X-SCITLDR: Cross-Lingual Extreme Summarization of Scholarly Documents

Sotaro Takeshita
University of Mannheim
Mannheim, Germany
sotaro.takeshita@uni-mannheim.de

Tommaso Green
University of Mannheim
Mannheim, Germany
tommaso.green@uni-mannheim.de

Niklas Friedrich
University of Mannheim
Mannheim, Germany
nfriedri@mail.uni-mannheim.de

Kai Eckert
Hochschule der Medien
Stuttgart, Germany
eckert@hdm-stuttgart.de

Simone Paolo Ponzetto
University of Mannheim
Mannheim, Germany
ponzetto@uni-mannheim.de

Abstract
The number of scientific publications nowadays is rapidly increasing, causing information overload for researchers and making it hard for scholars to keep up to date with current trends and lines of work. Consequently, recent work on applying text mining technologies for scholarly publications has investigated the application of automatic text summarization technologies, including extreme summarization, for this domain. However, previous work has concentrated only on monolingual settings, primarily in English. In this paper, we fill this research gap and present an abstractive cross-lingual summarization dataset for four different languages in the scholarly domain, which enables us to train and evaluate models that process English papers and generate summaries in German, Italian, Chinese and Japanese. We present our new X-SCITLDR dataset for multilingual summarization and thoroughly benchmark different models based on a state-of-the-art multilingual pre-trained model, including a two-stage `summarize and translate' approach and a direct cross-lingual model. We additionally explore the benefits of intermediate-stage training using English monolingual summarization and machine translation as intermediate tasks and analyze performance in zero- and few-shot scenarios.

CCS Concepts
• Computing methodologies → Natural language processing; Natural language generation; Language resources.

Keywords
Scholarly document processing, Summarization, Multilinguality

1 Introduction
For years, the number of scholarly documents has been steadily increasing [4], thus making it difficult for researchers to keep up to date with current publications, trends and lines of work. Because of this problem, approaches based on Natural Language Processing (NLP) have been developed to automatically organize research papers so that researchers can consume information in ways more efficient than just reading a large number of papers. For instance, citation recommendation systems provide a list of additional publications given an initial ‘seed’ paper, in order to reduce the burden of literature reviewing [3, 37]. One approach is to identify relevant sentences in the paper based on automatic classification [24]. This approach to information distillation is taken further by fully automatic text summarization, where a long document is used as input to produce a shorter version of it covering essential points [10, 57], possibly a TLDR-like ‘extreme’ summary [5]. Similar to the case of manually-created TLDRs, the function of these summaries is to help researchers quickly understand the main content of a paper without having to look at the full manuscript or even the abstract.

Just like in virtually all areas of NLP research, most successful approaches to summarization rely on neural techniques using supervision from labeled data. This includes neural models to summarize documents in general domains such as news articles [33, 49], including cross- and multi-lingual models and datasets [48, 52], as well as specialized ones e.g., the biomedical domain [39].

For the task of summarizing research papers, most available datasets are in English only, e.g., CSPubSum/CSPubSumExt [10] and ScisummNet [57], with community-driven shared tasks also having concentrated on English as de facto the only language of interest [6, 23]. But while English is the main language in most of the research communities, especially those in the science and technology domain, this limits the accessibility of summarization technologies for the researchers who do not use English as the main language (e.g., many scholars in a variety of areas of humanities and social and political sciences). We accordingly focus on the problem of cross-lingual summarization of scientific articles – i.e., produce summaries of research papers in languages different than the one of the original paper – and benchmark the ability of state-of-the-art multilingual transformers to produce summaries for English research papers in different languages. Specifically, we propose the new task of cross-lingual extreme summarization of scientific papers (CL-TLDR), since TLDR-like summaries have
shown much promise in real-world applications such as search engines for academic publications like Semantic Scholar\(^1\).

In order to evaluate the difficulty of CL-TLDR and provide a benchmark to foster further research on this task, we create a new multilingual dataset of TLDRs in a variety of different languages (i.e., German, Italian, Chinese, and Japanese). Our dataset consists of two main portions: a) a translated version of the original dataset from Cachola et al. [5] in German, Italian and Chinese to enable comparability across languages on the basis of post-edited automatic translations; b) a dataset of human-generated TLDRs in Japanese from a community-based summarization platform to test performance on a second, comparable human-generated dataset. Our work complements seminal efforts from Fatima and Strube who compile an English-German cross-lingual dataset from the Spektrum der Wissenschaft / Scientific American and Wikipedia, in that we focus on extreme summarization, build a dataset of expert-derived multilingual TLDRs (as opposed to leads from Wikipedia), and provide additional languages.

The contributions of this work are as follows.

- We propose the new task of cross-lingual extreme summarization of scientific articles (CL-TLDR).
- We create the first multilingual dataset for extreme summarization of scholarly papers in four different languages.
- Using our dataset, we benchmark the difficulty of cross-lingual extreme summarization with different models built on top of state-of-the-art pre-trained language models [29, 32], namely a two-step approach based on monolingual summarization and translation, and a multilingual transformer model to directly generate summaries cross-lingually.
- Our dataset may not be sufficient to successfully fine-tune large pre-trained language models for direct cross-lingual summarization. Consequently, to tackle this data scarcity problem, we additionally present experiments using intermediate fine-tuning techniques, which have shown to be effective to improve performance of pre-trained multilingual language models on many downstream NLP tasks [18, 20, 42, 43, inter alia].

The remainder of this paper is organized as follows. We first summarize in Section 2 seminal work on monolingual extreme summarization for English from Cachola et al. [5], on which our multilingual extension builds upon. We next introduce our new dataset for cross-lingual TLDR generation in Section 3. We present our cross-lingual models and benchmarking experiments in Section 4 and 5, respectively. Section 6 provides an overview of relevant previous work in monolingual and multilingual summarization. We wrap up our work with concluding remarks and directions for future work in Section 7.

