International Journal on Digital Libraries (2024) 25:249–271
https://doi.org/10.1007/s00799-023-00373-2

Cross-lingual extreme summarization of scholarly documents

Sotaro Takeshita[1] · Tommaso Green[1] · Niklas Friedrich[1] · Kai Eckert[2] · Simone Paolo Ponzetto[1]

Received: 7 November 2022 / Revised: 25 May 2023 / Accepted: 4 June 2023 / Published online: 10 August 2023

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. Recent work has tried to address this problem by developing methods for automated summarization in the scholarly domain, but concentrated so far only on monolingual settings, primarily English. In this paper, we consequently explore how state-of-the-art neural abstract summarization models based on a multilingual encoder–decoder architecture can be used to enable cross-lingual extreme summaries of scholarly texts. To this end, we compile a new abstractive cross-lingual summarization dataset for the scholarly domain in four different languages, 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 pipeline approach that independently summarizes and translates, as well as 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. Finally, we investigate how to make our approach more efficient on the basis of knowledge distillation methods, which make it possible to shrink the size of our models, so as to reduce the computational complexity of the summarization inference.

Keywords Scholarly document processing · Summarization · Multilinguality

1 Introduction

For years, the number of scholarly documents has been steadily increasing [9], 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 [8, 61]. One approach is to identify relevant sentences in the paper based on automatic classification [42]. This approach to information distillation is taken one step further by fully automatic text summarization, where a long document is used as input to produce a shorter version of it covering essential points [17, 95], possibly a TLDR[1]-like 'extreme' summary [10]. 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 tech-

✉ Sotaro Takeshita
sotaro.takeshita@uni-mannheim.de

Tommaso Green
tommaso.green@uni-mannheim.de

Niklas Friedrich
nfriedri@mail.uni-mannheim.de

Kai Eckert
k.eckert@hs-mannheim.de

Simone Paolo Ponzetto
ponzetto@uni-mannheim.de

[1] Data and Web Science Group, University of Mannheim, Mannheim, Germany

[2] Department of Computer Science, Mannheim University of Applied Sciences, Mannheim, Germany

[1] TLDR stands for "too long; didn't read" and is often used in online communication and texts to indicate a short summary that makes it possible to avoid reading a longer text.

niques using supervision from labeled data. For the task of summarizing research papers, most available datasets are in English only, e.g., CSPubSum/CSPubSumExt [17] and ScisummNet [95], with community-driven shared tasks also having concentrated on English as *de facto* the only language of interest [12, 40]. 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 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.[2]

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. [10] 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 [26], who compile an English-German cross-lingual dataset from the Spektrum der Wissenschaft/Scientific American and Wikipedia. We focus on extreme summarization, build a dataset of expert-derived multilingual TLDRs (as opposed to leads from Wikipedia) and provide additional languages.

Contributions. Our work provides the following contributions on the research topic of cross-lingual summarization for the scholarly domain.

- 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** from computer science in four different languages.
- We use our dataset to **benchmark the difficulty of cross-lingual extreme summarization** with different models built on top of state-of-the-art pre-trained language models [49, 57].
- We additionally **investigate whether cross-lingual summarization models using large pre-trained language models can be improved with intermediate fine-tuning techniques**, which have shown to be effective to improve performance of pre-trained multilingual language models on many downstream NLP tasks [29, 32, 70, 71, *inter alia*].

We build upon our original paper [86] and extend it in a number of ways:

- We benchmark **the choice of the multilingual encoder–decoder** by comparing performance of our original models using mBART [57] with those using mT5 [92].
- We study **the role of the stacking order in the summarization and translation pipeline approach**, so as to establish whether we can achieve better cross-lingual summaries by first translating and then summarizing, or vice versa.
- We further analyze **the code-switching capabilities of our model** by quantifying how much our multilingual models are able to retain English technical terminology in the translated summaries.
- We investigate **the application of a knowledge distillation method** [82] on our direct cross-lingual summarization models to explore the possibility of shrinking the model sizes while keeping the original summarization output quality.

