SimCSum: Joint Learning of Simplification and Cross-lingual Summarization for Cross-lingual Science Journalism

Mehwish Fatima[†], **Tim Kolber**[†], **Katja Markert**[‡] and **Michael Strube**[†]
[†]Heidelberg Institute for Theoretical Studies, Heidelberg, Germany
[‡]Department of Computational Linguistics, Heidelberg University
Heidelberg, Germany
(mehwish.fatima|tim.kolber|michael.strube)@h-its.org
markert@cl.uni-heidelberg.de

Abstract

Cross-lingual science journalism is a recently introduced task that generates popular science summaries of scientific articles different from the source language for non-expert readers. A popular science summary must contain salient content of the input document while focusing on coherence and comprehensibility. Meanwhile, generating a cross-lingual summary from the scientific texts in a local language for the targeted audience is challenging. Existing research on cross-lingual science journalism investigates the task with a pipeline model to combine text simplification and cross-lingual summarization. We extend the research in cross-lingual science journalism by introducing a novel, multi-task learning architecture that combines the aforementioned NLP tasks. Our approach is to jointly train the two high-level NLP tasks in SIMCSUM for generating cross-lingual popular science summaries. We investigate the performance of SIMCSUM against the pipeline model and several other strong baselines with several evaluation metrics and human evaluation. Overall, SIMCSUM demonstrates statistically significant improvements over the state-of-the-art on two non-synthetic cross-lingual scientific datasets. Furthermore, we conduct an in-depth investigation into the linguistic properties of generated summaries and an error analysis.

1 Introduction

Cross-lingual science journalism is a recently introduced task that produces science summaries in a target language from scientific documents in a source language while emphasizing simplification (Fatima and Strube, 2023). A real-world example of cross-lingual science journalism is Spektrum der Wissenschaft[1]. It is the German version of Scientific American and an acclaimed bridge between local readers and the latest scientific research in Germany. Spektrum's target audience are non-expert adults, so their journalists summarize complex scientific concepts in easy-to-understand terms in their local language. These scientific summaries are distinctive from other scientific texts/abstracts due to their length, which is more concise than regular scientific articles, containing non-complex and simplified terms, and are generated in a different language from their source.

Previous work on science journalism, including monolingual and cross-lingual, is quite limited. Monolingual science journalism has been investigated as a downstream task of abstractive summarization (Dangovski et al., 2021; Zaman et al., 2020) with customized monolingual datasets (Zaman et al., 2020; Goldsack et al., 2022). These datasets are, unfortunately, not suitable for cross-lingual science journalism. Moreover, cross-lingual science journalism has been investigated as a fusion of cross-lingual summarization and text simplification with a pipeline model (Fatima and Strube, 2023) with cross-lingual scientific datasets (Fatima and Strube, 2021). In the dawn of cross-lingual summarization, various pipeline models (Ouyang et al., 2019; Zhu et al., 2019, 2020) with synthetic cross-lingual datasets have been introduced to explore the task. Later, cross-lingual summarization models have been focused towards Multi-Task Learning (MTL) (Cao et al., 2020; Bai et al., 2021, 2022) and direct cross-lingual summarization with non-synthetic datasets (Ladhak et al., 2020; Fatima and Strube, 2021).

Due to the limited prior work in cross-lingual science journalism, this task needs further investigation. Fatima and Strube (2023) introduce the cross-lingual science journalism task with a pipeline-based model - **S**ELECT, **S**IMPLIFY and **R**EWRITE (SSR). SSR consists of three components - an extractive summarizer as SELECT, a reinforcement simplification model as SIMPLIFY, and a cross-lingual abstractive summarizer as REWRITE. SSR is a plug-and-play model which shows promising

[1]https://www.spektrum.de/magazin

Proceedings of The 4th New Frontiers in Summarization Workshop, pages 24–40

results for cross-lingual science journalism. However, this model has a few limitations. The SELECT component is not a trainable model and cannot be fine-tuned for the training data. The SIMPLIFY and REWRITE components require independent training and hyper-parameter settings. Therefore, these models cannot learn the mutual representations for simplification and cross-lingual summarization. We find this as a research gap. To the best of our knowledge, there are no available models for cross-lingual science journalism that can be trained at once on training data for simplification and cross-lingual summarization.

To fill this gap, we propose an MTL-based model - SIMCSUM that jointly trains for simplification and cross-lingual summarization to generate cross-lingual popular science summaries. SIMCSUM consists of one shared encoder and two independent decoders for each task based on a transformer architecture, where we consider cross-lingual summarization as our main task and simplification as our auxiliary task. The proposed model also leads us to two important research questions. (**RQ1**) Can jointly trained models learn and perform better than pipeline models for cross-lingual science journalism? (**RQ2**) Which linguistic features can effectively measure conciseness and readability to compare cross-lingual science journalism models? To investigate these research questions, we empirically evaluate the performance of SIMCSUM against SSR and several existing cross-lingual summarization models on two cross-lingual scientific datasets. We conduct a human evaluation to find the linguistic qualities of generated summaries. We further analyze the outputs for various lexical, readability and syntactic-based linguistic features. We also perform an error analysis to assess the quality of outputs.

2 Related Work

2.1 Scientific Summarization

This section focuses on the datasets for scientific summarization. Most science summarization datasets are collected from English scientific papers paired with abstracts: ARXIV (Kim et al., 2016; Cohan et al., 2018), PUBMED (Cohan et al., 2018; Nikolov et al., 2018), MEDLINE (Nikolov et al., 2018) and science blogs (Vadapalli et al., 2018b,a). Some work has been conducted for extreme summarization with monolingual dataset (Cachola et al., 2020), extended for cross-lingual extreme summarization (Takeshita et al., 2022). The extreme summarization task generates a one/two-line summary from a scientific abstract/paper, which makes it different from science journalism.

Cross-lingual scientific summarization is an understudied area due to its challenging nature. We find two studies: a synthetic dataset from English to Somali, Swahili, and Tagalog with round trip translation (Ouyang et al., 2019), two real cross-lingual datasets from Wikipedia Science Portal and Spektrum der Wissenschaft for English-German (Fatima and Strube, 2021).

2.2 Cross-lingual Summarization

This section focuses on MTL-based cross-lingual summarization. Zhu et al. (2019) develop an MTL model for English-Chinese cross-lingual summarization. They develop two variations of the transformer model (Vaswani et al., 2017), where the encoder is shared, and two decoders are independent. Cao et al. (2020) present a MTL model for cross-lingual summarization by joint learning of alignment and summarization. Their model consists of two encoders and two decoders, each dedicated to one task while sharing contextual representations. The authors evaluate their model on synthetic cross-lingual datasets for the English-Chinese language pairs. Takase and Okazaki (2022) introduce an MTL framework for cross-lingual abstractive summarization by augmenting (monolingual) training data with translations for three pairs: Chinese-English, Arabic-English, and English-Japanese. The model consists of a transformer encoder-decoder model with prompt-based learning in which each training instance is affixed with a special prompt to signal example type. Bai et al. (2021) develop a variation of multi-lingual BERT for English-Chinese cross-lingual abstractive summarization. The model is trained with a few shots of monolingual and cross-lingual examples. Bai et al. (2022) extend their work by introducing a MTL model to improve cross-lingual summaries by combining cross-lingual summarization and translation rates. They add a compression scoring method at the encoder and decoder of their model. They augment their datasets for different compression levels of summaries. One variation consists of cross-lingual and monolingual summarization decoders, while the other consists of cross-lingual and translation decoders.

Most of these studies focus on English-Chinese synthetic datasets emphasizing summarization and

translation (Zhu et al., 2019; Xu et al., 2020; Chen et al., 2023; Bai et al., 2021, 2022). By architecture, SIMCSUM is similar to Zhu et al. (2019) model as it also consists of one shared encoder and two task-specific decoders.

