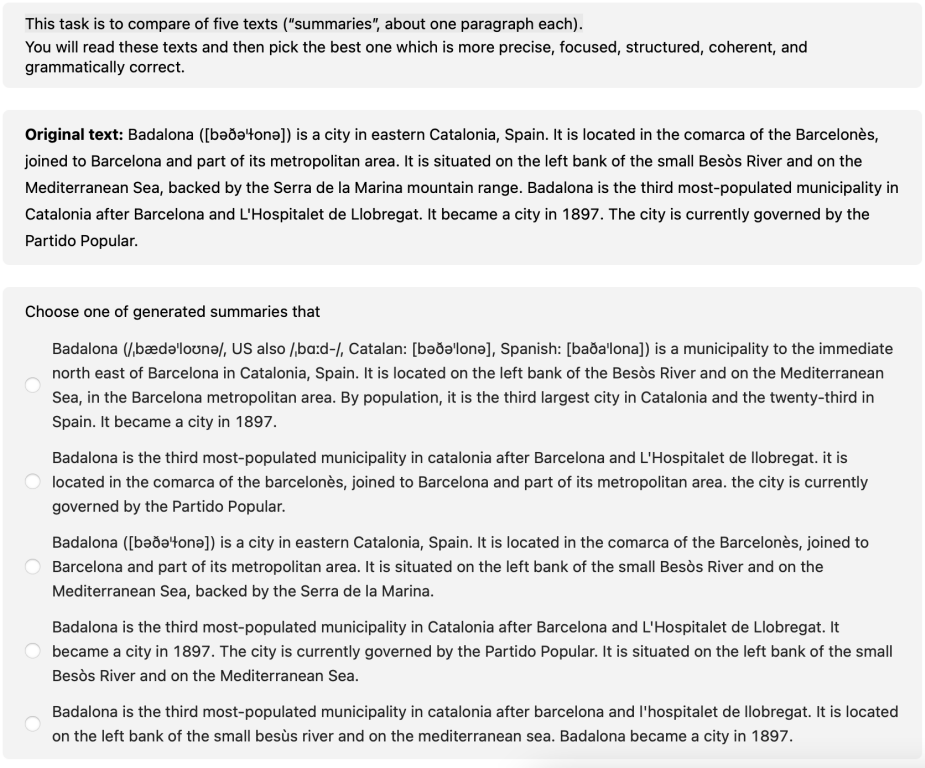


Figure 3. An example of a crowdsourcing task for Barcelona city. Human annotators were asked to select the most informative summary.

- *Fluency* evaluates the naturalness and readability of the generated summary. A summary is fluent if it has smooth and effortless expression, with proper grammar, syntax, and punctuation.

To ensure the reliability of our evaluation, we asked three crowdworkers to assess each summary using these criteria. We then computed the average scores for all the generated summaries.

5. Results

To answer $\mathbf{Q}_1$, we adopted various evaluation metrics to assess the quality of LLMs-generated abstracts. Automated summarization techniques such as Self-ROUGE and BERTScore were used to quantify the models' performance. A human evaluation was also conducted to assess the quality of the summaries generated under different aspects.

5.1. Automated evaluation of LLMs-generated summaries

Self-ROUGE evaluation. Table 2 presents the evaluation results of ROUGE scores for all LLMs-generated summaries and short abstracts. We observe that $\text{BART}_{\text{large-50}}$