While the first three new contributions are meant to extend the experimental part so as to provide a more complete and in-depth analysis of our original experiments, the last one focuses on improving its scope of application. This is because the large size of the cross-lingual models we use in our experiments can hinder building scalable real-world applications around them. To address this point, we follow the recent trend in 'green' and scalable NLP [65] and explore how to reduce the computational inference costs of our summarization models using knowledge distillation. This is especially essential for our overarching future vision of coupling summarization with semantification techniques within the broader vision of the VADIS project, which aims at improving accessibility of social science publications by connecting survey data and text from research papers [44].

The remainder of this paper is organized as follows. Section 2 provides an overview of relevant previous work in monolingual and multilingual summarization, as well as the broader field of scholarly document mining. We summarize in Sect. 3 seminal work on monolingual extreme summa-

[2] https://www.semanticscholar.org/product/tldr

rization for English from Cachola et al. [10], on which our multilingual extension builds upon. We next introduce our new dataset for cross-lingual TLDR generation in Sect. 4. We present our cross-lingual models and benchmarking experiments in Sects. 5 and 6, respectively. We wrap up our work with concluding remarks and directions for future work in Sect. 7.

2 Related work

2.1 Datasets and resources

General-domain summarization datasets. News article platforms play a major role when collecting data for summarization [35, 78], since article headlines provide ground-truth summaries. Narayan et al. [66] propose a news domain summarization dataset with highly compressed summaries to provide a more challenging summarization task (i.e., extreme summarization). Sotudeh et al. [84] 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 [18, 53, 74], few datasets for cross-lingual summarization are available. Zhu et al. [99] 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. [68] 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, as opposed to providing silver standards in the form of automatic translations without any human correction. Recently, Ladhak et al. [46] presented a large-scale multilingual dataset for the evaluation of cross-lingual abstractive summarization systems that are 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. Perez-Beltrachini and Lapata [69] automatically constructed datasets for cross-lingual summarization in four European languages by exploiting the structure of Wikipedia. Besides cross-lingual corpora, there are also large-scale multilingual summarization datasets for the news domain [80, 87]. 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 summarization datasets for the scholarly domain and most of them are in English. SCITLDR [10], the basis for our work on multilingual summarization, presents a dataset for research papers (see Sect. 3 for more details). Collins et al. [17] use author-provided summaries to construct an extractive summarization dataset from computer science papers, with over 10,000 documents. Cohan et al. [14] regard abstract sections in papers as summaries and create large-scale datasets from two open-access repositories (arXiv and PubMed). Yasunaga et al. [95] efficiently create a dataset for the computational linguistics domain by manually exploiting the structure of papers. Meng et al. [62] 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. [59] release 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 [26], 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 include 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 summarization, 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.

2.2 Models

Scholarly document mining. In recent years, there has been much interest from the NLP community in developing text mining techniques that bring order and provide novel ways to better access scientific publications [76]. Previous work has addressed a wide range of tasks, including citation linking [2, 3] and recommendation [34, 38], summarization [1, 77] (*inter alia*, see below) and argumentation mining [4, 5, 31]. But while there have been full-fledged projects on mining scientific publications [72], scholarly document processing has arguably gained much traction lately [7, 16], due to the ever growing need to efficiently access large amounts of published information, e.g., in the COVID-19 pandemic [24, 89]. Most recent contributions range from scholarly specific search platforms [47] all the way through novel reading interfaces [27] and full-fledged infrastructures [11, 44] leveraging advancements in data-driven AI, NLP and semantification techniques (e.g., document understanding and information extraction).

Automated text summarization. Summarization is a longstanding task in NLP [33, 67]. While early efforts focused mostly on extractive summarization [55], e.g., using an unsupervised graph-based approach [63], abstractive summarization has gained ever more traction in recent years starting with work using sequence-to-sequence models [75]. Just like in virtually all areas of NLP research, most successful current 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 [56, 81], including cross- and multilingual models and datasets [80, 87], as well as specialized ones e.g., the biomedical domain [64]. Work on cross-lingual summarization has historically received little attention until recent years [90], arguably to due to the availability of new resources (Sect. 2.1) as well as neural multilingual summarizers.