2.3 Science Journalism

This section focuses on science journalism models. Zaman et al. (2020) develop an extension of PGN (See et al., 2017) by modifying the loss function, so the model is trained for joint simplification and summarization. It is not a MTL model but a summarization model with an added loss for simplification. Moreover, the model is trained for monolingual science journalism on a customized dataset that contains simplified summaries from the Eureka Alert science news website. Dangovski et al. (2021) introduce monolingual science journalism as a downstream task of abstractive summarization and story generation. They apply BERT-based models with a prompting method for data augmentation on a monolingual dataset collected from Science daily press releases and scientific papers. They use three existing models for their work: SCI-BERT, a CNN-based sequence-to-sequence model and a story generation model.

Fatima and Strube (2023) propose cross-lingual science journalism as a downstream task of text simplification and cross-lingual summarization. They investigate the task on two non-synthetic cross-lingual scientific datasets with a pipeline-based model - **S**ELECT, **S**IMPLIFY and **R**EWRITE (SSR). To the best of author's knowledge, there is no other cross-lingual science journalism model to-date.

3 Proposed Model

Our model jointly trains for **Sim**plification and **C**ross-lingual **Sum**marization (SIMCSUM). We first define MTL and our tasks, and then discuss the architecture of our proposed model.

3.1 Multi-Task Learning

MTL is an approach in deep learning which improves generalization by learning different noise patterns from data related to different tasks. We define our MTL-based model trained on two tasks: simplification and cross-lingual summarization. We adopt hard parameter sharing as it improves the positive transfer and reduces the risk of overfitting (Ruder, 2017).

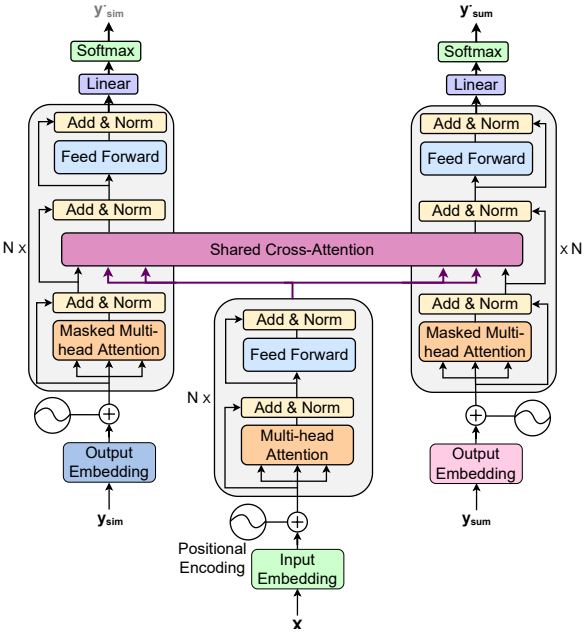

Figure 1: Architecture of SIMCSUM.

3.2 Summarization

We define single-document abstractive summarization as follows. Given a text $X = \{x_1, \cdots, x_m\}$ with m number of sentences comprising of a set of words (vocabulary) $W_X = \{w_1, \cdots, w_X\}$, an (encoder-decoder-based) abstractive summarizer generates a summary $Y = \{y_1, \cdots, y_n\}$ with n sentences that contain salient information of X, where $m \gg n$ and Y consisting of a set of words $W_Y = \{w_1, \cdots, w_Y | \exists\, w_i \notin W_X\}$. The decoder learns the conditional probability distribution over the given input and all previously generated words, where t denotes the time step.

$$P_\theta(Y|X) = \log P(y_t|\ y_{<t}, X) \tag{1}$$

Cross-lingual summarization adds another dimension of language for simultaneous translation and summarization. Given a text $X^l = \{x_1^l, \cdots, x_m^l\}$ in a language l with m sentences comprising of a vocabulary $W_X^l = \{w_1^l, \cdots, w_X^l\}$, a cross-lingual summarizer generates a summary $Y^k = \{y_1^k, \cdots, y_n^k\}$ in a language k that contains salient information in X, where $m \gg n$ and Y consisting of a vocabulary $W_Y^k = \{w_1^k, \cdots, w_Y^k | \exists\, w_i \notin W_X^l\}$. The conditional probability is the same as in Eq.1, the only difference being that the language on the decoder side is different from the encoder side.

3.3 Simplification

We define the document-level (lexical and syntactic) simplification task as follows. Given a

Algorithm 1 Training of SIMCSUM for Simplification and Cross-lingual Summarization

Input:
for each $d \in trainset$ **do**
▷ *Process each instance d of dataset D for tuples I of input x and targets for each task* $\mathcal{T}$
Create $I\langle x, y_{\mathcal{T}}\rangle$
end for

Initialize model parameters θ
Set maximum Epoch Ep
for epoch 1 to Ep **do**
for $b \in trainset$ **do**
▷ *b is a mini-batch containing I from* $trainset$
▷ SIMCSUM *consists of Encoder E, two Decoders* $D_{\mathcal{T}}$
Feed x to E and get the cross-attention
Feed $y_{\mathcal{T}}$ to $D_{\mathcal{T}}$
Feed the cross-attention to $D_{\mathcal{T}}$ [eq. (2)]
$t \leftarrow 0$
while θ_t is not converged **do**
$t \leftarrow t + 1$
Compute $\mathcal{L}(\theta)$ [eq. (3)]
Compute gradient $\nabla(\theta_t)$
Update $\theta_t \leftarrow \theta_{t-1} - \eta\nabla(\theta)$
end while
end for
end for

text $X = \{x_1, \cdots, x_m\}$ with m sentences comprising of a vocabulary $W_X = \{w_1, \cdots, w_X\}$, a simplification model generates the output text $Y = \{y_1, \cdots, y_n\}$ that retains the primary meaning of X, yet more comprehensible as compared to X, where $m \approx n$ and Y consisting of a vocabulary $W_Y = \{w_1, \cdots, w_Y | \exists\, w_i \notin W_X\}$. The conditional probability is also the same as in Eq.1.

3.4 SimCSum

We illustrate the framework of SIMCSUM[2] in Figure 1. SIMCSUM jointly trains on simplification and cross-lingual summarization. SIMCSUM adopts hard parameter sharing where the encoder is shared between the tasks while having two task-specific decoders. The decoders only share the cross-attention layer, and the loss is combined to update the parameters (θ). We opt for two decoders because each task's output language and length differ. The training method is described in Algorithm 1. Here we discuss the further details of SIMCSUM. For all mathematical definitions, $\mathcal{T} \in \{sim, sum\}$ denotes a task.

3.4.1 Architecture

Considering the excellent text generation performance of multi-lingual Bart (MBART) (Liu et al., 2020), we implement the SIMCSUM model based on it and modify it for two decoding sides for each task. Each encoder and decoder stack consists of 12 layers.

Self-Attention. Each layer of encoder/decoder has its self-attention, consisting of keys, values, and queries generated from the same sequence.

$$A(Q, K, V) = Softmax(\frac{Q \cdot K^T}{\sqrt{d_k}}) \cdot V$$

where Q is a query, K^T is transposed K (key) and V is the value. All parallel attentions are concatenated to generate multi-head attention scaled with a weight matrix W.

$$MH(Q, K, V) = Concat(A_1, \cdots, A_h) \cdot W^O$$

Cross-attention. The cross-attention connects the encoder and decoder and provides the decoder with a weight distribution at each step, indicating the importance of each input token in the current context. We concatenate the cross-attention of both decoders.

$$A(E, D_{\mathcal{T}}) = Concat(Softmax(\frac{D_{\mathcal{T}} \cdot E^T}{\sqrt{d_k}}) \cdot E) \tag{2}$$

where E is the encoder representation, $D_{\mathcal{T}}$ is the task-specific decoder contextual representation, and d_k is the model size.

