A Novel Wikipedia based Dataset for Monolingual and Cross-Lingual Summarization

Mehwish Fatima and **Michael Strube**
Heidelberg Institute for Theoretical Studies gGmbH, Heidelberg, Germany
(mehwish.fatima|michael.strube)@h-its.org

Abstract

Cross-lingual summarization is a challenging task for which there are no cross-lingual scientific resources currently available. To overcome the lack of a high-quality resource, we present a new dataset for monolingual and cross-lingual summarization considering the English-German pair. We collect high-quality, real-world cross-lingual data from *Spektrum der Wissenschaft*, which publishes human-written German scientific summaries of English science articles on various subjects. The generated *Spektrum* dataset is small; therefore, we harvest a similar dataset from the *Wikipedia Science Portal* to complement it. The Wikipedia dataset consists of English and German articles, which can be used for monolingual and cross-lingual summarization. Furthermore, we present a quantitative analysis of the datasets and results of empirical experiments with several existing extractive and abstractive summarization models. The results suggest the viability and usefulness of the proposed dataset for monolingual and cross-lingual summarization.

1 Introduction

The summarization research has recently shifted from monolingual summarization (MS) to cross-lingual summarization (CLS) (Ouyang et al., 2019; Duan et al., 2019; Zhu et al., 2019). However, due to the absence of real cross-lingual datasets, recent CLS studies (Shen et al., 2018; Ouyang et al., 2019; Zhu et al., 2019; Pontes et al., 2020) are conducted on existing monolingual news datasets and off-the-shelf machine translation (MT) systems which may introduce noise into pseudo-cross-lingual summarization (PCLS) data. As these CLS studies rely on only news data, the trained summarization models may not work well for other domains such as scientific texts. Although some efforts have been made for investigating the MS task on scientific papers (Vadapalli et al., 2018b; Nikolov et al., 2018; Cohan et al., 2018; Dangovski et al., 2019); however, there is no study on scientific text for CLS to date. Another aspect of consideration is that most CLS studies intend to generate the summaries in English from a local language but not vice versa to facilitate the local readers.

This paper aims to address this issue by developing a summarization dataset containing scientific texts of the English-German language pair from two resources, *Spektrum der Wissenschaft* (SPEKTRUM) and the *Wikipedia Science Portal* (WSP). The paper explores the CLS task by using scientific English documents to generate German summaries for the local readers. To the best of the authors' knowledge, the collected WSP or WIKIPEDIA dataset represents the largest CLS dataset of the English-German pair so far. We believe that the novel WIKIPEDIA dataset encourages new avenues of research in the less explored areas of CLS.

Contributions: This paper has several contributions, including data collection, dataset generation, statistical analysis of the datasets, and an empirical evaluation of MS and CLS. We collect our primary dataset from SPEKTRUM, consisting of 1,510 English science articles with human-written German summaries. In addition, we propose a novel scoring method, which validates the data present in the SPEKTRUM dataset before data extraction. To complement the SPEKTRUM dataset, we harvest our second dataset from WSP, containing 51,312 English and German science articles. The collection of data from two different resources ensures diversity in the written text and topics. It is worth noting that the WIKIPEDIA dataset can also be used for MS, which distinguishes it from existing datasets. We perform a detailed statistical analysis of the dataset that highlights the interesting patterns. Furthermore, we conduct an empirical evaluation with several extractive baselines and existing abstractive summarization models to validate the usability of

Proceedings of the Third Workshop on New Frontiers in Summarization, pages 39–50

our dataset for MS and CLS. Moreover, linguistic quality is evaluated on a subset of the output summaries of the MS and CLS experiments by human judges.

2 Related Work

2.1 Wikipedia Summarization Datasets

Researchers generally believe that WIKIPEDIA is a viable source for data collection, and generating summaries of WIKIPEDIA text is a challenging task.

2.1.1 Monolingual Datasets

WIKIPEDIA has been widely used for the creation of MS datasets such as English multi-document summarization (Zopf et al., 2016; Antognini and Faltings, 2020; Gholipour Ghalandari et al., 2020), English and German single and multi-document summarization (Hättasch et al., 2020), and German single-document summarization (Frefel, 2020). As these datasets are designed for MS, it makes them inadequate for cross-lingual evaluation.

2.1.2 Multilingual Datasets

The TAC MultiLing shared task is held biennially (2011-15) for multilingual multi-document summarization (Giannakopoulos et al., 2011; Giannakopoulos, 2013; Giannakopoulos et al., 2015). These corpora are composed of English Wikinews and translated into 9 languages. The final corpus (MultiLing'15) size is 1500 documents in total for all languages. Ladhak et al. (2020) also create a multilingual dataset named WikiLingua from WikiHow in 18 languages. However, the author conducted experiments to generate English summaries from non-English articles. Although these datasets are multilingual, they are non-scientific, thus cannot be used for cross-lingual summarization of scientific texts. Moreover, the small size of the MultiLing makes it difficult to use for cross-lingual neural models.