### 2 English Monolingual TLDR Summarization

The main part of our new cross-lingual dataset consists of an automatically-translated, post-edited version of the SCITLDR dataset from Cachola et al. [5], who presented seminal work on the topic of extreme summarization of scientific publications for English.

SCITLDR is a dataset composed of pairs of research papers and corresponding summaries: in contrast to other existing datasets, this dataset is unique because of its focus on extreme summarization, i.e., very short, TLDR-like summaries, and consequently high compression ratios – cf. the compression ratio of 238.1% of SCITLDR vs. 36.5% of CLPubSum [10]. An example of a TLDR summary is presented in Table 1, where we see how information from different summary-relevant sections of the paper (typically, in the abstract, introduction and conclusions) is often merged to provide a very short summary that is meant to help readers quickly understand the key message and contribution of the paper.

The original SCITLDR dataset consists of 5,411 TLDRs for 3,229 scientific papers in the computer science domain: it is divided into a training set of 1,992 papers, each with a single gold-standard TLDR, and dev and test sets of 619 and 618 papers each, with 1,452 and 1,967 TLDRs, respectively (thus being multi-target in that a document can have multiple gold-standard TLDRs). The summaries consist of TLDRs written by authors and collected from the peer review platform OpenReview\(^2\), as well as human-generated summaries from peer-review comments found on the same platform. While this dataset encourages the development of scholarly paper summarization systems, the original version only supports English as the target language, even though research across different fields can be conducted by researchers who use various languages from all over the world. Therefore, in this work, we extend the summaries in SCITLDR to support four additional languages, namely German, Italian, Chinese and Japanese.

### 3 X-SCITLDR

In this section, we describe the creation of the X-SCITLDR dataset and briefly present some statistics to provide a quantitative overview. Our dataset is composed of two main sources:

- a manually post-edited version of the original SCITLDR dataset [5] for German, Italian and Chinese (X-SCITLDR-PostEdit).
- a human-generated dataset of expert-authored TLDRs harvested from a community-based summarization platform for Japanese (X-SCITLDR-Human).

---

\(^1\)https://www.semanticscholar.org/product/tldr

\(^2\)https://openreview.net
The paper presents a multi-view framework for improving sentence representation in NLP tasks using generative and discriminative objective architectures.

Das Papier präsentiert einen Multi-View-Rahmen zur Verbesserung der Satzrepräsentation in NLP-Aufgaben...

Der Artikel präsentiert einen Multi-View-Rahmen zur Verbesserung der Satzrepräsentation in NLP-Aufgaben...

The paper provides a full characterization of permutation invariant and equivariant linear layers for graph data.

L'articolo fornisce una caratterizzazione completa degli strati lineari invarianti di permutazione ed equivarianti per i dati del grafico.

L'articolo fornisce una caratterizzazione completa dei layer lineari invarianti o equivarianti per la permutazione per i dati del grafo.

The paper proposes a framework for constructing spherical convolutional networks based on a novel synthesis of several existing concepts.

Das Papier schlägt einen Rahmen für die Konstruktion von sphärischen Faltungsnetzen vor, der auf einer neuartigen Synthese mehrerer bestehender Konzepte beruht.

Die Arbeit schlägt einen Rahmen für die Konstruktion von sphärischen Convolutional Networks vor, der auf einer neuartigen Synthese mehrerer bestehender Konzepte beruht.

We present a novel iterative algorithm based on generalized low rank models for computing and interpreting word embedding models.

Presentiamo un nuovo algoritmo iterativo basato su modelli generalizzati di basso rango per il calcolo e l'interpretazione dei modelli di incorporazione delle parole.

Presentiamo un nuovo algoritmo iterativo basato su modelli generalizzati di basso rango per il calcolo e l'interpretazione dei modelli di word embedding.

Table 2: Example of a post-editing correction (wrong sense): 'Papier' means a generic piece of paper but not a research paper in German ('Artikel'). Similarly, English 'graph' needs to be translated as 'grafo' as opposed to 'grafico' (English: 'diagram').

Table 3: Example of a post-editing correction (terminological English-preserving translation). 'Convolutional network' can be translated in German as 'faltendes Netz' or 'Faltungsnetz', whereas 'word embedding' can be translated as both 'incorporazione' or 'immersione delle parole' in Italian. We reduce variability in summaries by keeping the English domain-specific term in the target-language summaries.

X-SCITLDR-PostEdit. Given the overall quality of automatic translators [13], we opt for a hybrid machine-human translation process of post-editing [19] in which human annotators correct machine-generated translations as post-processing to achieve higher quality than when only using an automatic system. Although current machine translation systems arguably provide nowadays high-quality translations, a manual correction process is still necessary for our data, especially given their domain specificity. In Tables 2 and 3, we present examples of how translations are corrected by human annotators, and the reasons for the correction. These can be grouped into two cases:

a) Wrong translation due to selected wrong sense (Table 2). In this case, the machine translation system has problems selecting the domain-specific sense and translation of the source term.

b) Translation of technical terms (Table 3). To avoid having the same technical term being translated in different ways, we reduce the sparsity of the translated summaries and simplify the translation task by preserving technical terms in English.

Both cases indicate the problems of the translation system with domain-specific terminology. For the underlying translation system, we use DeepL. After the automatic translation process, we asked graduate students in computer science courses who are native speakers in the target language to fix incorrect translations.