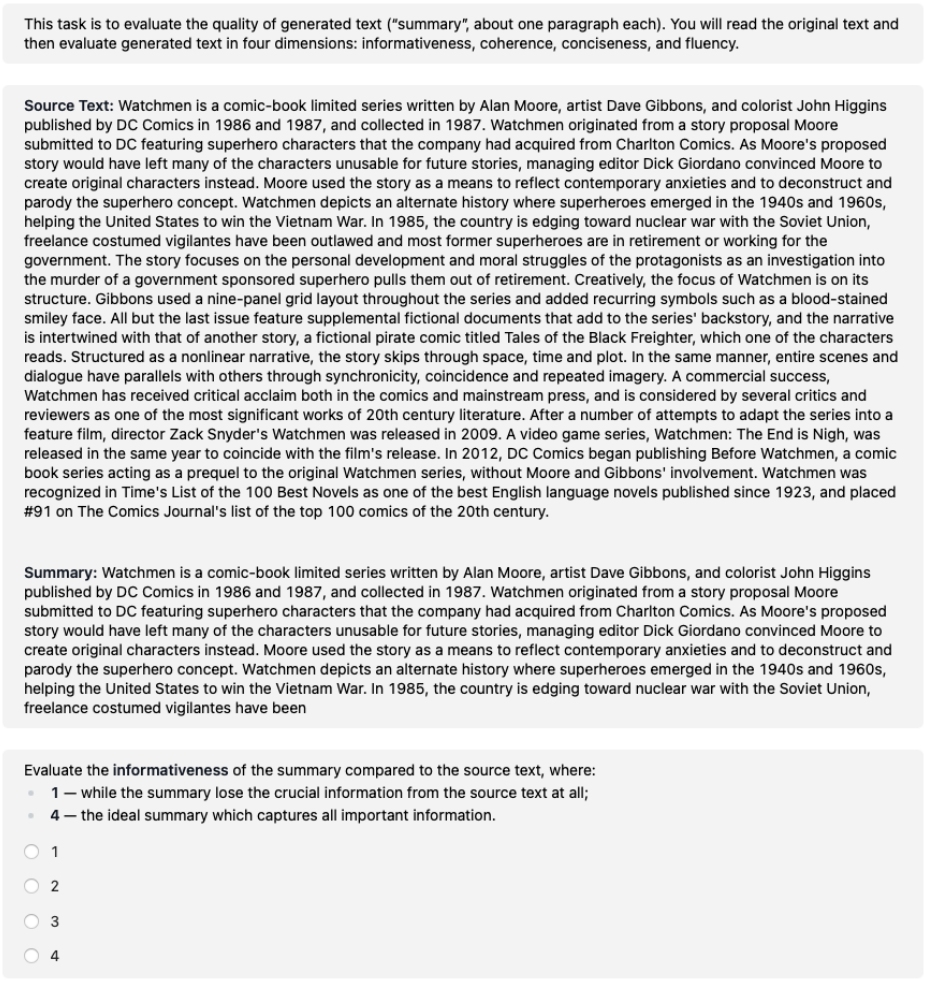


Figure 4. An example of a crowdsourcing task for evaluating the informativeness of generated summary.

generates high-quality summaries for most languages, except for Dutch where the shortened abstracts outperform the LLMs-generated summaries. Using a common threshold of p-value $= 0.05$ for significance testing[7], the results indicate a significant difference in score values (ROUGE-1 p-value ≤ 0.05; ROUGE-2 p-value ≤ 0.06) of $\text{BART}_{\text{large-50}}$ and short abstract.

BERTScore evaluation. Table 3 presents the evaluation results of LLMs-generated summaries and short abstracts using F_{BERT} as computed in Equation (3). Among all models, $\text{BART}_{\text{large-50}}$ achieves the best performance for most languages, indicating its effectiveness in generating high-quality summaries of DBPEDIA abstracts in multiple languages. However, for English, the quality of short abstracts is better by $+3.39\%$.