2.2 Scientific Summarization Datasets

Kim et al. (2016) build a dataset of introduction-abstract pairs from ARXIV papers for abstractive summarization. Vadapalli et al. (2018b,a) collect a parallel corpus of 87K pairs of research paper titles, abstracts and corresponding blog titles for title generation. Nikolov et al. (2018) create two datasets from scientific articles, abstract-title pairs from MEDLINE for title generation and body-abstract pairs from PUBMED for abstract generation. Cohan et al. (2018) also collect a scientific dataset from ARXIV (194K) and PUBMED (216K) articles for abstractive summarization. Dangovski et al. (2019) create a corpus of 60K science articles from ScienceDaily for summary generation. The datasets mentioned above consist of scientific papers, but all of them were made for MS, which makes them unsuitable for cross-lingual evaluation.

2.3 Cross-lingual Summarization Datasets

Zhang et al. (2016) perform cross-lingual multi-document sentence summarization for the English and Chinese language pair. They use the LDC news dataset and Google translators to get the parallel sentence pairs. Nguyen and Daumé III (2019) collect a dataset from descriptions of news articles from Global Voices in 15 languages. A few researchers work with MT and existing monolingual datasets to achieve the goal of CLS. Pontes et al. (2020) perform cross-lingual multi-sentence compression for the English-French pair. They use the MultiLing'11 dataset for the French language. The dataset is translated with Google translate to generate the English counterpart. Ouyang et al. (2019) propose a Translate-then-Summarize (TRANS-SUM) based CLS model for an inherent monolingual NYT dataset. They use the round-trip translation (RTT) method to convert the English dataset into Somali, Swahili, and Tagalog and then into noisy English. Zhu et al. (2019) also apply on RTT for CLS considering the English-Chinese language pair. There is limited prior work of CLS by using translation corpus during training. Shen et al. (2018); Duan et al. (2019) perform the cross-lingual headline generation and sentence summarization for the English-Chinese language pair. They use the English Gigaword corpus along with the English-Chinese translation corpus for training.

Though these studies focused on CLS, there is no real cross-lingual dataset except for MultiLing and WikiLingua. The datasets are, however, limited in scope and cannot be used for our experiments. To the best of the authors' knowledge, there is no real dataset created with the sole purpose of cross-lingual abstractive summarization of scientific text.

3 Dataset Construction

3.1 Spektrum Data

SPEKTRUM is the German equivalent of the "Scientific American", which began publishing in 1978[1].

[1] Spektrum.de/das-innere-spektrum

The SPEKTRUM magazine is one of the divisions of the Springer Nature publishing group. It is published on a monthly basis and covers many core areas of science, such as archaeology, astronomy, biology, chemistry, *etc*. The SPEKTRUM science journalists present complex English scientific research to non-scientist common readers in a local language (German). SPEKTRUM is therefore viewed as a mediator between scientific publications and the general public.

3.1.1 Data Collection

We have formally contacted and requested SPEKTRUM to release their data for the research purpose. In response to our request, we have meetings with SPEKTRUM's managing director and head of digital production. As a result, we have received a subset of SPEKTRUM data in XML format. The released SPEKTRUM raw data contains German summaries and URLs to their source documents. It consists of 20,556 summaries for the period of December 2000 to February 2019. To process the data, we develop an XML parser that parses the ids, dates, titles, keywords, summaries, and URLs. In some cases, the provided summaries have only one URL associated with them, while in others, there can be multiple URLs. Multiple URLs make detecting and extracting source articles challenging.

Further discussion with SPEKTRUM and manual inspection of a subset of data indicate that there is only one source URL, and the remaining links are for further reading. Before finding the source URLs, all URLs to social media platforms (Facebook, Twitter, YouTube), German websites and non-functional links are filtered out. As a result of filtering, only 5,590 instances are left with functional URLs. Upon further inspection, it is discovered the functional URLs are either PDF or HTML links.

For the instances with PDF links, the extraction of the source article is straightforward. A script is written for PDF URLs to download, extract and parse the text from the articles. We use Beautiful Soup[2] for extraction, and Tika[3] for parsing the text.

For the instances with multiple HTML links, we devise a novel two-step scoring method to find the best-fitting URLs consisting of scientific structure scoring and keyword matching scoring. *Scientific Structure:* This method checks the structure of text for scientific headings - Abstract, Introduction, Results, Discussion, References and Acknowledgements. The score ranges between 0 to 6 by assigning one point for each heading present in the text. The URLs that score four or higher are selected by assuming that a scientific article has at least four of the six headings. To further validate the selected URLs, a keyword matching scoring method is applied. *Keyword Matching:* This method uses the parsed German keywords from raw SPEKTRUM data and the English title of the HTML page to calculate a ratio for matched keywords. The German keywords are translated into English via Google translate[4]. The ratio of matched keywords is defined as the total number of keyword occurrences in page title divided by the total number of German keywords. The URLs with positive scores were selected for the extraction. After the scoring, the HTML pages are downloaded via the module request and extracted with Beautiful Soup[2]. The final extracted instances from PDF and HTML links are 3,554 in total with their German summaries.

3.1.2 Manual Cleaning

After the extraction, the English articles are further manually inspected to filter the incomplete extractions, garbage text, texts other than English, and shorter than the German summary. We manually cleaned the data by two annotators over a period of two weeks. Following manual cleaning, the final data consists of 1,510 English articles and German summaries written by experts in science journalism.