Summarization of scientific documents. In recent years, there has been much work on the problem of summarizing scientific publications and community-driven evaluation campaigns such as the CL-SciSumm shared tasks [12, 40]. Previous work on summarization has focused on specific features of scientific documents such as using citation contexts [13, 97] or document structure [15, 19]. Complementary to these efforts is a recent line of work on automatically generating visual summaries or graphical abstracts [93, 94]. In our work, we build upon recent contributions on using multilingual pre-trained language models for cross-lingual summarization [46] and extreme summarization for English [10] and bring these two lines of research together to propose the new task of cross-lingual extreme summarization of scientific documents.

Knowledge distillation for summarization models. While massively large pretrained language models achieve strong results on various summarization tasks, the enormous sizes hinder their deployment in real-world applications. Knowledge distillation [36] offers a chance to reduce the model size by transferring knowledge of the original teacher model to a smaller student without large performance drops. Because of its practicality, there has been a lot of work exploring how to utilize this framework for various NLP tasks [41, 79] as well as for summarization. Shleifer and Rush [82] perform comparative experiments of three different knowledge distillation methods for summarization models to better understand how they affect training and inference time as well as final summary quality. Zhang et al. [98], on the basis of their observation of how attention layers behave in summarization models, propose to modify the attention temperature parameter in the teacher model to generate pseudo-labels that are easier to learn for the student model. Li et al. [52] present a controlled study to understand the interaction between model quantization and distillation and report significant speed improvements. In our work, we utilize a simple yet effective knowledge distillation method called ‘shrink and fine-tune’ investigated by Shleifer and Rush [82] to understand its effects on our new cross-lingual extreme summarization task.

Table 1 An example of a TLDR summary for a research paper. Source: https://openreview.net/forum?id=0XXpJ4OtjW

Abstract: We propose a method for meta-learning reinforcement learning algorithms by searching over the space of computational graphs which compute the loss function for a value-based model-free RL agent to optimize. [...] **Introduction:** Designing new deep reinforcement learning algorithms that can efficiently solve across a wide variety of problems generally requires a tremendous amount of manual effort. [...] **Conclusion:** In this work, we have presented a method for learning reinforcement learning algorithms. We design a general language for representing algorithms which compute the loss function for [...]
TLDR: We meta-learn RL algorithms by evolving computational graphs which compute the loss function for a value-based model-free RL agent to optimize.

3 SCITLDR: English monolingual extreme summarization of scientific documents

Our work builds heavily on seminal work on extreme summarization of scientific publications from Cachola et al. [10], who first introduced an English monolingual dataset for this task and used it to benchmark a variety of state-of-the-art summarization models.

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 versus 36.5% of CLPubSum [17]. 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

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’)

a) German

Original Summary	The **paper** presents a multi-view framework for improving sentence representation in NLP tasks using generative and discriminative objective architectures.
Automatic Translation	Das Papier präsentiert einen Multi-View-Rahmen zur Verbesserung der Satzrepräsentation in NLP-Aufgaben . . .
Postedited Version	Der Artikel präsentiert einen Multi-View-Rahmen zur Verbesserung der Satzrepräsentation in NLP-Aufgaben . . .

b) Italian

Original Summary	The paper provides a full characterization of permutation invariant and equivariant linear layers for **graph data**.
Automatic Translation	L’articolo fornisce una caratterizzazione completa degli strati lineari invarianti di permutazione ed equivarianti per i dati del grafico .
Postedited Version	L’articolo fornisce una caratterizzazione completa dei layer lineari invarianti o equivarianti per la permutazione per i dati del grafo .