X-SCITLDR-Human. We complement the translated portion of the original TLDR dataset with a new dataset in Japanese crawled from the Web. For this, we collect TLDRs of scientific papers from

3https://www.deepl.com/translator
a community-based summarization platform, arXivTimes\textsuperscript{4}. This Japanese online platform is actively updated by users who voluntarily add links to papers and a corresponding user-provided short summary. The posted papers cover a wide range of machine learning related topics (e.g., computer vision, natural language processing and reinforcement learning). This second dataset portion allows us to test with a dataset for extreme summarization of research papers in an additional language and, crucially, with data entirely written by humans, which might result in a writing style different from the one in X-SCITLDR-PostEdit. That is, we can use these data not only to test the capabilities of multilingual summarization in yet another language but, more importantly, test how much our models are potentially overfitting by too closely optimizing to learn the style of the X-SCITLDR-PostEdit summaries.

In Table 4, we present various statistics of our X-SCITLDR dataset for both documents and summaries from the original English (EN) SCITLDR data\textsuperscript{5} and our new dataset in four target languages\textsuperscript{6}. SCITLDR and X-SCITLDR-PostEdit (DE/IT/ZH) have a comparably high compression ratio (namely, the average number of words per document to the average number of words per summary) across all four languages, thus indeed requiring extreme cross-lingual compression capabilities. While summaries in German, Italian and Chinese keep the compression ratio close to the original dataset in English, summaries in the Japanese dataset come from a different source and consequently exhibit rather different characteristics, most notably longer documents and summaries. Manual inspection reveals that Japanese documents come from a broader set of venues than SCITLDR, since arXivTimes includes many ArXiv, ACL and OpenReview manuscripts (in contrast to SCITLDR, whose papers overwhelmingly come from ICLR, cf. [5, Table 9]), whereas Japanese summaries often contain more than one sentence. Despite having both longer documents and summaries, the Japanese data still exhibit a high compression ratio (cf. datasets for summarization of both scientific and non-scientific documents having typically a compression ratio <40%), which indicates their suitability for evaluating extreme summarization in the scholarly domain.

### 4 Models for Cross-lingual TLDR

We next present a variety of models that we use to benchmark the feasibility and difficulty of the task of cross-lingual extreme summarization of scientific papers (henceforth: CL-TLDR). Our cross-lingual models are able to automatically generate summaries in a target language given abstracts in English. For this, we build upon the original work from [5] and focus on abstractive summarization, since this has been shown to outperform extractive summarization in a variety of settings.

**BART / mBART.** In our experiments, we use BART [29] and its multilingual variant mBART [32] as underlying summarization models. They are both transformer-based [53] pre-trained generative language models, which are trained with an objective to reconstruct noised text in an unsupervised sequence-to-sequence fashion. While BART only uses an English corpus for pre-training, mBART learns from a corpus containing multiple languages. These pre-trained BART/mBART models can be further trained (i.e., fine-tuned) in order to be applied to downstream tasks of interest like, for instance, summarization, translation or dialogue generation. We use BART/mBART as our underlying models, since these have been shown in previous work to perform well on the task of extreme summarization [29]. We follow Ladhak et al. [27] and use BART/mBART as components of two different architectures, namely: a) two-step approach to cross-lingual summarization, i.e., summarization via BART and translation using Machine Translation (MT) (Section 4.1); b) a direct cross-lingual summarization system obtained by fine-tuning mBART with input articles from English, and summaries from the target language (Section 4.2).

#### 4.1 Two-stage Cross-lingual Summarization: Summarize and Translate

A first solution to the CL-TLDR task is to combine a monolingual summarization model with machine translation in the target language (we call this approach EnSum-MT). This approach is composed of two stages. The model first takes an English text as input and generates a summary in English: the English summary is then automatically translated into the target language using machine translation (Figure 1).\textsuperscript{7} This model does not rely on any cross-lingual signal: it merely consists of two independent modules for translation and summarization and does not require any

\textsuperscript{4}https://arxivtimes.herokuapp.com

\textsuperscript{5}Slight differences with respect to the statistics from Cachola et al. [5, Table 1], e.g., different average number of words per summary (21 vs. 23.88), are due to a different tokenization (we use SpaCy: https://spacy.io).

\textsuperscript{6}Vocabulary sizes are computed after lemmatization with SpaCy.

\textsuperscript{7}Like for the creation of the multilingual portion of our dataset, we opt again for DeepL for all our languages (cf. Section 3).
A second approach to CL-TLDR is to directly perform cross-lingual summarization. We call this method X-SCITLDR: Cross-Lingual Extreme Summarization of Scholarly Documents. Our experiments, we investigate two strategies for intermediate task and cross-lingual fine-tuning. Based on this observation, in our experiments, we inject additional texts in a domain/task similar to the target domain/task can boost the performance on the downstream task by injecting additional summarization data.

Cross-lingual dataset to train the summarization model. While this system is conceptually simple, such a pipeline approach is known to cause an error propagation problem \[60\], since errors of the first stage (i.e., summarization) get amplified in the second stage (i.e., translation) leading to overall performance degradation.

4.2 Direct Cross-lingual Summarization

A second approach to CL-TLDR is to directly perform cross-lingual summarization using a pre-trained multilingual language model (we call this method CLSum). For this, we investigate the use of a pre-trained multilingual denoising autoencoder, namely mBART \[32\], and use cross-lingual training data from our new X-SCITLDR dataset to fine-tune mBART and generate summaries in the target languages given abstracts in English, as depicted in Figure 2. We follow Liu et al. \[32\] and control the target language by providing a language token to the decoder.

Intermediate task and cross-lingual fine-tuning. Our training dataset is relatively small compared to datasets for general domain summarization \[21, 40\]. To mitigate this data scarcity problem, we investigate the effectiveness of intermediate fine-tuning, which has been reported to improve a wide range of downstream NLP tasks (see, among others, \[18, 20, 42, 43\]). Gururangan et al. \[20\], for instance, show that training pre-trained language models on texts in a domain/task similar to the target domain/task can boost the performance on the downstream task by injecting additional related knowledge into the models. Based on this observation, in our experiments, we investigate two strategies for intermediate fine-tuning: intermediate task and cross-lingual fine-tuning.