3.4.2 Training Objective

We train our model end-to-end to maximize the conditional probability of the target sequence given a source sequence. We define the task-specific loss as follows.

$$\mathcal{L}_{\mathcal{T}}(\theta) = \sum_{n=1}^{N} \log P(y_{\mathcal{T}_t} | y_{\mathcal{T}_{<t}}, x; \theta)$$

where x represents the input, y is the target, N is the mini-batch size, t is the time step and θ denotes learnable parameters. We define the total loss of our model by task-specific losses where $\lambda_{\mathcal{T}}$ is an assigned weight to each task.

$$\mathcal{L}(\theta) = \sum \lambda_{\mathcal{T}} \cdot \mathcal{L}_{\mathcal{T}}(\theta) \tag{3}$$

3.5 Comparison with SSR

Here, we discuss the key differences between SSR (Fatima and Strube, 2023) and SIMCSUM. SSR is a component-based approach that combines three distinct modules: SELECT, SIMPLIFY and REWRITE. Each of these components addresses

[2] https://github.com/MehwishFatimah/SimCSum

a specific aspect of cross-lingual science journalism. It allows for fine-grained control over each step of the process, making it easier to analyze the individual contributions of these components. However, SIMCSUM is a unified model, where a single model is trained to handle both simplification and summarization. It needs to learn to balance and optimize these tasks simultaneously. It leverages synergies between the two tasks, potentially improving performance by learning shared representations between simplification and summarization. Furthermore, SSR introduces a degree of complexity due to its multi-component nature but allows for more control and transparency in the overall process. While SIMCSUMcan simplify the pipeline as it employs a single jointly trained model, potentially reducing complexity and taking benefits of shared representations.

4 Experiments

4.1 Datasets

4.1.1 Summarization

WIKIPEDIA. It is harvested from Wikipedia Science Portal for English and German (Fatima and Strube, 2021). Wikipedia Science Portal contains articles in various science fields. The WIKIPEDIA dataset consists of 50,132 English articles (avg. 1572 words) and German summaries (avg. 100 words).

SPEKTRUM. It is collected from Spektrum der Wissenschaft (Fatima and Strube, 2021). Spektrum is a famous science magazine (Scientific American) in Germany. It covers various topics in diverse science fields: astronomy, biology, chemistry, archaeology, mathematics, physics, *etc*. The SPEKTRUM dataset contains 1510 English articles (avg. 2337 words) and German summaries (avg. 361 words).

4.1.2 Simplification

We construct a synthetic WIKIPEDIA dataset for the simplification task by applying Keep-It-Simple (KIS) (Laban et al., 2021). To create the simplified WIKIPEDIA, we fine-tune KIS on WIKIPEDIA English articles as KIS is an unsupervised model and does not require parallel data. The simplified WIKIPEDIA consists of the original English articles paired with simplified English articles.

4.2 Split and Usage

We use WIKIPEDIA for training, validation and testing (80/10/10), while we use SPEKTRUM for zero-shot adaptability as a case study. All PLM baselines are trained on WIKIPEDIA where each instance I in the training set consists of $< x, y >$ where x is the input English text and y is the target German summary. SIMCSUM is trained on WIKIPEDIA where each instance I in the training set contains $< x, y_{sim}, y_{sum} >$ where x denotes the input English article and y_{sim} refers to the simplified English article and y_{sum} is the target German summary.

4.3 Models

Baselines. Almost all cross-lingual MTL models in §2 are based on translation and summarization, and none of them applies simplification. So we select several PLMs that accept long input texts as baselines. We fine-tune the following baselines: (1) mT_5 (Xue et al., 2021), (2) MBART (Liu et al., 2020), (3) PEGASUS (Zhang et al., 2020a), (4) LongFormer-Encoder-Decorder (LONG-ED) (Beltagy et al., 2020), and (5) XLSUM (Hasan et al., 2021) and (6) BIGBIRD (Zaheer et al., 2020). We also consider SSR as a strong baseline for cross-lingual science journalism.

SimCSum. We set $\lambda_{Sum} = 0.75$ for SIMCSUM based on the best results on the WIKIPEDIA validation set.

4.4 Training and Inference

The libraries, hardware and training time details are presented in Appendix A. Here, we discuss hyper-parameters.

Baselines. We fine-tune all models for a maximum of 25 epochs and average the results of 5 runs for each model. We use a batch size of 4-16, depending on the model size. We use a learning rate (LR) of $5e^{-5}$ and 100 warm-up steps to avoid over-fitting of the fine-tuned models. We use the Adam optimizer with a LR linearly decayed LR scheduler. The encoder language is set to English, and the decoder language is German.

SimCSum. We adopt similar settings as used for baselines, except for the batch size fixed to 4. We only generate tokens from the Summarization decoder side in the inference period. We use beam search of size 5 and a tri-gram block during the decoding stage to avoid repetition.

4.5 Evaluation

Automatic. We evaluate all models with three metrics. ROUGE (Lin, 2004) is a standard metric for summarization. BERT-score (Zhang et al., 2020b)

MODELS	R1	R2	RL	BS	FRE
GOLD	-	-	-	-	36.93
mT_5	26.79	12.65	23.40	69.12	45.42
MBART	31.43	13.20	25.12	70.52	44.67
PEGASUS	29.30	13.93	24.62	69.83	43.39
XLSUM	31.91	13.30	24.14	70.04	37.83
BIGBIRD	29.23	13.72	24.60	69.19	41.42
LONG-ED	15.11	06.82	13.67	63.94	24.48
SSR	32.17	13.56	25.01	70.60	**48.34**†
SIMCSUM	**34.50**†	**14.36**†	**25.85**†	**71.60**†	46.86

Table 1: The WIKIPEDIA results for all baselines and SIMCSUM. GOLD denotes the reference summaries. **Bold**† denotes the best overall results with significant improvements ($p < .001$).

MODELS	R1	R2	RL	BS	FRE
GOLD	-	-	-	-	40.76
mT_5	09.21	00.75	06.50	58.52	38.18
MBART	16.16	01.47	13.89	62.11	39.17
PEGASUS	11.49	00.95	08.01	60.56	37.93
XLSUM	17.10	01.63	09.79	62.25	33.83
BIGBIRD	12.28	01.04	08.65	59.97	36.24
LONG-ED	01.32	00.11	01.18	51.85	30.16
SSR	**21.14**†	**04.34**†	**15.15**†	63.09	41.01
SIMCSUM	20.98	03.82	14.16	**63.47**†	**41.03**†

Table 2: The SPEKTRUM results for all baselines and SIMCSUM. GOLD denotes the reference summaries. **Bold**† denotes the best overall results with significant improvements ($p < .001$).

(BS) is a recent metric for summarization and simplification as an alternative metric to n-gram-based metrics and applies contextual embeddings. For readability, we use Flesch Kincaid Reading Ease (FRE) (Kincaid et al., 1975) (Appendix B §B.2).

Human. We conduct a human evaluation to compare the outputs of SIMCSUM with MBART (baseline) for the same linguistic properties. Our annotators are two university students from the Computational Linguistics department with fluent German and English skills. It is worth mentioning that human evaluation of long cross-lingual scientific text is challenging and costly because it requires bi-lingual annotators with a scientific background.