[7]We tested if BARTlarge-50 has higher score values than short abstract using hypotheses (H0: No difference in score values) (H1: BARTlarge-50 has higher score values)

Table 2. Self-ROUGE evaluation results: *ROUGE-1* (R1), and *ROUGE-2* (R2)

	English		Spanish		German		French		Italian		Dutch	
	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2
Short-abstracts	0.58	**0.62**	0.51	0.45	0.68	0.60	0.70	0.61	0.64	0.54	**0.66**	**0.57**
T5	0.52	0.40	0.57	0.45	0.57	0.45	0.65	0.55	0.67	0.59	0.59	0.49
$\text{BART}_{\text{large-50}}$	**0.61**	0.53	**0.72**	**0.66**	**0.78**	**0.74**	**0.83**	**0.79**	**0.83**	**0.80**	0.63	0.55
$\text{BART}_{\text{large-CNN}}$	**0.61**	0.53	0.58	0.47	0.61	0.51	0.64	0.55	0.67	0.58	0.58	0.45
$\text{BART}_{\text{weak-sup}}$	0.49	0.34	0.30	0.16	0.24	0.10	0.25	0.12	0.26	0.13	0.34	0.20

Table 3. BERTScore (F1) evaluation results

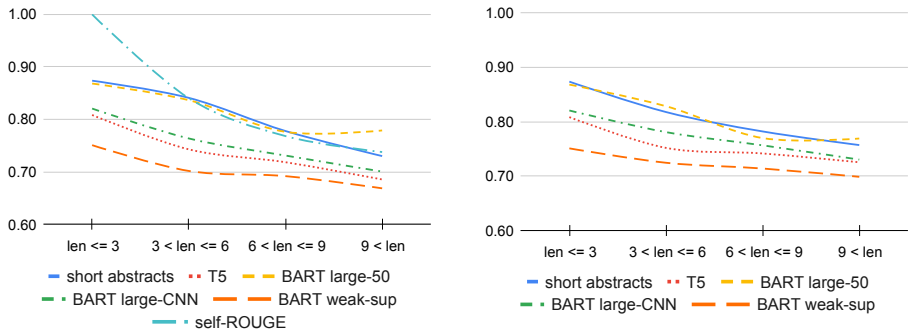
	English	Spanish	German	French	Italian	Dutch
Short-abstract	**0.87**	**0.75**	0.86	0.72	0.86	0.81
T5	0.83	0.68	0.75	0.70	0.81	0.70
$\text{BART}_{\text{large-50}}$	0.84	**0.75**	**0.89**	**0.84**	**0.88**	**0.84**
$\text{BART}_{\text{large-CNN}}$	0.84	0.72	0.76	0.74	0.80	0.73
$\text{BART}_{\text{weak-sup}}$	0.83	0.66	0.67	0.66	0.72	0.68

Therefore, we performed an in-depth analysis based on the number of sentences in each abstract. We grouped the DBPEDIA abstracts used in our experiments into four categories: i) *up to 3 sentences*, same as to the short abstracts consisting of the first three sentences of the original articles (40% of original abstracts), ii) *from 4 to 6 sentences*, which is twice the length of short abstracts (25% of original abstracts), iii) *from 7 to 9 sentences*, which adds three more sentences to the previous group (15% of original abstracts), and iv) *more than 9 sentences*, which forms the final bin (19% of original abstracts). As shown in Figure 5a, $\text{BART}_{\text{large-50}}$ model achieves comparable BERTScores to short abstracts for DBPEDIA abstracts up to 9 sentences and surpasses them for longer abstracts. For other models, we observed that BERTScore decrease as original texts become longer. As shown in Figure 5b BERTScore for short abstracts and summaries generated by $\text{BART}_{\text{large-50}}$ compared to Self-ROUGE summaries are similar. These plots indicate that $\text{BART}_{\text{large-50}}$ summaries achieve higher BERTScore scores than short abstracts, especially for longer texts. Overall, our results conclude that $\text{BART}_{\text{large-50}}$ is an effective resource for generating high-quality summaries of DBPEDIA abstracts depending on their lengths and can help guide future research studies.

5.2. Human Evaluation of LLM-generated summaries of DBPEDIA *abstracts*

We conducted two crowdsourcing experiments to evaluate the generated summaries in six languages: English, Spanish, German, French, Italian, and Dutch.