- **Intermediate task fine-tuning (CLSum+EnSum).** We explore the benefits of using additional summarization data other than the summaries in the target language and augment the training dataset with English data, i.e., the original SCITLDR data. That is, before fine-tuning on summaries in the target language (e.g., German), we train the model on English TLDR summarization as auxiliary monolingual summarization task to provide additional summarization capabilities (Figure 3).

- **Cross-lingual intermediate fine-tuning (CLSum+MT).** Direct cross-lingual summarization requires the model to learn both translation and summarization skills, arguably a difficult task given our small dataset. To alleviate this problem, we investigate a model that is trained on machine translation, before being fine-tuned on the summarization task. For this, we automatically translate the English abstracts into the target language and use these synthetic data as training data for fine-tuning on the task of automatically translating abstracts (Figure 4).

5 Experiments

**Input documents.** We follow Cachola et al. \[5\] and rely in all our experiments on an input consisting of abstracts only, since they showed that it yields similar results when compared to using the abstract, introduction, and conclusion sections together. Even more importantly, using only abstracts enables the applicability of our models also to those cases where only the abstracts are freely available and we do not have open access to the complete manuscripts. The average length of an abstract is 185.87 words for X-SCITLDR-PostEdit (EN/DE/IT/ZH) with an average compression ratio of 12.64%, and 190.92 words for X-SCITLDR-Human (Japanese) and a compression ratio of 39.76%.

**Evaluation metrics.** We compute performance using a standard metric to automatically evaluate summarization systems, namely ROUGE \[31\]. In the case of the post-edited portion of the X-SCITLDR dataset (X-SCITLDR-PostEdit, Section 3), the gold standard can contain multiple reference summaries for a given paper and abstract. Consequently, for Italian, German and Chinese we calculate ROUGE scores in two ways (\(\text{avg}\) and \(\text{max}\)) to account for these multiple references \[5\]. For \(\text{avg}\), we compute ROUGE F1 scores with respect to the different references and take the average score, whereas for \(\text{max}\) we select the highest scoring one. The Japanese dataset does not contain multiple reference TLD technologies; hence, we compute standard ROUGE F1 only. We test for statistical significance using sample level Wilcoxon signed-rank test with \(\alpha = 0.05\) \[14\].

**Hyperparameter tuning.** To find the best hyperparameters for each model, we use the development data and run a grid search using ROUGE-1 \(\text{avg}\) as a reference metric. We run experiments with learning rate \(\epsilon \in \{1 \cdot 10^{-5}, 3 \cdot 10^{-5}\}\) and random seed \(\in \{1122, 22\}\) during fine-tuning, number of beams for beam search \(\in \{2, 3\}\) and repetition penalty rate \(\in \{0.8, 1.0\}\) during decoding. For all settings, we set the batch size equal to 1 and perform 8 steps of gradient accumulation. We use the AdamW optimizer \[34\] with weight decay of 0.01 for 5 epochs without warm-up.

**Training strategy.** To prevent the model from losing the knowledge acquired in pre-training during fine-tuning (i.e., catastrophic forgetting \[26\]), we freeze the parameters of the embedding and decoder layers during intermediate task and cross-lingual fine-tuning \[36\], while we update the entire model during the final fine-tuning for downstream CL-TLDR. Since mBART requires large memory space when we update the entire model, we utilize the DeepSpeed library \[9\] to meet our infrastructure requirements.

**Research questions.** We organize the presentation and discussion of our results around the following research questions:

1. In preliminary experiments mBART often failed to generate in the target language after fine-tuning it on our cross-lingual dataset, thus confirming the need to augment with translation data (see also previous findings from Ladhak et al. \[27\]).
2. [https://www.deepspeed.ai](https://www.deepspeed.ai)
Due to the success of deep learning, i.e., different portions of our dataset (X-SCITLDR-PostEdit vs. -Human, Section 3), the differences in performance figures between X-SCITLDR-PostEdit (German, Italian and Chinese) and X-SCITLDR-Human (Japanese) are due to the different nature of the multilingual data, and how they were created. Post-edited data like those in German, Italian and Chinese are indeed automatically translated, and tend to better align to the also automatically translated English summaries, as provided as output of the EnSum-MT system. That is, since both summaries - the post-edited reference ones and those automatically generated and translated - go through the same process of automatic machine translation, they naturally tend to have a higher lexical overlap, i.e. a higher overlap in terms of shared word sequences. This, in turn, receives a higher reward from ROUGE, since this metric relies on n-gram overlap between system and reference summaries. While two-stage cross-lingual summarization seems to better align with post-edited reference TLDRs, for human-generated Japanese summaries, we observe an opposite behavior. Japanese summaries indeed have a different style than those in English (and their post-edited multilingual versions from X-SCITLDR-PostEdit) and accordingly have a lower degree of lexical overlap with translated English summaries from EnSum-MT.