5 Results

5.1 WIKIPEDIA

We report F-score of ROUGE and BERT-score and FRE of all models in Table 1. The first block includes the fine-tuned PLM models, the second block presents the pipeline baseline, and the last block includes SIMCSUM. From Table 1, we note that SIMCSUM outperforms all baselines for every metric except FRE. We compute the statistical significance of the results with the Mann-Whitney two-tailed test for a p-value $p < .001$. Interestingly, WIKIPEDIA summaries are not simplified compared to SPEKTRUM summaries; still, SIMCSUM performs better on WIKIPEDIA than the baselines. We interpret that the simplification auxiliary task helps SIMCSUM to learn a better contextual representation and produce more relevant German words. We deduce from the results that joint learning of simplification and cross-lingual summarization improves the quality of summaries.

Among the baselines, almost all models demonstrate comparable performance except LONG-ED. For R1, SSR perform better than other models, however, MBART and XLSUM perform also similar. PEGASUS takes the lead for R2, and MBART shows higher performance for RL. SSR and MBART take the lead for BS among the baselines. For FRE, a score between $30 - 50$ is the readability level best understood by college graduates. The WIKIPEDIA summaries fall in this range. For FRE, SSR performs the best among all models. Interestingly, almost all baselines except BIGBIRD and XLSUM demonstrate good performance.

5.2 Case Study: SPEKTRUM

Table 2 presents the results of all models on SPEKTRUM. We find a similar pattern that SIMCSUM outperforms all baselines except SSR for ROUGE-scores. We also compute the statistical significance of these results with the same procedure. The SPEKTRUM results are on the lower side compared to the WIKIPEDIA results due to zero-shot adaptability, especially for R2. This is because the ROUGE score computes n-gram overlap (Ng and Abrecht, 2015). The SPEKTRUM summaries have higher FRE scores compared to WIKIPEDIA. Interestingly, we note that all baselines perform lower than the GOLD summaries. However, the SIMCSUM score is similar to the GOLD summaries.

Human Evaluation. We compare the SIMCSUM and MBART outputs for analyzing linguistic qualities because SIMCSUM's architecture is based on MBART. We provide 30×2 (for each model) random summaries with their original texts. We ask two annotators to evaluate each document for three linguistic properties on a Likert scale from $1 - 5$. The first five samples are used to calibrate the an-

MODELS	FLUENCY	RELEVANCE	SIMPLICITY
mBART	2.28 (0.64)	1.64 (0.70)	1.86 (0.56)
SIMCSUM	2.62 (0.87)	2.76 (0.78)	2.88 (0.81)

Table 3: The SPEKTRUM human evaluation for mBART and SIMCSUM. The average scores (Krippendorff's α) for each linguistic feature are presented here.

notations of annotators, and then each annotator provides independent judgments on the rest of the samples.

Table 3 shows the human evaluation results. The samples used for calibration are not used for computing the scores (guidelines in Appendix C). We compute the inter-rater reliability by using Krippendorff's α[3]. We find that SIMCSUM improves the fluency, relevance and readability of outputs. We present a few comparative examples of SIMCSUM and mBART in Appendix E.

6 Readability Analysis: SPEKTRUM

The automatic evaluation of models show higher performance of SIMCSUM for ROUGE and BERTscore for WIKIPEDIA. However, it is not the case for FRE. Interestingly, for SPEKTRUM, SIMCSUM shows better performance for FRE. So we decide to further investigate readability with lexical diversity, readability scores and syntactic analysis to determine the quality of generated summaries. These types of analyses are well-known in NLP for textual analysis (Aluisio et al., 2010; Hancke et al., 2012; Vajjala and Lučić, 2018; Mosquera, 2022; Weiss and Meurers, 2022). The lexical diversity and readability scores are computed over all SPEKTRUM's reference summaries (Gold) and outputs of mBART, SSR and SIMCSUM. The gold summaries' score is a guideline for how similar the models' outputs are to gold summaries.

6.1 Lexical Diversity

Lexical diversity estimates the overall language distribution and computes cohesion through synonyms in the text. It is a good indicator of the readability of a text. We calculate Shannon Entropy Estimation (SEE) (Shannon, 1948) and Measure of Textual Lexical Diversity (MTLD) (McCarthy, 2005) to find lexical diversity (see Appendix B.1 for the formula).

SEE presents a text's "informational value" and language diversity. It is a language-dependent feature, and its value varies for different languages.

[3]https://github.com/LightTag/simpledorff

FEATURES	GOLD	mBART	SSR	SIMCSUM
Lexical Diversity				
SEE ↓	4.25	4.26	4.25	4.25
MTLD ↑	201	65.13	90.32	91.75
Readability scores				
CLI ↓	18.45	21.64	19.98	20.96
ARI ↓	18.99	21.07	21.16	20.26

Table 4: Lexical diversity and readability features' average scores.

Higher SEE scores suggest higher lexical diversity. We aim to get similar SEE as Gold summaries. Table 4 shows SEE scores of all three models which are similar to Gold summaries suggesting the similar informational value of all summaries.

MTLD is considered a robust version of the type-token ratio (TTR) and calculates lexical diversity without considering text length. Higher MTLD represents the greater vocabulary richness. Table 4 presents MTLD scores of all three models. The gold summaries have the highest scores, while SIMCSUM is the second highest, SSR has a slightly lower score than SIMCSUM, and mBART has the lowest score. These scores suggest that the lexical richness of all groups is not similar, in contrast to SEE results. The SIMCSUM outputs are more lexically diverse than the SSR and mBART outputs. We deduce from the improved SIMCSUM scores that joint learning of simplification and cross-lingual summarization impacts word generation. These results also suggest that MTLD provides a better estimation of lexical diversity for our summaries.

6.2 Readability Scores

Readability scores measure comprehension levels of the text. One of the syllables-based readability scores is already presented in §5. Coleman and Liau (1975) suggests that word length in letters is a better predictor of readability than syllables. We calculate Coleman Liau Index (CLI) (Coleman and Liau, 1975) and Automated Readability Index (ARI) (Senter and Smith, 1967) as these do not rely on syllables (see Appendix B.1 for the formula).

CLI computes scores on word lengths. ARI computes scores on characters, words and sentences. For both CLI and ARI, the lower score is better as it shows the ease of reading and understanding. From Table 4, we note that Gold summaries have the lowest score, SIMCSUM and SSR have similar scores, while mBART has the highest score. We deduce from these scores that joint learning of simplifica-

FEATURES↓	GOLD	MBART	SSR	SIMCSUM
ASL	24.09	24.15	22.65	20.97
ADD	3.60	4.16	3.95	3.91
ADW	0.93	0.95	0.94	0.94
ATH	8.32	8.72	8.60	8.57

Table 5: Syntactic features' average scores.

ERROR TYPES	MBART	SIMCSUM
Non-German words	83	35
Wrong name entities	1	2
Unfaithful information	3	3

Table 6: Error occurrences for MBART and SIMCSUM summaries which may contain multiple errors.

tion and cross-lingual summarization has an impact on both word and sentence level because CLI only focuses on words, while ARI includes sentences also.

6.3 Syntactic Analysis

Syntactic analysis elaborates on how words and phrases are related in a sentence structure. We perform it with constituency trees on 25×3 (for each model) random summaries from MBART, SIMCSUM and the gold summaries. The total number of sentences for MBART is 70, for SSR is 120, for SIMCSUM is 80 and for gold is 131. Table 5 presents four syntactic features (see Appendix B.2 for definitions).

We note from the average sentence length (ASL) that SIMCSUM produces shorter sentences among all models, which exhibits syntactic simplicity. A small average dependency distance (ADD) shows that words with a dependency relation are close together, making the text easier to understand. Table 5 shows that SIMCSUM summaries have a smaller average dependency than SSR and MBART, much closer to gold summaries. Fewer dependents per word (ADW) make a text less ambiguous and thus easier to follow. Table 5 shows both SSR and SIMCSUM outputs have fewer dependents than the MBART outputs and are similar to gold summaries. The average tree height (ATH) explains the syntactic structural complexity of a sentence. Table 5 shows that SIMCSUM outputs are less structurally complex than SSR and MBART outputs, however, gold summaries have the least average tree height. We deduce from the syntactic analysis that joint learning of simplification and cross-lingual summarization positively impacts the syntactic properties of summaries.