In the *first experiment*, we presented 100 abstracts per language to native speakers and asked them to choose the most comprehensive summary between a short abstract, or LLMs-generated summaries by $\text{BART}_{\text{large-CNN}}$, $\text{BART}_{\text{large-50}}$, or T5. For each abstract, we used a majority vote of three annotators to select the best summary. Table 4 shows the percentage of summaries chosen by the annotators for each language and model. We observe that 36% of the human annotators preferred the generated summaries by $\text{BART}_{\text{large-CNN}}$, 45% preferred the summaries generated by the T5 model in German, and



(a) BERTScore similarity with original DBPEDIA abstracts

(b) BERTScore similarity with self-ROUGE summaries

Figure 5. BERTScore for abstracts with different sentence lengths in 6 languages.

Table 4. Human evaluation of the LLM-generated summaries in 6 languages. The average rate of annotators' agreement = 0.71

	English	**Spanish**	**German**	**French**	**Italian**	**Dutch**
Short-abstracts	28%	**48%**	32%	**36%**	**46%**	35%
T5	4%	2%	**45%**	25%	12%	9%
$\text{BART}_{\text{large-50}}$	22%	42%	6%	24%	34%	**36%**
$\text{BART}_{\text{large-CNN}}$	**36%**	8%	6%	15%	7%	16%
$\text{BART}_{\text{weak-sup}}$	9%	0%	11%	0%	1%	4%

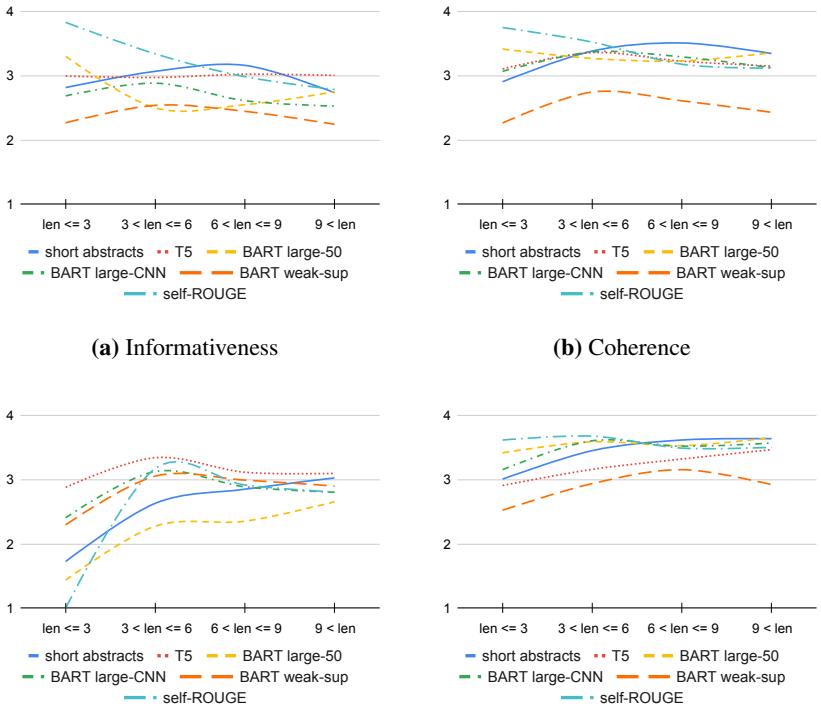
36% selected the $\text{BART}_{\text{large-50}}$-generated summaries in Dutch. For Spanish, French, and Italian languages, the annotators selected short abstracts instead. These results suggest that the length of DBPEDIA abstracts influences human preferences. For shorter abstracts (less than five sentences), human annotators preferred short abstracts. For longer abstracts (more than five sentences), they selected the LLMs-generated summaries. This implies that short abstracts are informative enough in the case of full DBPEDIA abstracts with short content and do not need further summarization. In contrast, longer DBPEDIA abstracts can be summarized efficiently using pre-trained large language models.