To better understand the behavior of the system in light of the different performance on post-edited vs. human-generated data, we manually inspected the output of the two systems. Table 7 shows an example of automatically generated summaries for a given input abstract: it highlights that summaries generated using our cross-lingual models (CLSum) tend to be shorter and consequently ‘abstracter’ than those created by translating English summaries (EnSum-MT). This, in turn, can hurt the performance of the cross-lingual models more in that, while we follow standard practices and used ROUGE F1, this metric has been found unable to address the problems with ROUGE recall, which rewards longer summaries, in the ranges of typical summary lengths produced by neural systems [51]. Table 8 presents the average summary lengths in different languages for our MT-based and cross-lingual systems: the numbers show that for languages in the X-SCITLDR-PostEdit portion of our dataset summaries are indeed shorter. Japanese cross-lingually
Table 5: Results on the X-SCITLDR-PostEdit portion of our cross-lingual TLDR dataset (ROUGE-1/-2 and -L): English to German, Italian or Chinese TLDR-like summarization using post-edited, automatically-translated summaries of the English data from Cachola et al. [5]. Best results per language and metric are bolded. Statistically significant improvements of the cross-lingual models (CLSum/+EnSum/+MT) with respect to the ‘summarize and translate’ pipeline (EnSum-MT) are marked with †.

Table 6: Results on the X-SCITLDR-Human portion of our cross-lingual TLDR dataset (Rouge-1/-2 and -L): English to Japanese TLDR-like summarization using human-generated summaries from ArXivTimes. Best results per metric are bolded. Statistically non-significant improvements of the cross-lingual models (CLSum/+EnSum/+MT) with respect to EnSum-MT are marked with an asterisk (∗).

Table 7: Example of gold-standard summaries and automatically generated versions.

Manual inspection reveals that German summaries tend to be penalized more because of differences in word compounding between reference and generated summaries: while there exist proposals to...
address this problem in terms of language-specific pre-processing [17], we opt here for a standard evaluation setting equal for all languages. Moreover, German summaries tend to contain less English terms than, for instance, Italian summaries (6.78 vs. 4.88 English terms per summary on average in the test data), which seems to put the cross-lingual model at an advantage (cf. English terminology in EnSum-MT vs. CL-Sum in Table 7). The performance gap between EnSum-MT and CLSum is the largest on the Chinese dataset, which shows that it is more challenging for mBART to learn to summarize from English into a more distant language [28].

**RQ2: The potential benefits of intermediate fine-tuning.** In the second set of experiments, we compare the performance of our ‘base’ cross-lingual model with those using intermediate-stage training for learning the summarization and translation tasks from additional data. Specifically, we compare target-language fine-tuning of mBART (CLSum) with additional intermediate task fine-tuning on English monolingual summarization (CLSum+EnSum) and cross-lingual intermediate fine-tuning using machine translation on synthetic data (CLSum+MT). The rationale behind these experiments is that in the direct cross-lingual setting the model needs to acquire both summarization and translation capabilities, which requires a large amount of cross-lingual training data, and thus might be hindered by the limited size of our dataset.

Including additional training on summarization based on English data (CLSum+EnSum) has virtually no effects on the translated portion of SCITLDR (Table 5) for German, and even degrades performance on Italian and Chinese. This is likely because English TLDR summaries are well aligned with their post-edited translations and virtually bring no additional signal while requiring the decoder to additionally translate into an additional language (i.e., English and the target language). On the contrary, CLSum+EnSum improves on all ROUGE metrics for Japanese (Table 6): this is because, as previously mentioned, the ArXivTimes data have a different style from SCITLDR and thus English TLDRs provide an additional training signal that help to improve results for the summarization task.

Including MT-based pre-training, i.e., fine-tuning mBART on machine translation from English into the target language, and then on cross-lingual summarization (CLSum+MT) improves over simple direct cross-lingual summarization (CLSum) on all languages – a finding in line with results from Ladkh et al. [27] for WikiHow summarization. This highlights the importance of fine-tuning the encoder-decoder for translation before actual fine-tuning for the specific cross-lingual task, thus injecting general translation capabilities into the model.

**RQ3: Zero- and few-shot experiments.** To better understand the contribution of intermediate fine-tuning and to analyze performance in the absence of multilingual summarization training data (i.e., in zero-shot settings), we present experiments in which we compare: a) using mBART with no fine-tuning (CLSum); b) fine-tuning mBART on English SCITLDR data only and evaluate performance on X-SCITLDR in our four languages; c) fine-tuning mBART on synthetic translations of abstracts only and testing on X-SCITLDR. These experiments are meant to quantify the zero-shot cross-lingual transfer capabilities of the cross-lingual models (i.e., can we train on English summarization data only without the need of a multilingual dataset?) as well as to explore how much we can get away with summarization data at all (i.e., what is the performance of a system that is trained to simply translate abstracts?).

We present our results in Table 10. The performance figures indicate that the zero-shot cross-lingual transfer performance of CLSum+EnSum is extremely low for all our four languages, with reference performance on English TLDRs from Cachola et al. [5] being 31.1/10.7/24.4 R1/R2/RL (cf. Table 10, BART ’abstract-only’), and barely improves over no fine tuning at all (CLSum). This suggests that robust cross-lingual transfer in our summarization task is more difficult than in other language understanding tasks (see for example the much higher average performance on the XTREME tasks [22]). The overall very good performance of CLSum+MT seems to suggest that robust cross-lingual summarization performance can still be achieved without multilingual summarization data through the ‘shortcut’ of fine-tuning a multilingual pre-trained model to translate English abstracts, since these can indeed be seen as summaries (albeit of a longer length than our TLDR-like summaries) and can thus be merely translated so as to provide a strong baseline.