7 Error Analysis

To further explore the challenges of generating cross-lingual science summaries, we consider the base model - MBART and SIMCSUM for analyzing the produced errors. We randomly select 25×2 (for each model) summaries from the SIMCSUM and MBART outputs. We note three main categories of errors in the manual inspection. Table 6 presents the occurrences of these errors in each model. Appendix D presents some examples from the error analysis and its guidelines.

Non-German Words. This is the error type where the models either produce non-existent German words or partially English-German or words in another language. We find that MBART is more prone to produce such errors. We note that it is due to the imbalance between the pre-trained and fine-tuned dataset sizes. These models are pre-trained on many languages and usually fine-tuned on comparatively smaller data. SIMCSUM tends to produce fewer errors due to data augmentation (simplification data) during the training.

Wrong Named Entities. This is the error type where the models produce wrong named entities, such as cities or country names and persons' first and last names. We find that both models tend to produce such errors, however, the percentage of such errors is quite low. We note that the models overestimate or underestimate the probability of word sequences present in data.

Unfaithful Information. This is the error type where we find some (new) information in generated summaries that is not faithful to the source documents. We note that this error is caused by long inputs where the model tends to hallucinate and generates some content that cannot be verified from the source. We find that SIMCSUM makes similar errors as MBART for this error type.

8 Conclusions

In this paper, we investigate a recently introduced task - cross-lingual science journalism. We propose a novel multi-task model - SIMCSUM- that combines two high-level NLP tasks, simplification and cross-lingual summarization. SIMCSUM jointly trains for reducing linguistic complexity and cross-lingual abstractive summarization. Our empirical investigation shows the significantly superior performance of SIMCSUM over the pipeline-based SSR

model and other baselines on two non-synthetic cross-lingual scientific datasets. This is confirmed by human evaluation. Furthermore, our in-depth linguistics analysis shows how multi-task learning in SIMCSUM has lexical and syntactic impacts on the generated summaries. We perform error analysis to find what kind of errors has been produced by the model. In the future, we plan to add modules for linguistically informed simplification.

9 Limitations

We proposed SIMCSUM for the cross-lingual science journalism task and verified its performance for WIKIPEDIA and SPEKTRUM datasets for the English-German language pair. We believe that SIMCSUM is adaptable for other domains and languages. However, we have not verified it experimentally and limited our experiments to English-German scientific texts.

Our model jointly trains on two high-level NLP tasks, which takes slightly more time than its base model - MBART, as it has more parameters to learn during the training. However, it takes much less time in training than SSR. Furthermore, our model is trained on synthetic simplification data, which may create a dependency on the simplification model - Keep-it-Simple (Laban et al., 2021). Therefore, we plan to add linguistically informed simplification modules in our model as our future work. We also find during error analysis that both MBART and SIMCSUM have problems (repetition or unfaithful information) with long inputs, which need further investigation that how we can mitigate such errors.

10 Ethical Consideration

Reproducibility. We discussed all relevant parameters, training details, and hardware information in §4.4 and Appendix A.

Legal Consent. We obtained legal consent from Spektrum der Wissenschaft to use their dataset. We adopted the public implementations with mostly recommended settings, wherever applicable.

Acknowledgements

The work was carried out at Heidelberg Institute for Theoretical Studies (supported by the Klaus Tschira Foundation), Heidelberg, Germany, under the collaborative Ph.D. scholarship scheme between the Higher Education Commission of Pakistan (HEC) and Deutscher Akademischer Austausch Dienst (DAAD). We would like to express our sincere gratitude to the reviewers whose feedback and insights greatly contributed to the quality of this work. We also extend our appreciation to the human annotators whose valuable contributions were essential for the success of this project.

References

Sandra Aluisio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. 2010. Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 fifth workshop on innovative use of NLP for building educational applications*, pages 1–9.

Yu Bai, Yang Gao, and He-Yan Huang. 2021. Cross-lingual abstractive summarization with limited parallel resources. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6910–6924.

Yu Bai, Heyan Huang, Kai Fan, Yang Gao, Yiming Zhu, Jiaao Zhan, Zewen Chi, and Boxing Chen. 2022. Unifying cross-lingual summarization and machine translation with compression rate. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1087–1097.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

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.

Yue Cao, Hui Liu, and Xiaojun Wan. 2020. Jointly learning to align and summarize for neural cross-lingual summarization. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 6220–6231.

Yulong Chen, Huajian Zhang, Yijie Zhou, Xuefeng Bai, Yueguan Wang, Ming Zhong, Jianhao Yan, Yafu Li, Judy Li, Xianchao Zhu, et al. 2023. Revisiting cross-lingual summarization: A corpus-based study and a new benchmark with improved annotation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9332–9351.

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.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine Scoring. *Journal of Applied Psychology*, 60(2):283.

Rumen Dangovski, Michelle Shen, Dawson Byrd, Li Jing, Desislava Tsvetkova, Preslav Nakova, and Marin Soljacic. 2021. We Can Explain Your Research in Layman's Terms: Towards Automating Science Journalism at Scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12728–12737, Online.

Mehwish Fatima and Michael Strube. 2021. A novel Wikipedia based dataset for monolingual and cross-lingual summarization. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 39–50, Online and in Dominican Republic. Association for Computational Linguistics.

Mehwish Fatima and Michael Strube. 2023. Cross-lingual science journalism: Select, simplify and rewrite summaries for non-expert readers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1843–1861.

Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2022. Making science simple: Corpora for the lay summarisation of scientific literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10589–10604.

Julia Hancke, Sowmya Vajjala, and Detmar Meurers. 2012. Readability classification for german using lexical, syntactic, and morphological features. In *Proceedings of COLING 2012*, pages 1063–1080.

Tahmid Hasan, Abhik Bhattacharjee, Md Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M Sohel Rahman, and Rifat Shahriyar. 2021. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703.

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.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

Philippe Laban, Tobias Schnabel, Paul Bennett, and Marti A. Hearst. 2021. Keep it simple: Unsupervised simplification of multi-paragraph text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6365–6378, Online. 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.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Philip M McCarthy. 2005. *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. Ph.D. thesis, The University of Memphis.

Alejandro Mosquera. 2022. Tackling data drift with adversarial validation: An application for german text complexity estimation. In *Proceedings of the GermEval 2022 Workshop on Text Complexity Assessment of German Text*, pages 39–44.

Jun-Ping Ng and Viktoria Abrecht. 2015. Better summarization evaluation with word embeddings for rouge. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1925–1930.

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.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

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.

RJ Senter and Edgar A Smith. 1967. Automated readability index. Technical report, Cincinnati Univ OH.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Sho Takase and Naoaki Okazaki. 2022. Multi-task learning for cross-lingual abstractive summarization. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3008–3016.

Sotaro Takeshita, Tommaso Green, Niklas Friedrich, Kai Eckert, and Simone Paolo Ponzetto. 2022. X-scitldr: cross-lingual extreme summarization of scholarly documents. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*, pages 1–12.

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.

Sowmya Vajjala and Ivana Lučić. 2018. Onestopenglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 297–304.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz 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.

Zarah Weiss and Detmar Meurers. 2022. Assessing sentence readability for german language learners with broad linguistic modeling or readability formulas: When do linguistic insights make a difference? In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 141–153.

Ruochen Xu, Chenguang Zhu, Yu Shi, Michael Zeng, and Xuedong Huang. 2020. Mixed-lingual pretraining for cross-lingual summarization. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 536–541.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297.