Furthermore, the data is preprocessed for lower case conversion, word and sentence tokenization with NLTK toolkit[5]. The markup tags are used to preserve the structural information on the section and sentence level. The final version of the dataset is stored in JSON format. Unfortunately, this data is insufficient to train the summarization models. Therefore, we decided to collect a similar nature cross-lingual dataset from WIKIPEDIA. Moreover, SPEKTRUM data cannot be published due to the magazine's policies.

3.2 Wikipedia Data

WIKIPEDIA is considered a reliable source for mono- and multi-lingual data acquisition (Antognini and Faltings, 2020; Gholipour Ghalandari et al., 2020; Hättasch et al., 2020; Frefel, 2020).

[2] Pypi/Beautifulsoup
[3] Pypi/Tika
[4] Pypi/Googletrans
[5] Pypi/Nltk

As well as maintaining a consistent format for the articles[6], data is available in several forms for researchers, including data dumps, databases, DBpedia and WIKI-API[7]. These features make WIKIPEDIA an ideal source for cross-lingual summarization data. As a result, we select WSP for scientific cross-lingual data collection. The WSP is a popular, crowd-sourced science encyclopedia available in many languages and is enormous in volume (≈6M articles in English and 2.4M in German). Numerous articles cover various topics such as biology, agriculture, technology, linguistics, and so on.

3.2.1 Data Collection

Figure 1 illustrates the process of collecting monolingual and cross-lingual data from WIKIPEDIA. The Figure 1 shows how English and German articles are connected as well as how they are split to form summaries (lead) and texts. According to WIKIPEDIA's guidelines[6], the lead is the first paragraph of an article that summarizes it. Note that WIKIPEDIA lead is different from a news-style lead or "lede".

WIKI-API[8] is used for data collection which provides an efficient way for extracting an article from a given category, getting existing inter-language links for that article, and extracting sections of that article. Before extraction, the following steps are taken to collect valid data. (i) A list of science subcategories is generated from the main categories. (ii) This list is further processed to generate another list of English and German articles titles. In total, there were 238,766 titles from various science subcategories. (iii) The title list is further checked to find empty titles for both languages. The empty condition is: if the lead is absent, only the lead is present, or only the title is present. Such titles are removed from the list. (iv) Finally, the updated list is used to extract the original articles.

The extracted data is preprocessed to remove the noise and white spaces. The data is converted into lower case and then tokenized for words and sentences with the NLTK toolkit[5]. The markup tags are used to preserve the structural information on the section and sentence level. The final version of the dataset is stored in JSON format. The final dataset is released under the Creative Commons Attribution-ShareAlike 3.0 Unported License for the summarization community[9].

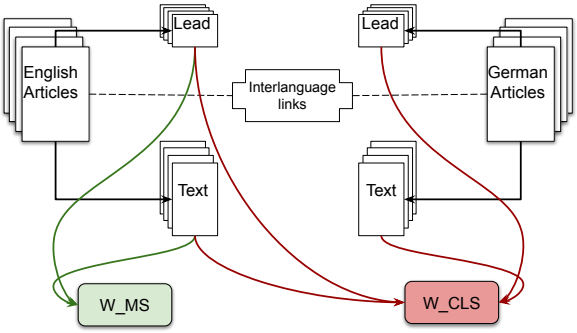

Figure 1: WIKIPEDIA data collection

3.2.2 Manual Verification

The majority of corpus construction studies (Antognini and Faltings, 2020; Ladhak et al., 2020; Frefel, 2020) have omitted the manual verification of collected data due to its complexity. Only Hättasch et al. (2020) performed human verification on a subset of 39 summaries from three different parts of the dataset (Harry potter, English and German Star Wars) for one parameter of interest. For cross-lingual science articles, manual verification poses different challenges such as it requires bilingual comprehension of various scientific topics.

To verify the cross-lingual mappings, we randomly select 20 articles from cross-lingual data. The articles with German summaries are given to two native German speakers (judges) who are also fluent in English. They are asked to evaluate the German summaries based on two different parameters, *i.e.*, (i) relevance and (ii) length. The relevance determines if the German summary is related to the English article, and if not, it is given a score of zero. The length refers to how long or short a summary is. Summaries that are long are given a score of one. Zero is assigned to short summaries (one to two sentences). Considering the length parameter is important because our final objective is to summarize the SPEKTRUM dataset, and we want to have a similar dataset. In terms of relevance, both judges agreed that German summaries are relevant to English articles. For the length, the sample German summaries get an average score of 0.74 with a substantial agreement (Fleiss's $\kappa = 0.76$) between judges. It is worth noting that short summaries (≈25%) make the data challenging as such short summaries are used in extreme summarization (Ca-

[6] Wikipedia/Manual_of_Style
[7] Wikimedia/Research:Data
[8] Pypi/Wikiapi
[9] https://github.com/MehwishFatimah/wsd

	WIKIPEDIA			SPEKTRUM	
	EN-TEXT TRAIN/VAL/TEST	EN-SUM TRAIN/VAL/TEST	DE-SUM TRAIN/VAL/TEST	EN-TEXT TEST	DE-SUM TEST
Split	80/10/10	80/10/10	80/10/10	–	–
Total vocabulary	22M/2.7M/2.7M	3.4M/.4M/.4M	2.9M/.3M/.3M	1M	.4M
Total words	64M/8M/7.9M	5.7M/.7M/.7M	4M/.5M/.5M	4.3M	.6M
Avg. words/doc	1572/1562/1542	139/140/140	100/101/101	2337	361
Standard deviation	1961/1935/1906	110/114/112	92/109/124	1510	250
Total sentences	2.5M/.3M/.3M	.2M/.03M/.03M	.2M/.02M/.02M	.19M	.03M
Avg. sentences/doc	61/61/60	06/04/06	05/05/06	102	69
Standard deviation	76/74/72	06/05/04	05/06/08	17	13
Compression ratio	–	20/20/20	17/18/17	–	30

Table 1: Statistics of the final version of the dataset.

chola et al., 2020).