In the *second experiment*, we performed another crowdsourcing evaluation to assess the quality of the generated summaries and short abstracts based on four criteria: *informativeness, coherence, conciseness*, and *fluency*. We used a 4-point scale, where 1 is the lowest and 4 is the highest rating. Each summary was compared with the original DBPEDIA abstract by three crowdworkers, following the same procedure as in the first experiment. The evaluation results in Table 5 demonstrate that the T5 model outperforms the other models in terms of informativeness and conciseness, whereas the $\text{BART}_{\text{large-50}}$ model performs better in terms of coherence and fluency.

Finally, we performed an in-depth analysis of the generated summaries and short abstracts based on their length, in the same manner, in Section 5.1. We used the same categorization of DBPEDIA abstracts based on the number of sentences Figure 6. We observe that T5 and $\text{BART}_{\text{large-50}}$ produced more informative and coherent summaries than short abstracts for DBPEDIA summaries with more than 9 sentences. Moreover,

Table 5. Human evaluation of the quality of the LLM-generated summaries (average scores of the English, German, and Dutch languages). The average rate of annotators' agreement = 0.69

Model	Informativeness	Coherence	Conciseness	Fluency
Short-abstract	2.94	3.28	2.55	3.42
T5	**2.99**	3.21	**3.12**	3.21
$BART_{large\text{-}50}$	2.77	**3.32**	2.18	**3.55**
$BART_{large\text{-}CNN}$	2.68	3.21	2.81	3.46
$BART_{weak\text{-}sup}$	2.37	2.51	2.81	2.88



(a) Informativeness **(b)** Coherence

(c) Conciseness **(d)** Fluency

Figure 6. Human evaluation of generated summaries in different criteria for DBPEDIA abstracts (English) with different sentence lengths.

$BART_{large\text{-}50}$ model created more fluent summaries than short abstracts for most categories. Interestingly, T5 model produced more concise summaries than short abstracts, regardless of their length. In summary, the human evaluation indicates that both models $BART_{large\text{-}50}$ and T5 can produce summaries of equivalent quality. In the automated evaluation using Self-ROUGE and BERTScore metrics, the $BART_{large\text{-}50}$ model generated better summaries than other models and short abstracts.

5.3. Automated and Human evaluation results correlation

To answer $\mathbf{Q}_2$, we measured the correlation between the scores of automatic metrics and human judgments using two non-parametric rank correlation coefficients: *Spearman's rank* and *Kendall's rank*. Specifically, *Spearman's rank correlation coefficient*, denoted by Spearman's ρ, assesses the linear association between two variables based on their ranks [38]. Similarly, *Kendall rank correlation coefficient*, denoted by Kendall's τ, evaluates the degree of agreement between two ranked variables [39]. We computed both coefficients for the single document task [40] and plotted them in Figure 7. These measures do not require any assumptions about the distribution of the variables or their joint distribution. Our correlation analysis indicates that BERTScore has the strongest relationship with human ratings of *informativeness*, with Spearman and Kendall coefficients of $\rho \leq 0.61$ and $\tau \leq 0.49$, respectively. Furthermore, ROUGE-1 has the highest correlation with human assessment of *coherence*, with Spearman and Kendall coefficients of $\rho \leq 0.31$ and $\tau \leq 0.25$, respectively. We also observe that *conciseness* has a negative correlation with human evaluation in all cases, with Spearman and Kendall coefficients of $\rho \geq -0.62$ and $\tau \geq -0.52$, while *fluency* has a negligible correlation with values close to 0. Therefore, BERTScore is a recommended measure to assess the *informativeness* of generated summaries, while ROUGE-1 can effectively capture the *coherence* dimension. However, automatic and human scores for *conciseness* were negatively correlated, suggesting a potential direction for exploring this relationship in future work. Additionally, none of the metrics showed a strong correlation with human judgments of *fluency*, implying an open challenge.