Farooq Zaman, Matthew Shardlow, Saeed-Ul Hassan, Naif Radi Aljohani, and Raheel Nawaz. 2020. Htss: A novel hybrid text summarisation and simplification architecture. *Information Processing & Management*, 57(6):102351.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020a. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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.

Junnan Zhu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2020. Attend, translate and summarize: An efficient method for neural cross-lingual summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1309–1321, Online. Association for Computational Linguistics.

A Training and Inference

Libraries. We train all models with Pytorch[4], Hugging Face[5] integrated with DeepSpeed[6] for parallel model training with ZeRO-2. We apply ZeRO-2[7] to enable model parallelism. ZeRO-2 reduces the memory footprints for gradients and optimizer because it shards the optimizer states and gradients across GPUs.

Hardware. For all models, we complete training and inference on 4 Tesla P40 GPUs each with 24GB memory.

Training Time. We use the maximum length capacity at the encoder side and set length of 200 tokens at the decoder side. mBART takes 1 day and 17 hours, mT_5 takes 10 hours, PEGASUS takes 1 day and 8 hours, XLSUM takes 1 day and 3 hours, LONG-ED takes almost 4 days, and BIGBIRD takes 2 days to complete 25 epochs. SIMCSUM takes 2 days to complete 25 epochs.

B Analysis: SPEKTRUM

B.1 Lexical Diversity

SEE is calculated with a frequency table as follows.

$$H(x) = \sum_{i=1}^{n} p(x_i) log_2 \frac{1}{p(x_i)}$$

where $H(x)$ is the total amount of information in an entire probability distribution. $P(x_i)$ refers to the frequency of a token appearing in the text, and $1/p(x)$ denotes the information of each case.

MTLD divides the texts into sequences having the same TTR and then calculates the mean length of the sequences.

B.2 Readability

FRE is calculated as follows:

$$FRE = 180 - ASL - (58.5 \times ASW)$$

where average sentence length (ASL) is the number of words divided by the number of sentences in the text. The average number of syllables per word (ASW) is the number of syllables divided by the number of words in the text. The numeric values are language-dependent constants.

CLI is calculated as follows:

$$CLI = 5.88 \times \frac{L}{W} - 29.6 \times \frac{S}{W} - 15.8$$

where L is the total number of characters (including numbers and punctuation), W is the total number of words, and S is the total number of sentences in a given text.

ARI is computed as follows:

$$ARI = 4.71 \times \frac{L}{W} + 0.5 \times \frac{W}{S} - 21.43$$

where L is the total number of characters (including numbers and punctuation), W is the total number of words, and S is the total number of sentences in a given text.

B.3 Syntactic Analysis

We use Stanza[8] to extract dependency relations and Stanford Parser [9] to extract constituency trees for each summary. Before tree generation, we replace all German umlauts (ä, ö, ü and ß) in the summaries with their replacements (ae, oe, ue and ss) due to encoding issues of the Stanford Parser.

Average Sentence Length. It is the number of tokens in the sentences averaged over the number of sentences in a summary.

Average Dependency Distance. It is the averaged dependency distance over the sentences, which means the distance between the dependency heads and their dependents.

Average Dependents per Word. It computes the average number of dependents for each word.

Average Tree Height. For computing the average tree height of a summary, we calculate the height of every tree and average it over the sentences.

C Human Evaluation

C.1 Task

We provided annotators with 30 examples of documents paired with a reference summary and two system-generated summaries. The models' identities had not been revealed. The annotators had to rate each model summary for the following linguistic properties after reading the English document and the German summaries. We asked annotators to use the first 5 examples to resolve the annotator's conflict and to find a common consensus for rating the linguistic aspects. However, the rest of the examples were annotated independently.

[4] https://pytorch.org/

[5] https://huggingface.co/

[6] https://www.microsoft.com/en-us/research/project/deepspeed/

[7] Initially, we used ZeRO-3 offload with FP16 evaluation, and the training became quite slow as it consumes a lot of time for offloading during evaluation.

[8] https://stanfordnlp.github.io/stanza/constituency.html

[9] https://nlp.stanford.edu/software/lex-parser.shtml

C.2 Linguistic Properties

We asked annotators to annotate each summary for the following linguistic properties.

Relevance. A summary delivers adequate information about the original text. Relevance determines the content relevancy of the summary.

Fluency. The words and phrases fit together within a sentence, and so do the sentences. Fluency determines the structural and grammatical properties of a summary.

Simplicity. Lexical (word) and syntactic (syntax) simplicity of sentences. A simple summary should have minimal use of complex words/phrases and sentence structure.

C.3 Scale

We use a Likert scale from 1 to 5 to score each property (1:worst | 2:bad | 3:ok | 4:good | 5:best). These scores should be assigned by comparing the outputs of both models.

D Error Analysis

D.1 Guidelines

We define our informal guidelines for the error analysis as follows. To find the errors in the MBART and SIMCSUM outputs, we compare them to each other, to the SPEKTRUM German gold summary and the original English text.

Non-German Words. To find them, it is sufficient to read through our model outputs and look up any unknown words. If one of the unknown words turns out to be a non-German word, we mark them in red.

Wrong named entities. We find wrong named entities by comparing the names in both system outputs to the reference summary. If the names differ, we verify with the original text that they refer to the same person and thus represent a mistake by the model, and we mark them in blue.

Unfaithful information. We note new/unfaithful information by looking up every piece of information in the model outputs in the reference summary. We search for this information in the original text, and if it is not present there, it is clear that the model produced new information that is not faithful to the source text. We mark this information in orange.

D.2 Examples

Target: im freigehege aufzuwachsen und sich dort im schlamm zu suhlen, stärkt offenbar das immunsystem von schweinen : verglichen mit artgenossen, die in einem stall gehalten wurden, hatten freilandschweine eine höhere anzahl von gesundheitsfördernden darmbakterien. zu diesem ergebnis kommen jetzt wissenschaftler um denise kelly von der university of aberdeen. auch gene, die auf entzündungsreaktionen hindeuten, waren bei tieren, die im außenbereich lebten, wesentlich weniger aktiv. [...] **Translation:** Growing up in an outdoor enclosure and wallowing in the mud there apparently strengthens the immune system of pigs : compared to conspecifics that were kept in a pen, free-range pigs had a higher number of health-promoting intestinal bacteria. scientists led by denise kelly from the university of aberdeen are now coming to this conclusion. genes that indicate inflammatory responses were also significantly less active in animals that lived outdoors. [...]
MBART: als fruehmikrobielle kolonisation wird **dieabdeckung** der brustdruese von schweinen bei der fruehen entwicklung des erwachsenen menschen bezeichnet. sie wird in der medizin, der mikrobiologie und der pathologie erforscht. **early microbial colonization of the gut reduced the incidence of infectious**, entzuendlichen und **autoimmunischen diseasen reduced the incidence of infectious, inflammatory and autoimmune diseases.** **Translation:** **the covering** of the mammary gland of pigs during the early development of the adult human is referred to as early microbial colonization. it is researched in medicine, microbiology and pathology. **early microbial colonization of the gut reduced the incidence of infectious**, inflammatory and **autoimmune diseases reduced the incidence of infectious, inflammatory and autoimmune diseases.**
SIMCSUM : die fruehzeitshygiene ist ein begriff aus der entwicklungsbiologie und bezeichnet das phaenomen, dass die fruehzeitliche besiedlung des darmes durch krankheitserreger verhindert wird. die fruehzeitshygiene unterscheidet sich von anderen entwicklungsbiologischen forschungsgebieten wie der entwicklungsphysiologie, der haematologie und der palaeontologie dadurch, dass in ihrer gesamtheit zur fruehen entwicklungsphase die mikrobielle vielfalt des darmes zaehlt. **Translation:** early hygiene is a term from developmental biology and describes the phenomenon that prevents early colonization of the intestines by pathogens. early hygiene differs from other developmental biological research areas such as developmental physiology, haematology and palaeontology in that the microbial diversity of the intestine counts in its entirety for the early development phase.