### Table 8: Average summary length (number of tokens).

| Language | EnSum-MT | CLSum |
|----------|----------|-------|
| DE       | 23.48    | 22.94 |
| IT       | 24.17    | 22.73 |
| ZH       | 25.90    | 19.76 |
| JA       | 30.50    | 56.76 |

### Table 9: Word-level Jaccard coefficients between automatically translated summaries and their post-edited versions.

| Language | Train | Val | Test |
|----------|-------|-----|------|
| DE       | 0.95  | 0.92| 0.92 |
| IT       | 0.79  | 0.78| 0.78 |
| ZH       | 0.96  | 0.95| 0.94 |

### Table 10: Performance in zero-shot settings. no intermediate fine-tuning (CLSum) vs. intermediate task (+EnSum) and cross-lingual fine-tuning (+MT).

| Lang | Model          | R1 (avg) | R2 (avg) | RL (avg) |
|------|----------------|----------|----------|----------|
| DE   | CLSum          | 2.67     | 0.46     | 2.58     |
|      | CLSum+EnSum    | 3.46     | 0.70     | 3.32     |
|      | CLSum+MT       | 14.42    | 2.04     | 10.75    |
| IT   | CLSum+EnSum    | 4.83     | 0.97     | 4.41     |
|      | CLSum+MT       | 16.11    | 3.48     | 12.38    |
| ZH   | CLSum+EnSum    | 0.79     | 0.10     | 0.76     |
|      | CLSum+MT       | 17.88    | 3.60     | 13.95    |
| JA   | CLSum+EnSum    | 2.37     | 0.68     | 2.17     |
|      | CLSum+MT       | 29.43    | 4.29     | 18.27    |
We finally present results for our models in a few-shot scenario to investigate performance using cross-lingual data with a limited number of example summaries (shots) in the target language. Figure 5 shows few-shot results averaged across all four target languages (detailed per-language results are given in Table 11) for different sizes of the training data from the X-SCITLDR dataset. The results highlight that, while CL-TLDR is a difficult task with the models having little cross-lingual transfer capabilities (as shown in the zero-shot experiments), performance can be substantially improved when combining a small amount of cross-lingual data, i.e. as little as 1% of examples for each target language, and intermediate training. As the amount of cross-lingual training data increases, the benefits of intermediate fine-tuning become smaller and results for all models tend to converge. This indicates the benefits of intermediate fine-tuning in the scenario of limited training data such as, e.g., low-resource languages other than those in our resource. Our few-shot results indicate that we can potentially generate TLDRs for a multitude of languages by creating little labeled data in those languages, and at the same time leverage via intermediate fine-tuning labeled resources for English summarization and machine translation (for which there exist plenty of resources).

### Table 11: Detailed few-shot performance on downstream CL-TLDR for each language and cross-lingual model with different percentages of X-SCITLDR training data for each target language.

| Lang | Model     | Rouge-1 (avg) |   | Rouge-2 (avg) |   | Rouge-L (avg) |   |
|------|-----------|---------------|---|---------------|---|---------------|---|
|      | 1% | 5% | 30% | 100% | 1% | 5% | 30% | 100% | 1% | 5% | 30% | 100% |
| DE   | CLSum     | 11.30 | 14.67 | 13.74 | 17.99 | 1.30 | 1.94 | 2.43 | 3.58 | 9.33 | 11.78 | 11.25 | 14.69 |
|      | CLSum+EnSum | 15.01 | 13.72 | 17.92 | 18.03 | 1.65 | 1.82 | 2.94 | 3.56 | 12.24 | 10.85 | 14.24 | 14.74 |
|      | CLSum+MT  | 9.30   | 16.01 | 18.42 | 18.28 | 1.19 | 2.83 | 3.89 | 3.99 | 7.90  | 13.04 | 15.07 | 15.07 |
| IT   | CLSum     | 9.51   | 17.18 | 18.25 | 21.20 | 1.50 | 3.17 | 4.27 | 6.15 | 8.20  | 13.75 | 15.20 | 17.54 |
|      | CLSum+EnSum | 16.36 | 18.06 | 21.03 | 20.47 | 2.46 | 3.35 | 5.92 | 6.14 | 13.12 | 14.22 | 17.51 | 17.39 |
|      | CLSum+MT  | 15.40 | 18.28 | 21.56 | 21.71 | 2.82 | 4.61 | 6.80 | 7.04 | 12.27 | 15.23 | 17.73 | 18.11 |
| ZH   | CLSum     | 0.76   | 9.97  | 20.45 | 23.03 | 0.09 | 1.13 | 4.30 | 5.76 | 0.73  | 8.77  | 17.55 | 20.27 |
|      | CLSum+EnSum | 4.47  | 13.18 | 23.06 | 22.62 | 0.31 | 2.03 | 5.57 | 5.52 | 4.23  | 11.35 | 19.84 | 19.88 |
|      | CLSum+MT  | 14.80 | 18.11 | 24.05 | 23.28 | 2.84 | 3.83 | 6.20 | 5.97 | 12.84 | 15.53 | 20.63 | 20.27 |
| JA   | CLSum     | 2.17   | 27.29 | 31.78 | 30.94 | 0.63 | 4.70 | 5.36 | 4.66 | 1.98  | 18.75 | 20.63 | 20.34 |
|      | CLSum+EnSum | 21.87 | 29.60 | 29.86 | 32.30 | 3.23 | 4.85 | 4.44 | 5.66 | 16.26 | 19.97 | 19.55 | 20.89 |
|      | CLSum+MT  | 28.25 | 30.78 | 32.23 | 32.30 | 4.31 | 5.59 | 5.54 | 5.47 | 19.33 | 20.61 | 20.57 | 20.85 |

### 6 Related Work

**General-domain summarization datasets.** News article platforms play a major role when collecting data for summarization [21, 47], since article headlines provide ground-truth summaries. Narayan et al. [40] propose a news domain summarization dataset with highly compressed summaries to provide a more challenging summarization task (i.e., extreme summarization). Sotudeh et al. [50] propose TLDR9+, another extreme summarization dataset that was collected automatically from a social network service.