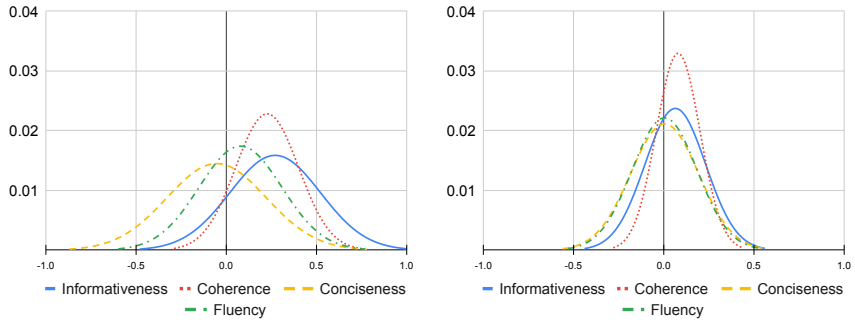
5.4. Supplemental Material Statement.

Our implementation is open source and can be accessed on the GitHub project.[8] We used the `transformer` library v4.25.1 from the `Huggingface hub` to implement our approach. We recommend following the official guideline[9] for setting up and loading the pre-trained language models (BART, $\text{BART}_{\text{large-CNN}}$ and T5).

6. Conclusion

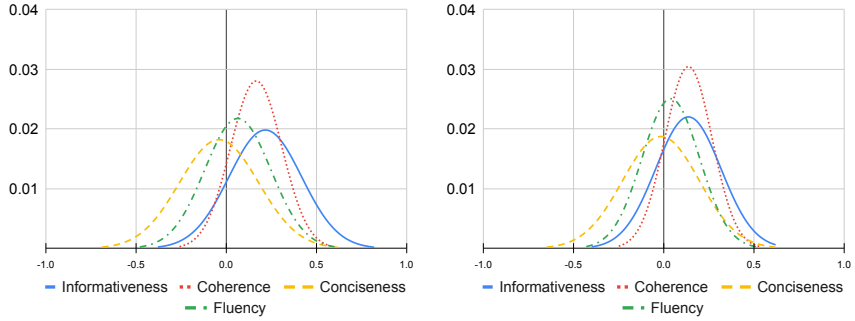
In this study, we explored using different language models for generating abstractive summaries of DBPEDIA abstracts. We observed that the existing shortened abstracts of DBPEDIA, which are obtained by truncating the full abstracts (i.e., selecting the first two-three sentences), may not cover all the relevant information. To overcome this limitation, we propose an abstractive summarization approach based on pre-trained language models such as BART and T5. We conducted various experiments on a multilingual dataset of DBPEDIA abstracts in six languages (English, Spanish, German, French, Italian, and Dutch). We employed automated metrics (Self-ROUGE, BERTScore) and human evaluation to investigate the best model for each language. Our results demonstrate that pre-trained language models can generate informative and concise summaries of DBPEDIA abstracts. However, selecting the most suitable model for each language is

[8] https://github.com/dice-group/DBpedia-Summarizer
[9] https://huggingface.co/docs/transformers/index



(a) Spearman's rank correlation coefficient for BERTScore

(b) Spearman's rank correlation coefficient for ROUGE-1



(c) Kendall's rank correlation coefficient for BERTScore

(d) Kendall's rank correlation coefficient for ROUGE-1

Figure 7. Correlation comparison between human and automated evaluations.

crucial. Furthermore, we found a correlation between automated and human evaluation for assessing *informativeness* with BERTScore and coherence with ROUGE-1. There is also a negative correlation for *conciseness* with human ratings. The evaluation of *fluency* is challenging without human involvement. We plan to investigate larger pre-trained language models in our future studies and fine-tune them on abstractive summarization datasets such as XL-Sum and Wikisum.