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

a) German

Original Text	The paper proposes a framework for constructing spherical **convolutional networks** based on a novel synthesis of several existing concepts.
Automatic Translation	Das Papier schlägt einen Rahmen für die Konstruktion von sphärischen Faltungsnetzen vor, der auf einer neuartigen Synthese mehrerer bestehender Konzepte beruht.
Postedited Version	Die Arbeit schlägt einen Rahmen für die Konstruktion von sphärischen Convolutional Networks vor , der auf einer neuartigen Synthese mehrerer bestehender Konzepte beruht.

b) Italian

Original Text	We present a novel iterative algorithm based on generalized low rank models for computing and interpreting **word embedding models**.
Automatic Translation	Presentiamo un nuovo algoritmo iterativo basato su modelli generalizzati di basso rango per il calcolo e l’interpretazione dei modelli di incorporazione delle parole .
Postedited Version	Presentiamo un nuovo algoritmo iterativo basato su modelli generalizzati di basso rango per il calcolo e l’interpretazione dei modelli di word embedding .

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 5411 TLDRs for 3229 scientific papers in the computer science domain: it is divided into a training set of 1992 papers, each with a single gold-standard TLDR, and dev and test sets of 619 and 618 papers each, with 1452 and 1967 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,[3] as well as human-generated summaries from peer-review comments found on the same platform.

In the following, we extend the original work of Cachola et al. [10] in two different ways, namely in terms of: (a) a new multilingual dataset for TLDR-like extreme summarization in languages other than English and (b) benchmarking of multilingual transformer-based pre-trained generative language models. We achieve this by creating a new cross-lingual dataset that consists of an automatically translated, post-edited version of the SCITLDR dataset to support four additional languages, namely German, Italian, Chinese and

[3] https://openreview.net

Table 4 Statistics of our dataset (X-SCITLDR)

	Documents				Summaries			
	# documents (train/dev/test)	# words	vocabulary size	average # words per doc	# words	vocabulary size	average # words per summary	compression ratio (%)
EN	1992/619/618	370,244	20,819	5000	47,574	6725	23.88	244.57
DE					43,929	13,808	22.05	264.87
IT					48,050	7127	24.12	242.14
ZH					47,711	7953	23.95	243.86
JA	1606/199/199	306,815	14,769	10,000	121,989	6706	75.91	131.73

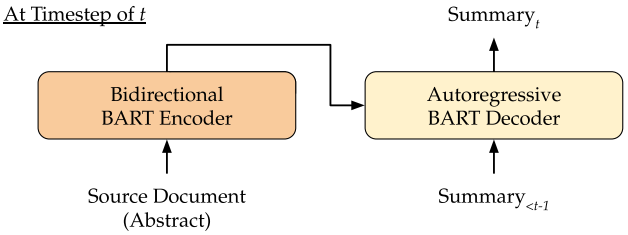

At Timestep of *t*
Summary$_t$
Bidirectional BART Encoder
Autoregressive BART Decoder
Source Document (Abstract)
Summary$_{<t-1}$

Fig. 1 Using monolingual BART for English text summarization: BART is fine-tuned to convert a given text (e.g., paper abstract) into a shorter summary in a regressive manner by generating one token at a time

Japanese. We then use these reference summaries to fine-tune pre-trained language models and produce multilingual summarization systems that are able to support languages other than English as the target language.

4 X-SCITLDR: a new dataset for cross-lingual extreme summarization of scientific papers

We first describe the creation of our X-SCITLDR dataset and briefly present some statistics to provide a quantitative overview. Our dataset is composed of two main sources:

- An automatically translated, manually post-edited version of the original SCITLDR dataset [10] for German, Italian and Chinese (X-SCITLDR-PostEdit).
- A manually generated dataset of expert-authored TLDRs harvested from a community-based summarization platform for Japanese (X-SCITLDR-Human).