3.3 Final Dataset

The extracted English and German articles are used to create the following sets.

1. **W-MS** - WIKIPEDIA monolingual dataset consisting of English texts and corresponding English summaries.

2. **W-CLS** - WIKIPEDIA cross-lingual dataset containing English texts and corresponding German summaries.

3. **S-CLS** - SPEKTRUM manually corrected test set consisting of English texts and corresponding German summaries.

4 Dataset Statistics

4.1 Overview

Table 1 provides statistics for the final version of the monolingual and cross-lingual datasets for train, val(idation), and two test sets (WSP and SPEKTRUM). There are pairs of text and summary in each set. Total articles are 41,049 (80%) in the train, 5,131 (10%) in the val, 5,132 (10%) in the W-CLS and 1,510 in the S-CLS test sets.

Table 2 presents a comparison of the proposed dataset with some of existing summarization datasets. Based on our observations, our datasets differ from existing datasets in various aspects, particularly cross-linguality.

4.2 Compression Ratio

The compression ratio is defined as the word ratio between a text and its summary (Grusky et al., 2018). Table 1 presents the compression ratio of English and German summaries. An English summary is typically 20% as long as an English text, and a German summary is 17.5% as long as an English text. It is important to note that while both languages belong to the Germanic family, they differ in inflection and compound words. Therefore, judging from these averages, it is difficult to determine whether English summaries are in fact longer than those in German.

4.3 Novel *N*-grams in Summaries

Table 3 presents the percentage of *n*-grams in the summaries that do not appear in the corresponding text for W-MS. The percentage of novel *n*-grams in the summaries serves as a measure of their abstractiveness. Approximately 25% of the summary unigrams for the train, val, and test sets are novel. The train, val, and test sets have almost 70% novel bigrams. The percentage of novel *n*-grams also increases as n (1−5) increases and reaches up to 93% for 5-grams. Furthermore, the Table 3 shows that the summaries have more novel words in them and that the dataset tends to be abstractive.

5 Experiments

5.1 Datasets and Baselines

We conduct an empirical evaluation of W-MS, W-CLS and S-CLS for the summarization task. For MS, we apply both extractive and abstractive methods to W-MS, with the extractive methods serving as baselines. For CLS, we apply the abstractive models to W-CLS and evaluate the models with two test sets: WIKIPEDIA and SPEKTRUM. For CLS baselines, we apply two existing pipeline methods to create PCLS data from W-MS by us-

	Newswire				Scientific				
Dataset	DM	CNN	NYT	NR	ARXIV	PM	**W-MS**	**W-CLS**	**S-CLS**
Avg. words/text	653	760	549	659	4900	3000	1559	1559	2337
Avg. words/sum	55	46	40	27	220	203	140	100	361
Compression ratio	12	16.5	13.8	24	22.5	15	20	18	30
Cross-lingual	No	No	No	No	No	No	No	**Yes**	**Yes**

Table 2: Comparison of W-MS, W-CLS and S-CLS to existing summarization datasets.

	1-g	**2-g**	**3-g**	**4-g**	**5-g**
Train	24.6	69.3	87.6	92.1	93.0
Val	24.5	69.1	87.4	91.9	92.7
Test	24.7	69.4	87.5	92.1	92.9

Table 3: Percentage of novel *n*-grams in W-MS summaries.

ing FairSeq[10]: (i) TRANS-SUM - Translate-then-Summarize, and (ii) SUM-TRANS - Summarize-then-Translate.

5.2 Methods

The following extractive methods are selected: (i) SUM-BASIC, (ii) LUHN, (iii) KL-SUM, (iv) LSA, (v) LEX-RANK, (vi) TEXT-RANK, and (vii) BERT[11,12].

The following abstractive models are chosen: (i) Attention based sequence to sequence model (S2S) (Bahdanau et al., 2015), (ii) Pointer generator network (PGN) (See et al., 2017), and (iii) Transformer based sequence to sequence model (TRF) (Vaswani et al., 2017). We select these models because these models show good results in previous studies (See et al., 2017; Ouyang et al., 2019; Duan et al., 2019). Moreover, we want to evaluate the performance of these models on our proposed dataset, therefore skipping the pre-trained embeddings and models.

The S2S and PGN models are applied with almost the same hyper-parameters as in See et al. (2017). Word embeddings are configured with 128 dimensions and hidden layers with 256 dimensions. The vocabulary size is 100K and 50K at the encoder and decoder sides, without the OOV words handling as used in the PGN model. In order to solve the OOV words, we choose BPE instead of the *n*-gram vocabulary. The Adam optimizer is used with a learning rate of 0.15 and a mini-batch of size 16. The models are trained for 40 epochs, and the validation loss is calculated to determine the best-trained model.