Aknowledegment

This work has been supported by the German Federal Ministry of Education and Research (BMBF) through the EuroStars project $E!114154$ PORQUE (grant no $01QE2056C$) and the KIAM project (grant no $02L19C115$). Additionally, this work has been partially supported by: the Department of Research and Universities of the Government of Catalonia ($SGR00930$), the Ministry of Science and Innovation of Spain with the project COMCRISIS (reference code $PID2019-109064GB-I00$), the EU-funded SoBigData++ project under Grant Agreement 871042 and MCIN/AEI /10.13039/501100011033 under the Maria de Maeztu Units of Excellence Programme ($CEX2021-001195-M$).

References

1. Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer, 2007.
2. Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. DBpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2):167–195, 2015.
3. Marius Kavaliauskas and Remigijus Venskutonis. Knowledge graphs in search engines: A systematic literature review. *Informatics*, 7(4):72, 2020.
4. Kaili Sun, Xudong Luo, and Michael Y. Luo. A survey of pretrained language models. In *Knowledge Science, Engineering and Management*, 2022.
5. Junyi Li, Tianyi Tang, Wayne Xin Zhao, and Ji-Rong Wen. Pretrained language model for text generation: A survey. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 4492–4499. ijcai.org, 2021. . URL https://doi.org/10.24963/ijcai.2021/612.
6. Zaid Lawal and Wei Lu. Evaluating the effectiveness of modern pre-trained language models for text summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6479–6490. Association for Computational Linguistics, July 2020. URL https://www.aclweb.org/anthology/2020.acl-main.720.
7. Yaser Keneshloo, Mahdi Namazifar, and Hamed Zamani. Fine-tuning pre-trained transformer models for abstractive text summarization. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, pages 2723–2732. Association for Computing Machinery, October 2020. . URL https://doi.org/10.1145/3340531.3411917.
8. Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.
9. Yichong Xu, Ruochen Xu, Dan Iter, Yang Liu, Shuohang Wang, Chenguang Zhu, and Michael Zeng. Inheritsumm: A general, versatile and compact summarizer by distilling from gpt, 2023.
10. Ayesha Ayub Syed, Ford Lumban Gaol, Alfred Boediman, Tokuro Matsuo, and Widodo Budiharto. A survey of abstractive text summarization utilising pretrained language models. In Ngoc Thanh Nguyen, Tien Khoa Tran, Ualsher Tukayev, Tzung-Pei Hong, Bogdan Trawiński, and Edward Szczerbicki, editors, *Intelligent Information and Database Systems*, pages 532–544, Cham, 2022. Springer International Publishing.
11. Yuntian Zhang, Hanjun Dai, Yiming Li, Zihang Liu, Jianfeng Gao, Jaime Carbonell, Caiming Xiong, and Ying Liu. Bart: Pre-training sequence to sequence models for language generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