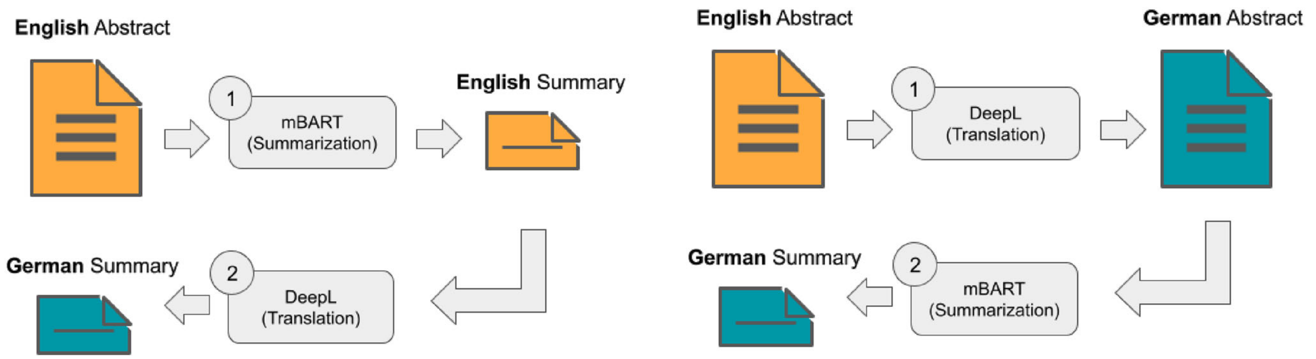
Besides allowing us to evaluate our models across languages with different sizes of pre-training data (e.g., mBART has been exposed to half as much Italian than Japanese or German, cf. Table 1 from [57]), using two different sources allows us to perform a 'cross-domain'-like evaluation between datasets from different sources, namely conference reviews (X-SCITLDR-PostEdit) versus expert community efforts (X-SCITLDR-Human), so as to evaluate the generalization capabilities of our models across different domains. Moreover, having a dataset comprising post-edited translated summaries and human-generated ones makes it possible to investigate performance across different summarization styles—since post-edited summaries are not guaranteed to be the same as the ones humans would have generated from scratch.

X-SCITLDR-PostEdit. Given the overall quality of automatic translators [20], we opt for a hybrid machine-human translation process of post-editing [30] 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.[4] 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.

[4] https://www.deepl.com/translator



(a) 'summarize and translate' (EnSum-MT)

(b) 'translate and summarize' (TGTSum-MT)

Fig. 2 Two pipelines for cross-lingual summarization using automatic machine translation and monolingual summarization in stacking orders: **(a)** 'summarize and translate' (Sect. 5.1) first summarizes an English abstract and then translates the generated summary into the target language; **(b)** 'translate and summarize' (Sect. 5.1) translates an English abstract to the target language and then summarizes it with a monolingual summarization model in the same target language

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 a community-based summarization platform, arXivTimes.[5] 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 or vice versa.

In Table 4, we present various statistics of our X-SCITLDR dataset for both documents and summaries from the original English (EN) SCITLDR data and our new dataset in four target languages.[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. [10, Table 9]), whereas Japanese summaries often contain more than one sentence. Despite having both longer documents and summaries, the Japanese data still exhibit a very high compression ratio (cf. datasets for summarization of both scientific and nonscientific documents having typically a compression ratio $< 40\%$), which indicates their suitability for evaluating extreme summarization in the scholarly domain.

5 CL-TLDR: cross-lingual extreme summarization of scholarly documents

We present in this section the different 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 [10] and focus on *abstractive* summarization, since this has been shown to outperform *extractive summarization* in a variety of settings. We first present the two transformer-based pre-trained generative language models used within our summarization systems, namely mBART and mT5, and show how to use them within two different architectures for CL-TLDR, namely a two-stage pipeline model (Sect. 5.1) and direct CL-TLDR approach (Sect. 5.2).

[5] https://arxivtimes.herokuapp.com

[6] Slight differences with respect to the statistics from Cachola et al. [10, 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). Vocabulary sizes are computed after lemmatization with SpaCy.