Almost the same hyper-parameters are applied for TRF as in Vaswani et al. (2017). Word embeddings have dimensions of 512 and hidden layers have dimensions of 786. The model consists of encoder and decoder stacks, each having 6 layers and 8 multi-attention heads at the decoder side. To make the results comparable among all models, the same vocabulary size of 100K and 50K at the encoder and decoder sides are selected. The Adam optimizer is used at a learning rate of 0.0001 and with a residual dropout of 0.1. For all abstractive models, a beam search of size 4 is applied in the inference phase. For all abstractive models, the encoder and decoder length is fixed to 400 and 100 words as in See et al. (2017). All abstractive models are trained on a single Tesla P40 GPU with 24GB RAM. For training and inference, the S2S and TRF models take around 6 days, and the PGN model takes 3 days.

5.3 Evaluation

For automatic evaluation, ROUGE metric is used for F-score, Precision and Recall. ROUGE relies on different metrics that include *n*-gram (R-N) and Longest Common Sub-sequence - LCS (R-L) overlap (Lin, 2004). Unigram and bigram overlap (R-1,2) provide a reasonable estimation of informativeness, while R-L estimates the summaries' fluency.

In order to further investigate the linguistic quality of system summaries, two native German speakers with fluent English have evaluated the summaries for two parameters (details are present in Section 6.3). It is worth to be noted that previous monolingual scientific summarization studies (Cohan et al., 2018; Dangovski et al., 2019) have not considered the human evaluations due to its demanding nature. For human evaluation of scientific articles, human judges must read and comprehend long domain-specific articles with summaries

[10] Github.com/fairseq
[11] (i-vi) Pypi/Sumy
[12] (vii) Pypi/Bertext

	R-1			R-2			R-L		
	F	P	R	F	P	R	F	P	R
Extractive models									
SUM-BASIC	28.67	22.82	38.68	07.15	05.53	10.12	**25.24**	20.06	34.02
TEXT-RANK	26.29	18.82	43.57	07.11	04.86	13.27	22.98	16.43	38.19
KL-SUM	24.96	17.73	42.13	06.40	04.42	11.58	21.64	15.35	36.65
LUHN	25.67	19.25	38.50	06.75	04.86	11.03	22.54	16.88	33.89
LEX-RANK	26.53	20.11	38.95	06.69	04.92	10.47	23.22	17.58	34.18
RANDOM	28.05	21.20	41.45	**07.51**	05.44	12.13	24.55	18.52	36.39
LSA	26.51	18.97	44.01	07.40	05.03	13.96	23.09	16.50	38.45
BERT	**28.74**	23.56	36.83	**07.51**	06.02	09.98	25.03	20.51	32.10
Abstractive models									
PGN	22.25	47.88	14.49	05.34	11.90	03.44	20.58	44.52	13.38
S2S	20.94	54.98	12.93	04.75	11.67	02.98	19.31	51.40	11.89
TRF	**25.53**	40.95	18.55	**06.29**	07.03	05.69	**22.76**	36.83	16.47

Table 4: Monolingual results of ROUGE evaluation of W-MS with different extractive and abstractive methods.

to evaluate the linguistic qualities of system summaries. It is more challenging to conduct cross-lingual evaluations as it requires bilingual comprehension for articles tailored to various science topics.

6 Results and Discussion

6.1 Monolingual Results

Table 4 presents the MS results of different extractive and abstractive models with the W-MS. For extractive methods, the BERT achieves the highest results for R-1 and R-2, whereas the SUM-BASIC performs well for R-L. Overall, all extractive techniques yield similar results. All abstractive models perform fairly well for R-1, R-2 and R-L. Nevertheless, the abstractive models have a slightly lower performance than the extractive models. In general, all the summarization methods perform worse for R-2 than R-1 and R-L.

We consider two factors when comparing monolingual extractive and abstractive results: (i) the impact of novel *n*-grams in the reference summaries, and (ii) the length of output summaries. Regarding the impact of novel *n*-grams, extractive methods are not impacted by the presence/absence of novel *n*-grams. For example, if we consider novel unigrams, as mentioned in Table 3, approximately 25% of the summary unigrams are not present in the corresponding text, but the remaining 75% unigrams can overlap. As the extractive methods extract the sentences from the actual text and maintain a good percentage of overlapped words. However, as abstractive models do not rely on extraction, their results can be influenced by the presence/absence of novel *n*-grams. Based on observation, the extractive results show that Recall is higher than Precision, indicating that the system summaries are longer than the reference summaries. From abstractive results, it can be observed that Precision is higher than Recall indicating that the system summaries tend to be shorter than reference summaries in contrast to extractive methods. Ideally, the system summaries should be similar to reference summaries. Nevertheless, since the models were evaluated on news datasets, they tend to produce short summaries.