Table D.1: An example of SPEKTRUM output, where MBART produces non-German words (marked as red) and SIMCSUM generates the summary with wrong attention on hygiene. The summaries are translated via Google translate.

Target: nicht nur stress und genetische faktoren tragen zu essstörungen bei, sondern auch vorübergehende hungerphasen. nun stießen forscher vom europäischen zentrum für hirnforschung in rom auf einen möglichen grund dafür - und zwar in den gehirnen von labormäuse, die sich auch durch schmerzen nicht vom naschen abbringen ließen. die psychologin rossella ventura und ihre kollegen trainierten mäuse darauf, in einer von zwei kammern nach einer belohnung in form von schokolade zu suchen. dann versetzten sie den tieren ausgerechnet in dem raum, wo die leckerei lockte, schmerzhafte elektroschocks. die mäuse mieden fortan die schokolade und hielten sich vorwiegend in der anderen kammer auf - außer jene, die zuvor einige zeit lang kaum futter bekommen hatten. [...] **Translation:** Not only stress and genetic factors contribute to eating disorders, but also periods of starvation. now researchers from the european center for brain research in rome have discovered a possible reason for this - namely in the brains of laboratory mice, which were not deterred from snacking even by pain. psychologist rossella ventura and her colleagues trained mice to look for a chocolate reward in one of two chambers. then they gave the animals painful electric shocks in the room where the treat lured them. From then on, the mice avoided the chocolate and mainly stayed in the other chamber - except for those that had hardly received any food for a while. [...]
MBART: **eating disorders** (engl. “ess-stoerungen”) nennt man eine vielzahl von ursachen und mechanismen, die zur auspraegung von essstoerungen fuehren koennen. **Translation:** **eating disorders** are a number of causes and mechanisms that can lead to the development of eating disorders.
SIMCSUM : als essverhalten oder magersucht werden die verhaltensweisen von menschen und tieren bezeichnet, die **excessive** (vergoren) oder chronisch (z.b. nach unfaellen oder erkrankungen) zu es verlangen oder dieses verhalten unterdruecken. essverhalten kann zu verhaltensstoerungen fuehren, die ursache fuer die entstehung von essstoerungen sind innere oder aeussere ursachen des essverhaltens. **Translation:** eating habits or anorexia are the behaviors of people and animals that demand **excessive** (fermented) or chronic (e.g. after accidents or illnesses) or suppress this behavior. eating behavior can lead to behavioral disorders, the cause of eating disorders are internal or external causes of eating behavior.

Table D.2: An example of SPEKTRUM output, where MBART produces non-German words (marked as red) and quite short summary. While SIMCSUM generates also generates a non-German word. The summaries are translated via Google translate.

Target: der tod kam auf dem höhepunkt der macht : cangrandei. della scala, der berühmteste spross der scaliger aus verona, hatte in zahlreichen kriegerischen auseinandersetzungen die herrschaft über ganz venetien an sich gerissen. doch kurz nach seinem triumphalen einzug in die norditalienische stadt treviso am 18. juli 1329 begannen mit durchfall und erbrechen die ersten symptome einer tödlichen erkrankung. [...] **Translation:** death came at the height of power : cangrandei. della scala, the most famous offspring of the scaliger family from verona, seized control of all of veneto in numerous military conflicts. but shortly after his triumphant entry into the northern italian city of treviso on july 18, 1329, the first symptoms of a fatal illness began with diarrhea and vomiting. [...]
MBART: **alberto** cangrande (* 1291 oder 1329 in verona ; † 29. oktober 1329 ebenda) war ein italienischer dichter, schriftsteller und politiker der flämischen republik. in der zeit der italienischen republik bestieg er von seinen brudern die dynastie der scaligeri. bei der eroberung der elfenbeinküste 1311 wurde er zum maßgebenden gründer der faction in rom. ab 1314 war er herrscher im königreich venedig. cangrande gilt als vater der italienischen **Translation:** **alberto** cangrande (b. 1291 or 1329 in verona ; † october 29, 1329 ibid) was an italian poet, writer and politician of the flemish republic. in the time of the italian republic he ascended from his brothers the dynasty of the scaligeri. when he conquered the ivory coast in 1311, he became the main founder of the faction in rome. from 1314 he was ruler of the kingdom of venice. cangrande is considered the father of the italian
SIMCSUM : cangrande della scala oder cangrande **cangrande** (* 9. märz 1291 in verona ; † 22. juli 1329 ebenda) war ein begründer der von den scaligeri beeinflussten dynastie, der unter den herrschern von verona von 1308 bis 1387 regierte. der sohn des veronaer stadtbegründers und stadtmeisters alberto della scala (genannt „ albert “ oder „ adelsmann “), der nur als unterstellt war, wurde nach seinem tod 1301 zum bruder von alberto verbannte. **Translation:** cangrande della scala or cangrande **cangrande** (born march 9, 1291 in verona ; † july 22, 1329 ibid) was a founder of the dynasty influenced by the scaligeri, who reigned under the rulers of verona from 1308 to 1387. the son of verona's founder and mayor alberto della scala (called " albert " or " nobleman "), who was only subordinate, was banished to alberto's brother after his death in 1301.

Table D.3: An example of SPEKTRUM output, where MBART generates a wrong named entity. SPEKTRUM, on the other hand, gets it right but generates a wrong alias for this person. The summaries are translated via Google translate.

Target: inoffizieller weltrekordhalter im nicht - frieren trotz eisiger kälte ist der niederländer wimhof : er lief bereits ziemlich nackt halbmarathons nördlich des polarkreises und nahm das längste eisbad seit beginn der aufzeichnungen - ohne sich ernsthaft zu beklagen. wissenschaftliche untersuchungen des eismannes hatten bisher keine körperlichen auffälligkeiten gefunden, die erklären könnten, warum hof mit kälte offensichtlich besser umgeht als der durchschnittsmensch. [...] **Translation:** the unofficial world record holder for not freezing despite the icy cold is the dutchman wimhof : he ran half marathons north of the arctic circle quite naked and took the longest ice bath since records began - without seriously complaining. scientific examinations of the ice cream man had not found any physical abnormalities that could explain why hof obviously deals with the cold better than the average person. [...]
MBART: wim hof (* 20. april 1959) ist ein niederländischer sportler, der sich unter anderem **mit eisschwimmern beschäftigt**. **Translation:** wim hof (born april 20, 1959) is a dutch athlete who is **involved with ice swimmers**, among other things.
SIMCSUM : wim hof (* 20. april 1959), auch als eismann bekannt, ist ein niederländischer **swimmer und ein ehemaliger olympiateilnehmer**. er hält seit dem **28. februar 2015** den bislang längsten direkten kontakt menschlichen körpers mit eis. **Translation:** Wim Hof (born April 20, 1959), also known as Eismann, is a Dutch **swimmer and a former Olympic competitor**. **since february 28, 2015**, he has been in the longest direct contact between the human body and ice.

Table D.4: An example of SPEKTRUM output, where both MBART and SIMCSUM produce unfaithful information. Marked in orange is unfaithful information to the original text. The summaries are translated via Google translate.

E SIMCSUM Examples and Results

Here we present some examples showing the difference between SIMCSUM and MBART.