**Cross-lingual summarization datasets.** While there are growing numbers of cross-lingual datasets for natural language understanding tasks [11, 30, 45], few datasets for cross-lingual summarization are available. Zhu et al. [60] propose to use machine translation to extend English news summarization to Chinese. To ensure dataset quality, they adopt round-trip translation by translating the original summary into the target language and back-translating the result to the original language for comparison, keeping the ones that meet a predefined similarity threshold. Ouyang et al. [41] create cross-lingual summarization datasets by using machine translation for low-resource languages such as Somali, and show that they can generate better summaries in other languages by using...
noisy English input documents with English reference summaries. Our work differs from these prior attempts in that our automatically translated summaries are corrected by human annotators, opposed to providing silver standards in the form of automatic translations without any human correction. Recently, Ladhak et al. [27] presented a large-scale multilingual dataset for the evaluation of cross-lingual abstractive summarization systems that is built out of parallel data from WikiHow. Even though it is a large high-quality resource of parallel data for cross-lingual summarization, this corpus is built from how-to guides: our dataset focuses instead on scholarly documents. Besides cross-lingual corpora, there are also large-scale multilingual summarization datasets for the news domain [48, 52]. The work we present here differs in that we focus on extreme summarization for the scholarly domain and we look specifically at the problem of cross-lingual summarization in which source and target language differ.

Datasets for summarization in the scholarly domain. There are only a few existing scholarly datasets for summarization the scholarly domain and most of them are in English. SCITLDR [5], the basis for our work on multilingual summarization, presents a dataset for research papers (see Section 2 for more details). Collins et al. [10] use author-provided summaries to construct an extractive summarization dataset from computer science papers, with over 10,000 documents. Cohan et al. [7] regard abstract sections in papers as summaries and create large-scale datasets from two open-access repositories (arXiv and PubMed). Yasunaga et al. [57] efficiently create a dataset for the computational linguistics domain by manually exploiting the structure of papers. Meng et al. [38] present a dataset which contains four summaries from different aspects for each paper, which makes it possible to provide summaries depending on requests by users. Lu et al. [35] is a large-scale dataset for multi-document summarization for scientific papers, for which models need to summarize multiple documents.

The work closest to ours has been recently presented by Fatima and Strube [16], who introduce an English-German cross-lingual summarization dataset collected from German scientific magazines and Wikipedia. This resource is complementary to ours in many different aspects. While both datasets are in the scientific domain, their data includes either articles from the popular science magazine Scientific American / Spektrum der Wissenschaft or articles from the Wikipedia Science Portal. In contrast, our dataset includes scientific publications written by researchers for a scientific audience. Second, our dataset focuses on extreme, TLDR-like datasets, which we argue is more effective in helping researchers browse through many potentially relevant publications in search engines for scholarly documents. Finally, our summaries are expert-generated, as opposed to relying on the ‘wisdom of the crowds’ from Wikipedia, and are available in three additional languages.

Summarization of scientific documents. In recent years, there has been much work on the problem of summarizing scientific publications, a task that belongs to the wider area of scholarly data and text mining [2, 46]. Scholarly document processing has gained much traction lately, due to the ever growing need to efficiently access large amounts of published information, e.g., in the COVID-19 pandemic [15, 54]. Research efforts in summarization have been arguably catalyzed by community-driven evaluation campaigns such as the CL-SciSumm shared tasks [6, 23]. Previous work on summarization has focused on specific features of scientific documents such as using citation contexts [9, 59] or document structure [8, 12]. Complementary to these efforts is a recent line of work on automatically generating visual summaries or graphical abstracts [55, 56]. In our work, we build upon recent contributions on using multilingual pre-trained language models for cross-lingual summarization [27] and extreme summarization for English [5], and bring these two lines of research together to propose the new task of cross-lingual extreme summarization of scientific documents.

7 Conclusion

In this paper, we presented X-SCITLDR, the first dataset for cross-lingual summarization of scientific papers. Our new dataset makes it possible to train and evaluate NLP models that can generate summaries in German, Italian, Chinese and Japanese from input papers in English. We used our dataset to investigate the performance of different architectures based on multilingual transformers, including a two-stage ‘summarize and translate’ approach and a direct cross-lingual model. We additionally explored the potential benefits of intermediate task and cross-lingual fine-tuning and analyzed the performance in zero- and few-shot scenarios. For future work, we plan to investigate how to apply additional techniques designed for cross-lingual text generation such as training with multiple decoders [60], automatically complementing our multilingual TLDRs with visual summaries [55], as well as devising new methods to include background knowledge such as, in our case, technical terminology and domain adaptation capabilities [58], into multilingual pre-trained models.

Our work crucially builds upon recent advances in multilingual pre-trained models [32] and cross-lingual summarization [27], and investigates how these methodologies can be applied for multilingual scholarly document processing. The application of NLP techniques for mining scientific papers has been primarily focused on the English language: while this work we want to put forward the vision of enabling scholarly document processing for a wider range of languages, ideally including both resource-rich and resource-poor languages in the longer term. Our vision of ‘Scholarly Document Processing for all languages’ is in line with current trends in NLP (cf., e.g., [44] and [1], inter alia): while our initial effort concentrated here on fairly resource-rich languages, in future work we plan to focus specifically on resource-poor languages where multilingual NLP can and is indeed expected to make a difference in enabling wider (and consequently more diverse and fairer [25]) accessibility to scholarly resources.

Downloads

The X-SCITLDR corpus and the code used in our CL-TLDR experiments is available under an open license at https://github.com/sobamchan/xscitldr.