6.2 Cross-lingual Results

Table 5 presents the CLS results with different abstractive models. We cannot compare our results with those of recent CLS studies since they used pseudo-cross-lingual data from the news domain. We overcome this problem by using two baselines, TRANS-SUM and SUM-TRANS, which have been used in recent studies. Among the baselines, the SUM-TRANS models perform better than the TRANS-SUM models. However, these baseline models do not perform well in comparison with real CLS data models (W-CLS). The CLS models show significantly ($p < 0.05$) improved results with W-CLS data as compared to SUM-TRANS and TRANS-SUM models ($p < 1 \times 10^{-6}$). The results supported our hypothesis (as mentioned in Section 1) that MT introduced noise to pseudo-cross-lingual data (PCLS). Consequently, the data noise acts as a bias and affects the neural models. The CLS models

	R-1			R-2			R-L		
	F	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**
TRANS-SUM									
S2S	14.18	30.49	09.24	01.51	02.55	01.07	12.77	27.67	08.30
PGN	15.81	31.35	10.57	02.86	05.58	01.92	14.69	29.14	09.82
TRF	16.15	32.56	11.41	03.66	05.84	02.72	15.25	32.06	09.29
SUM-TRANS									
S2S	15.04	32.85	09.75	01.48	02.56	01.04	13.64	28.03	08.82
PGN	18.24	28.62	13.38	04.14	10.98	02.55	16.04	30.45	11.11
TRF	19.31	26.77	14.18	04.23	11.67	02.84	17.37	31.74	12.18
W-CLS test set									
S2S†	18.37	37.93	12.12	04.04	09.91	02.54	16.55	34.57	10.88
PGN†	20.72	30.34	15.73	03.79	05.93	02.79	18.68	27.48	14.15
TRF†	**21.61**	26.81	18.10	**04.37**	05.16	03.79	**18.10**	22.42	15.18
S-CLS test set									
S2S$^{\dagger *}$	16.47	26.42	11.97	03.42	03.43	03.41	11.87	25.47	07.74
PGN$^{\dagger *}$	18.64	29.74	13.54	03.83	04.05	03.63	15.65	26.42	11.12
TRF$^{\dagger *}$	**20.81**	31.39	14.47	**04.19**	05.43	03.41	**17.54**	21.73	15.29

Table 5: Cross-lingual results of ROUGE evaluation with different abstractive methods. † denotes a significant improvement in the results and * denotes a significant difference in the results.

learn the mappings between encoder and decoder sides language distributions along with compression. Therefore, distortion in language distributions (*e.g.*, wrong translated tokens, UNK tokens) can affect mappings' learning. Therefore, it is better to train a CLS model with real cross-lingual data rather than PCLS. Overall, the abstractive models perform well for R-1, R-2 and R-L with the W-CLS.

We extend CLS experiments to S-CLS using the same models trained for W-CLS. In these experiments, we examine how on-the-ground cross-lingual summarization models perform on a real-world dataset. The CLS models under-perform on S-CLS set ($p < 1 \times 10^{-4}$) with p-value ($p < 0.05$). The slight drop in performance is probably due to the fact that the decoder is a conditional model that learns contextual representations from training data. Moreover, it seems that BPE vocabulary caters to the unseen words of S-CLS set, as both test sets (W-CLS and S-CLS) have not been used in vocabulary construction.

The CLS results suggest that neural models can learn cross-lingual mappings as well as compression. Using WIKIPEDIA dataset, the models learn the structural mappings between English and German languages and tend to maintain a logical structure of sentences in summaries. Comparatively, all models perform poorly with R-2 compared to R-1 and R-L. Due to the short summaries produced by the models, Precision is higher than Recall, which in turn affects the F-score. Earlier, we noted that these models are designed for news datasets, which do not require long summaries. We selected these neural models because they have demonstrated good performance in machine translation and summarization. However, their implementation has not been tested for cross-lingual texts that are long compared to the use-cases previously mentioned.

Both tasks are different in nature, so there can be no direct comparison between the MS and CLS. In Appendix A, Figures 2 and 3 show the examples of monolingual and cross-lingual system-generated summaries and their reference summary. In abstractive models, blue color represents the creation of new words by the models, red represents the incorrect information, yellow depicts the extractive parts and bold shows the repetition of words/phrases.

6.3 Human Evaluation

Our human judges evaluate the linguistic quality of output summaries. They are native Germans speakers with fluent English skills. For evaluation of the models, we randomly select 20 output summaries, their reference summaries, and the input articles (10 from PGN_MS and 10 from PGN_CLS). The evaluation is performed for two parameters: (i) correctness and ii) fluency on a scale of $1-3$

as scale range used by Ouyang et al. (2019). Correctness measure defines whether the original message is preserved coherently (relevance) in a non-redundant manner. Fluency measure determines the structural and grammatical properties of summaries. For MS, the average score for correctness is 2.10, and fluency is 2.65, with a moderate agreement (Fleiss's $\kappa = 0.60$ and 0.58) between judges. For CLS, the average score for correctness is 1.65, and fluency is 1.96, with a substantial agreement (Fleiss's $\kappa = 0.70$) for both scores between judges. From these results, it can be observed that the fluency of the models is good in maintaining an appropriate structure of the output summaries, while the correctness of the models is modest. The cross-lingual models tend to produce irrelevant content in some summaries.

7 Conclusions

Lack of cross-lingual experimental datasets impedes the progress of CLS research. In this paper, we present a new MS and CLS dataset extracted from SPEKTRUM and WIKIPEDIA. Our empirical investigation demonstrates the viability and amenability of the proposed dataset and also highlights the challenging nature of the dataset for recent summarization models. Our results demonstrate the significance of constructing real cross-lingual datasets for CLS. Furthermore, the English and German summaries scored reasonably well in terms of correctness and fluency based on human evaluations. We anticipate that the proposed dataset will encourage the study and research of MS and CLS. In the future, we will scale the CLS experiments using science domain pre-trained models.