Target: für menschen ist der einbeinige stand immer eine wackelpartie , vor allem bei geschlossenen augen . um nicht umzukippen , müssen die muskeln permanent die leichten schwankungen ausgleichen . nicht so bei flamingos : sie kostet es weniger kraft , auf einem bein zu stehen als auf zweien . deswegen können sie auch beruhigt auf einem bein ein schläfchen machen , ohne dabei umzukippen . [...] **Translation:** standing on one leg is always a shaky game for humans, especially with closed eyes. in order not to tip over , the muscles have to constantly compensate for the slight fluctuations . Not so with flamingos: it takes less strength for them to stand on one leg than on two. that's why you can take a nap on one leg without tipping over. [...]
MBART: **die biogerontologie (von griech. bíos „ leben " und lat. protes „ befestigt ") beschäftigt sich mit dem verhältnis von körpergewicht und körperhaltung. so untersucht die biogerontologie das verhältnis von körpergewicht und körperhaltung : welche gewichtszustände und welche muskeln notwendig sind, um einen pilz, der einer krankheit ausgesetzt ist, zu tragen? wie sehr wichtig es ist, den jeweiligen körpergewichtszustand zu messen. dies wird in der** **Translation:** **biogerontology (from greek bíos "life" and lat. protes "fixed") deals with the relationship between body weight and posture. this is how biogerontology examines the relationship between body weight and posture: which weight conditions and which muscles are necessary to carry a fungus that is exposed to a disease? how very important it is to measure the respective body weight condition. this will be in the**
SIMCSUM : flamingos (phoenicopteridae) oft sitzen auf einem bein, um eine muskelkontraktion zu erzeugen. sie haben die fähigkeit, das körpergewicht aufzunehmen und zu tragen, ohne dabei eine muskelaktive wirkung auszuüben. **Translation:** Flamingos (phoenicopteridae) often perch on one leg to produce muscle contraction. they have the ability to absorb and carry body weight without exerting a muscle-active effect.

Table E.1: An example of SPEKTRUM output, where SIMCSUM generates a better summary than MBART . In this case, the MBART summary misses the article's point by focusing on biogerontology which has no relevancy in this case, and SIMCSUM generates a summary that better captures the idea and main focus of the original text. Marked in orange is unfaithful information to the original text. The summaries are translated via Google translate.

Target: eine besondere strategie, seinen fressfeinden zu entkommen, hat der bombardierkäfer pheropsophus jessoensis entwickelt. er wehrt sich sogar noch, nachdem er eigentlich schon verloren hat, und das sogar recht erfolgreich, wie jetzt shinji sugiura und takuya sato von der universität kobe in den biology letters berichten. gerade verschluckt, schießt der käfer im magen der kröte einen strahl heißes, ätzendes sekret aus der afterdrüse - eine strategie, die den bombardierkäfern ihren namen eingebracht hat. normalerweise allerdings dient die waffe zur abschreckung. in diesem fall jedoch ätzen sich die käfer mit ihrer hilfe quasi zurück in die freiheit : das sekret lässt die kröten erbrechen ; und der eigentlich schon verspeiste käfer entkommt. [...] **Translation:** the bombardier beetle pheropsophus jessoensis has developed a special strategy for escaping its predators. he even defends himself after he has actually already lost, and even quite successfully, as shinji sugiura and takuya sato from the university of kobe now report in the biology letters. just swallowed, the beetle in the toad's stomach shoots a jet of hot, caustic secretion out of the anal gland - a strategy that has given the bombardier beetle its name. normally, however, the weapon serves as a deterrent. in this case, however, the beetles etch their way back to freedom with your help: the toads vomit the secretion; and the actually already eaten beetle escapes. [...]
MBART: **unter zoologischer kompatibilitaet versteht man in der biologie die faehigkeit von beutetieremischlingen, nach einem laubabschweif von artgenossen in den verdauungstrakt von artgenossen zu entweichen. die oekologische kompatibilitaet** ist ein wesentlicher bestandteil der erfolgreichen entstehung von beutegreifern.beim laubabschweifexperiment mit hilfe des insektenpheropsophus jessoensis aus zwei kaeferarten, dem bufo japonicus und dem b. torrenticola, konnte nachgewiesen werden, dass der kaefer pheropsophus jessoensis unter laboratorischen bedingungen eine in den magensaftgastropfen der jagdsekten einschleudern kann. **Translation:** in biology, **zoological compatibility means the ability of hybrid prey animals to escape into the digestive tract of conspecifics after the leaves have drifted away from conspecifics. ecological compatibility** is an essential part of the successful emergence of predators.in the leaf-dropping experiment with the help of the insect pheropsophus jessoensis from two beetle species, the bufo japonicus and the b. torrenticola, it could be proven that the beetle pheropsophus jessoensis can inject one into the gastric juice droplets of the hunting sects under laboratory conditions.
SIMCSUM : als wehrverhalten oderverteidigungsverhalten bezeichnet man in der verhaltensbiologie verschiedene verhaltensweisen von tieren, die wehren gegen die fressfeinde oder andere fresstiere bilden. nicht zu verwechseln ist das wehrverhalten **mit der phylogenetik, die sich aus der evolutionsgeschichte entwickelt hat.** **Translation:** in behavioral biology, defensive behavior is the term used to describe various behaviors of animals that defend themselves against predators or other predators. the defensive behavior should not be confused with **the phylogenetics that have developed from the history of evolution.**

Table E.2: An example of SPEKTRUM output, both SIMCSUM and MBART produce reasonable summaries. The SIMCSUM summary is preferred because it explains the basic idea of the beetle defending itself against predators. The MBART summary is filled with complex words that do not make sense or are unnecessary for this context. These words are marked in orange. The summaries are translated via Google translate.

Target: manche farben bleiben besser im gedächtnis haften als andere, doch auf die erinnerung an den entsprechend gefärbten gegenstand hat das keinen einfluss. das zeigt nun eine studie von psychologen um christof kuhbandner von der universität regensburg. die ergebnisse machen deutlich, dass der gegenstand selbst und dessen eigenschaften vom gehirn nicht immer als einheit betrachtet werden. die wissenschaftler testeten für ihre studie das gedächtnis ihrer versuchspersonen. dabei variierten sie sowohl, was zu merken war, als auch, wie es dargeboten wurde. insbesondere präsentierten sie wörter und gegenstände in unterschiedlichen farben. [...] **Translation:** some colors stick in the memory better than others, but this has no influence on the memory of the correspondingly colored object. this is now shown by a study by psychologists led by christof kuhbandner from the university of regensburg. the results make it clear that the object itself and its properties are not always viewed as a unit by the brain. for their study, the scientists tested the memory of their subjects. they varied both what was to be remembered and how it was presented. in particular, they presented words and objects in different colors. [...]
MBART: die farbpsychologie ist ein teilgebiet der psychologie und beschäftigt sich mit der wirkung von farben auf die wahrnehmung und das gedächtnis. **Translation:** color psychology is a branch of psychology and deals with the effect of colors on perception and memory.
SIMCSUM : die farbe ist ein farbmuster, das uns wahrnehmungen und die sinneswahrnehmung durch das sehen vermittelt. die farbwahrnehmung gehört zu den kognitionswissenschaften. **wir sehen sie nicht nur als eindimensionales spektrum möglicher farben, sondern auch als eindimensionales, oftmals auch als „ visuelles system “, das die sinneswahrnehmung mitanpasst.** [...]
Translation: color is a color pattern that conveys perceptions and sensory perception to us through seeing. Color perception is one of the cognitive sciences. **we see them not only as a one-dimensional spectrum of possible colors, but also as a one-dimensional, often also as a " visual system " that also adapts the sensory perception.** [...]

Table E.3: An example of SPEKTRUM output, where MBART performs better than SIMCSUM . MBART generates a summary that is too short but which better recapitulates the main idea. The orange marked words in the SIMCSUM summary are incoherent and are not faithful to the original text. The summaries are translated via Google translate.