Acknowledgments

The work presented in this paper is funded by the German Research Foundation (DFG) under the VADIS (PO 1900/5-1; EC 477/7-1) and JOIN-T2 (PO 1900/1-2) projects. We thank Ines Rebbein and the three anonymous reviewers for their helpful comments.
References

[1] David Ieofoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D’souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzabu, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azmne, Shamsud-Deen Momodu, Chizuru Emure, Joyce Nokotumba-Nahende, Perex Ogyo, Amosluwu Aweni, Catherine Gitau, Derguemie Mbaye, Jenskaabali Alahi, Seid Muhim Yamine, Tajuddem Gnawade, Ignatieus Ezeunci, Rubunago Andre Nyongabo, Jonathan Mukiiria, Verrah Otuende, Irofe Onrle, Davis David, Samba Ngom, Tock Adevuru, Paul Rayon, Motofelewu Adami, Gerald Munuki, Emmanuel Anebi, Chimakama Chukwueke, Nikruda Odu, Eric Peter Wairagala, Samuel Oyewale, Davide Puro, Badeley Feaye, Blessing Sibanda, Oreyangheh Anna, Bonaventure F. P. Dossou, Kechchi Ogejii, Thierno Brahama DIOP, Abdoulaye Diallo, Adekunle Akinfaderi, Tendai Marengekker, and Solomon Osei. 2021. MasakanaNER: Named Entity Recognition for African Languages. arXiv:2103.11811 [cs.CL].

[2] Ize Betagy, Arman Cohan, Guy Feigenblat, Dayne Freitag, Tirthankar Ghosal, Keith Hall, Drahomira Herrmannova, Petr Knoch, Kyle Lo, Philipp Mayr, Robert Patton, Michal Shmueli-Scheurer, Anita de Waard, Kuanan Wang, and Lucy Wang. 2021. Overview of the Second Workshop on Scholarly Document Processing. In Proceedings of the Second Workshop on Scholarly Document Processing. Association for Computational Linguistics, Online, 159–165. https://aclanthology.org/2021.sdp-1.12

[3] Chandra Bhagavatula, Sergey Feldman, Russell Power, and Waled Ammar. 2018. Content-Based Recommendation. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Association for Computational Linguistics, New Orleans, Louisiana, 235–245. https://doi.org/10.18653/v1/N18-1022

[4] Lutz Bornmann and Rüdiger Mutz. 2015. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. J. Asoc. Inf. Sci. Technol. 66, 11 (Nov. 2015), 2125–2122. https://doi.org/10.1002/asi.23329 Publisher. Wiley.

[5] Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel Weld. 2020. TDLR: Extreme Summarization of Scientific Documents. In Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics, Online, 4271–4280. https://doi.org/10.18653/v1/2020.findings-emnlp.428

[6] Muthu Kumar Chandrasekaran, Michihito Yasunaga, Dragomir Radev, Dayne Freitag, and Min-Yen Kan. 2019. Overview and Results: CL-SciSumm Shared Task 2019. arXiv:1907.09654 [cs.CL].

[7] Arman Cohan and Nazli Goharian. 2018. Scientific document summarization via citation contextualization and scientific discourse. Int J Digit Libr. 19, 2-3 (2018), 287–303. https://doi.org/10.1007/s10791-017-0216-8

[8] Ed Collins, Iz Beltagy, Arman Cohan, Guy Feigenblat, Dayne Freitag, Tirthankar Ghosal, Ronald Kemker, Marc McClure, Angelina Abitino, Tyler L. Hayes, and Christopher D. Manning. 2018. The Hitchhiker’s Guide to Reading and Comprehend. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: HLT 2018. Association for Computational Linguistics, New Orleans, Louisiana, 1693–1701. https://doi.org/10.18653/v1/N18-1288

[9] Suchin Gururangan, Ana Marasović, Michael Arandjelović, and Paulusi Suleyman. 2015. Teaching Machines to Read and Comprehend. In Advances in Neural Information Processing Systems. Curran Associates, Inc., Red Hook, NY, USA, 3997–4005. ISBN: 978-1-5107-1695-4 Publisher. MIT Press.

[10] Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. Wikisum: A New Benchmark Dataset for Cross-Lingual Abstractive Summarization. In Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics, Online, 4034–4048. https://doi.org/10.18653/v1/2020.findings-emnlp.360

[11] Pratik Joshi, Sebastain Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Nalini Johnson. 2020. XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalisation. In Proceedings of the 37th Annual Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119), Hal Daumé III and Aarti Singh (Eds.). PMLR, Online, 4141–4150. https://proceedings.mlr.press/v119/hdaume302.html

[12] Robert Doherty, Andrew Mani, and Dominik Pfeffer. 2018. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Online and in Dominican Republic, 39–50. https://doi.org/10.18653/v1/2018.acl-main.56

[13] Ronald Kemker, Marc McClure, Angelina Abitino, Tyler L. Hayes, and Christopher D. Manning. 2018. Measuring Catastrophic Forgetting in Neural Networks. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 32nd Innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018. Shaia A. McInerney and Khanan Q. Weinberger (Eds.). AAAI Press, New Orleans, Louisiana, USA, 3390–3398. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16410

[14] Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. Wikisum: A New Benchmark Dataset for Cross-Lingual Abstractive Summarization. In Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics, Online, 4034–4048. https://doi.org/10.18653/v1/2020.findings-emnlp.360

[15] Anne Esteve, Anupriya Kale, Romain Paulus, Karuma Hashimoto, Wenpeng Yin, Dragomir Radev, and Richard Socher. 2021. COVID-19 information retrieval with deep-learning based semantic search, question answering, and abstractive summarization. npj Digital Medicine 4, 1 (12 Apr 2021), 68. https://10.1038/s41746-021-00437-7

[16] Preeti Garg, Sai Prasad, and Dipankar Raychaudhuri. 2020. Multimodal abduction for Zero-Shot Cross-Lingual Transfer with Multilingual Transformers. (May 2020). http://arxiv.org/abs/2005.06349

[17] Mike Lewis, Yinhan Liu, Naman Goyal, Meghana Ghandhi, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence pre-training for Natural Language Generation, Translation, and Comprehension. In Proceedings of the 58th Annual Meeting of