Acknowledgments

We thank *Spektrum der Wissenschaft* for providing us their data. We would also like to acknowledge the assistance of Nadia Arslan, Fabian Düker and Jason Brockmeyer in the data extraction and manual evaluation. This work is carried out at the Heidelberg Institute for Theoretical Studies (supported by the Klaus Tschira Foundation), Heidelberg, Germany. In addition, the first author is funded by the Higher Education Commission of Pakistan (HEC) - the Deutscher Akademischer Austausch Dienst (DAAD).

References

Diego Antognini and Boi Faltings. 2020. GameWikiSum: a novel large multi-document summarization dataset. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6645–6650, Marseille, France. European Language Resources Association.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel Weld. 2020. TLDR: Extreme summarization of scientific documents. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4766–4777, Online. Association for Computational Linguistics.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Rumen Dangovski, Li Jing, Preslav Nakov, Mićo Tatalović, and Marin Soljačić. 2019. Rotational unit of memory: A novel representation unit for RNNs with scalable applications. *Transactions of the Association for Computational Linguistics*, 7:121–138.

Xiangyu Duan, Mingming Yin, Min Zhang, Boxing Chen, and Weihua Luo. 2019. Zero-shot cross-lingual abstractive sentence summarization through teaching generation and attention. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3162–3172, Florence, Italy. Association for Computational Linguistics.

Dominik Frefel. 2020. Summarization corpora of Wikipedia articles. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6651–6655, Marseille, France. European Language Resources Association.

Demian Gholipour Ghalandari, Chris Hokamp, Nghia The Pham, John Glover, and Georgiana Ifrim. 2020. A large-scale multi-document summarization dataset from the Wikipedia current events portal. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1302–1308, Online. Association for Computational Linguistics.

George Giannakopoulos. 2013. Multi-document multilingual summarization and evaluation tracks in ACL 2013 MultiLing workshop. In *Proceedings of the*

MultiLing 2013 Workshop on Multilingual Multi-document Summarization, pages 20–28, Sofia, Bulgaria. Association for Computational Linguistics.

George Giannakopoulos, Mahmoud El-Haj, Benoît Favre, Marina Litvak, Josef Steinberger, and Vasudeva Varma. 2011. TAC2011 MultiLing Pilot Overview. In *Proceedings of the Fourth Text Analysis Conference, TAC 2011*, Gaithersburg, Maryland, USA. NIST.

George Giannakopoulos, Jeff Kubina, John Conroy, Josef Steinberger, Benoit Favre, Mijail Kabadjov, Udo Kruschwitz, and Massimo Poesio. 2015. MultiLing 2015: Multilingual summarization of single and multi-documents, on-line fora, and call-center conversations. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 270–274, Prague, Czech Republic. Association for Computational Linguistics.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Benjamin Hättasch, Nadja Geisler, Christian M. Meyer, and Carsten Binnig. 2020. Summarization beyond news: The automatically acquired fandom corpora. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6700–6708, Marseille, France. European Language Resources Association.

Minsoo Kim, Dennis Singh Moirangthem, and Minho Lee. 2016. Towards abstraction from extraction: Multiple timescale gated recurrent unit for summarization. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 70–77, Berlin, Germany. Association for Computational Linguistics.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online. Association for Computational Linguistics.

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

Khanh Nguyen and Hal Daumé III. 2019. Global Voices: Crossing borders in automatic news summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 90–97, Hong Kong, China. Association for Computational Linguistics.

Nikola I Nikolov, Michael Pfeiffer, and Richard HR Hahnloser. 2018. Data-driven Summarization of Scientific Articles. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jessica Ouyang, Boya Song, and Kathy McKeown. 2019. A robust abstractive system for cross-lingual summarization. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2025–2031, Minneapolis, Minnesota. Association for Computational Linguistics.

Elvys Linhares Pontes, Stéphane Huet, Juan-Manuel Torres-Moreno, and Andréa Carneiro Linhares. 2020. Compressive Approaches for Cross-language Multi-document Summarization. *Data & Knowledge Engineering*, 125:101763.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Shi-qi Shen, Yun Chen, Cheng Yang, Zhi-yuan Liu, and Mao-song Sun. 2018. Zero-shot Cross-lingual Neural Headline Generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(12):2319–2327.

Raghuram Vadapalli, Bakhtiyar Syed, Nishant Prabhu, Balaji Vasan Srinivasan, and Vasudeva Varma. 2018a. Sci-blogger: A step towards automated science journalism. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*, pages 1787–1790. ACM.

Raghuram Vadapalli, Bakhtiyar Syed, Nishant Prabhu, Balaji Vasan Srinivasan, and Vasudeva Varma. 2018b. When science journalism meets artificial intelligence : An interactive demonstration. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 163–168, Brussels, Belgium. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Jiajun Zhang, Yu Zhou, and Chengqing Zong. 2016. Abstractive Cross-language Summarization via Translation Model Enhanced Predicate Argument Structure Fusing. *IEEE/ACM Transac-*

tions on Audio, Speech, and Language Processing, 24(10):1842–1853.