12. Yuting Zhang, Ting Liu, Min Zhang, Yantao Chen, and Guilin Gao. Summarizing real-world entities in knowledge graphs: A survey. *Artificial Intelligence Review*, 52 (4):493–535, 2019.
13. Mohamed Elgharib, Jing He, Kukka Tero, Diana Inkpen, and Jimmy Chen. Automatic summarization of knowledge graph entities. In *Proceedings of the International Conference on Web Intelligence*, pages 600–607. ACM, 2017.
14. Avaneesh Kumar Yadav, Ranvijay, Rama Shankar Yadav, and Ashish Kumar Maurya. State-of-the-art approach to extractive text summarization: a comprehensive review. *Multimedia Tools and Applications*, pages 1–63, 2023.
15. Som Gupta and Sanjai Kumar Gupta. Abstractive summarization: An overview of the state of the art. *Expert Systems with Applications*, 121:49–65, 2019.
16. Abigail See, Peter J Liu, and Christopher D Manning. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017.
17. Shashi Narayan, Shay B Cohen, and Mirella Lapata. Fine-tuning pre-trained transformer models for abstractive summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.
18. Apar Garg, Saiteja Adusumilli, Shanmukha Yenneti, Tapas Badal, Deepak Garg, Vivek Pandey, Abhishek Nigam, Yashu Kant Gupta, Gyan Mittal, and Rahul Agarwal. News article summarization with pretrained transformer. In *International Advanced Computing Conference*, pages 203–211. Springer, 2020.
19. Elena Lloret, Laura Plaza, and Ahmet Aker. The challenging task of summary evaluation: an overview. *Language Resources and Evaluation*, 52:101–148, 2018.
20. Elena Lloret, Laura Plaza, and Ahmet Aker. Analyzing the capabilities of crowdsourcing services for text summarization. *Language resources and evaluation*, 47: 337–369, 2013.
21. Neslihan Iskender, Tim Polzehl, and Sebastian Möller. Best practices for crowd-based evaluation of German summarization: Comparing crowd, expert and automatic evaluation. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, Online, November 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.eval4nlp-1.16.
22. Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.
23. Alexander Fabbri, Simeng Han, Haoyuan Li, Haoran Li, Marjan Ghazvininejad, Shafiq Joty, Dragomir Radev, and Yashar Mehdad. Improving zero and few-shot abstractive summarization with intermediate fine-tuning and data augmentation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 704–717, Online, June 2021. Association for Computational Linguistics. . URL https://aclanthology.org/2021.naacl-main.57.
24. Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100, 2018.

25. Mikhail Koroteev. On the usage of semantic text-similarity metrics for natural language processing in russian. In *2020 13th International Conference "Management of large-scale system development" (MLSD)*, pages 1–4, 2020. .
26. Yizhu Liu, Qi Jia, and Kenny Zhu. Length control in abstractive summarization by pretraining information selection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6885–6895, 2022.
27. Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. Recent advances in natural language processing via large pre-trained language models: A survey. *CoRR*, abs/2111.01243, 2021. URL https://arxiv.org/abs/2111.01243.
28. Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020.
29. Haifeng Wang, Jiwei Li, Hua Wu, Eduard Hovy, and Yu Sun. Pre-trained language models and their applications. *Engineering*, 2022. ISSN 2095-8099. . URL https://www.sciencedirect.com/science/article/pii/S2095809922006324.
30. Ayesha Ayub Syed, Ford Lumban Gaol, Alfred Boediman, Tokuro Matsuo, and Widodo Budiharto. A survey of abstractive text summarization utilising pretrained language models. In *Intelligent Information and Database Systems: 14th Asian Conference, ACIIDS 2022, Ho Chi Minh City, Vietnam, November 28–30, 2022, Proceedings, Part I*, pages 532–544. Springer, 2022.
31. GS Ramesh, Vamsi Manyam, Vijoosh Mandula, Pavan Myana, Sathvika Macha, and Suprith Reddy. Abstractive text summarization using t5 architecture. In *Proceedings of Second International Conference on Advances in Computer Engineering and Communication Systems: ICACECS 2021*, pages 535–543. Springer, 2022.
32. Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*, 2020.
33. Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28, 2015.
34. Ankit Sahu and Sriram G Sanjeevi. Better fine-tuning with extracted important sentences for abstractive summarization. In *Proceedings of the International Conference on Communication, Control and Information Sciences (ICCISc)*, volume 1, pages 1–6. IEEE, 2021.
35. Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*, 2017.
36. Chin-Yew Lin and FJ Och. Looking for a few good metrics: Rouge and its evaluation. In *Ntcir workshop*, 2004.
37. Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR, 2020.
38. Jerome L Myers, Arnold D Well, and RF Lorch Jr. Research design and statistical analysis routledge. *New York.[Google Scholar]*, 2010.

39. Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.
40. Chin-Yew Lin and Eduard Hovy. Manual and automatic evaluation of summaries. In *Proceedings of the ACL-02 workshop on automatic summarization*, pages 45–51, 2002.