Junnan Zhu, Qian Wang, Yining Wang, Yu Zhou, Jiajun Zhang, Shaonan Wang, and Chengqing Zong. 2019. NCLS: Neural cross-lingual summarization. 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)*, pages 3054–3064, Hong Kong, China. Association for Computational Linguistics.

Markus Zopf, Maxime Peyrard, and Judith Eckle-Kohler. 2016. The next step for multi-document summarization: A heterogeneous multi-genre corpus built with a novel construction approach. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1535–1545, Osaka, Japan. The COLING 2016 Organizing Committee.

A Example Summaries

Monolingual Article: united nations conference on sustainable development background in 1992 , the first conference of its kind , the united nations conference on environment and development (unced) , commonly referred to as the rio conference or earth summit , succeeded in raising public awareness of the need to integrate environment and development . the conference drew 109 heads of state to rio de janeiro , brazil , to address what were dubbed urgent problems of environmental protection and socio-economic development . [...]

Reference: the united nations conference on sustainable development (uncsd) , also known as rio 2012 , or earth summit 2012 was the third international conference on sustainable development aimed at reconciling the economic and environmental goals of the global community. hosted by brazil in rio de janeiro from 13 to 22 june 2012 , [...]

BERT: united nations conference on sustainable development background in 1992 , the first conference of its kind , the united nations conference on environment and development (unced) , commonly referred to as the rio conference or earth summit , succeeded in raising public awareness of the need to integrate environment and development . [...]

LEX-RANK: the world conference on human rights , for example , focused on the right of people to a healthy environment and the right to development ; controversial demands that had met with resistance from some member states until the earth summit . it also created new international institutions , among them the commission on sustainable development , tasked with the follow-up to the rio conference and led to the reform of the global environment facility . [...]

PGN: the united nations environment programme (unep–1898) is an international non-governmental organization devoted to the united nations convention on climate change in the united states . it is the world 's fourth-largest programme in the united states and **the united nations environment programme (unep) and the united nations environment programme (unep)** . the international commission on climate change is the **world 's fourth-largest programme .**

TRF: the first conference of its kind , the united nations conference on environment and development (unced) was held in 1992 . the conference drew heads of state to rio de janeiro , brazil to address what were dubbed urgent problems of environmental protection and socio-economic development . the earth summit influenced subsequent un conferences and set the global green agenda .

Figure 2: Example of monolingual system-generated summaries.

Cross-lingual Article: d' arrest discovered ngc525 using his 11-inch refractor telescope at copenhagen . he located the galaxy 's position with a total of two observations . as he also noted the mag 11-12 star just 2' northwest , his position is fairly accurate . the galaxy was later catalogued by john louis emil dreyer in the new general catalogue , where it was described as very faint , very small , 11th or 12th magnitude star 5 seconds of time to west . the galaxy appears very dim in the sky as it only has an apparent visual magnitude of 13.3 and thus can only be observed with telescopes . [...]

Reference: ngc525 ist eine linsenförmige galaxie vom hubble - typ s0 im sternbild fische auf der ekliptik . sie ist schätzungsweise 99 millionen lichtjahre von der milchstraße entfernt und hat einen durchmesser von etwa 40 . 000 lichtjahren . im selben himmelsareal befinden sich u.a. die galaxien ngc516 , ngc524 , ic101, ic102 . das objekt wurde am 25 . september 1862 von dem deutsch-dänischen astronomen heinrich ludwig d' arrest entdeckt .
Translation: ngc525 is a lenticular galaxy of the hubble type s0 in the constellation pisces on the ecliptic . it is estimated to be 99 million light years from the milky way and about 40 , 000 light years across. the galaxies ngc516 , ngc524 , ic101 , ic102 are located in the same area of the sky . the object was discovered on september 25 , 1862 by the german danish astronomer heinrich ludwig d' arrest .

TRANS-SUM-PGN: ngc142 ist eine unregelmäßige galaxie im sternbild eridanus . sein d <unk> l wurde mit hilfe der kugelsternhaufenluminosität auf 31 , 01 ± 0 , 21 geschätzt , was etwa 52 ms entspricht . er ist das hellste <unk> <unk> .
Translation: ngc142 is an irregular galaxy in the constellation of eridanus . its d<unk>l was estimated with the help of the globular cluster luminosity to 31 . 01 ± 0 . 21 , which corresponds to about 52 ms . it is the brightest <unk> <unk> .

W-CLS-PGN: ngc499 ist eine elliptische galaxie mit aktivem galaxienkern vom hubble - typ e0 im sternbild fische am nordsternhimmel . sie ist schätzungsweise 22 millionen lichtjahre von der milchstraße entfernt und hat einen durchmesser von etwa 70 . 000 lichtjahren .
Translation: ngc499 is an elliptical galaxy with an active galaxy core of the hubble type e0 in the constellation pisces in the north star sky . it is an estimated 22 million light years from the milky way and has a diameter of around 70 , 000 light years .

W-CLS-TRF: d' arrest entdeckte ngc990 mit seinem 11 - zoll - refraktorteleskop . die galaxie wurde von john ratter im neuen katalog katalogisiert , wo sie als sehr kleiner stern beschrieben wurde .
Translation: d' arrest discovered ngc990 with its 11 - inch refractor telescope . the galaxy was cataloged by john ratter in the new catalog , where it was described as a very small star .

Figure 3: Example of cross-lingual system-generated summaries.