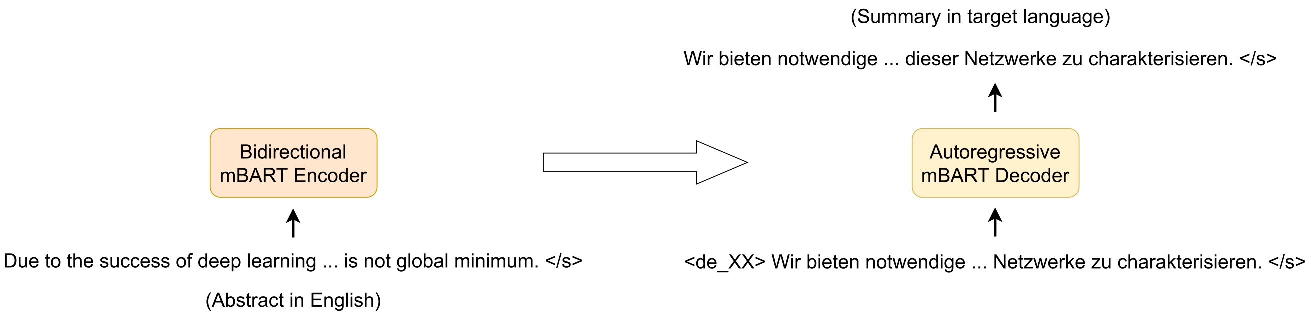

(Summary in target language)
Wir bieten notwendige ... dieser Netzwerke zu charakterisieren. </s>
Bidirectional mBART Encoder
Autoregressive mBART Decoder
Due to the success of deep learning ... is not global minimum. </s>
<de_XX> Wir bieten notwendige ... Netzwerke zu charakterisieren. </s>
(Abstract in English)

Fig. 3 mBART learns to take an English abstract and generate a summary in the target language (here, German). We can control the target language by providing a language token (< de > in the figure)

Pretrained mBART
Intermediate Task Fine-Tuning (monolingual English summarization)
Target-Language Summarization Fine-Tuning
Pretrained mBART
Cross-lingual Intermediate Fine-Tuning (En-X Machine Translation)
Target-Language Summarization Fine-Tuning

(a) Intermediate task fine-tuning (CLSum+EnSum)

(b) Cross-lingual intermediate fine-tuning (CLSum+MT)

Fig. 4 Two approaches to perform the intermediate fine-tuning to mitigate the training data scarcity issue. **(a)** CLSum+Ensum inserts an additional training stage between pre-training and cross-lingual summarization fine-tuning in which the pre-trained language model learns to summarize in English. **(b)** CLSum+MT inserts an additional training stage between pre-training and downstream cross-lingual summarization by fine-tuning the pre-trained language model to translate from English into the target language

Base models: mBART and mT5. In our experiments, we use BART [49] and its multilingual variant mBART [57] as underlying summarization models. They are both transformer-based [88] 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—cf. Fig. 1.

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 [49]. We follow Ladhak et al. [46] 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) (Sect. 5.1); (b) a direct cross-lingual summarization system obtained by fine-tuning mBART with input articles from English and summaries from the target language (Sect. 5.2).

In addition to mBART, we quantify performance using different pre-trained language models, in order to benchmark how stable our results are across different transformer-based encoder–decoder models. For this we additionally evaluate performance using mT5 [92], which, akin to mBART, is a large pre-training language model designed to generate texts in multiple languages. When compared with mBART, mT5 has two major differences, namely a) the noising function used during pre-training—i.e., while mBART learns to recover masked spans and shuffled sentences in texts, mT5 only uses span masking as a noising function—and b) the overall model size—i.e., mBART and mT5—contains 610 and 1229 million of parameters, respectively, which impacts computational costs (e.g., required computational time, memory consumption).

5.1 Two-stage cross-lingual summarization

A first solution to the CL-TLDR task is to combine a monolingual summarization model with a machine translation system. This approach is composed of two stages, namely translation and (monolingual) summarization. In this work, we investigate two variants of this setting, namely 'summarize and translate' and 'translate and summarize', whose difference is the different stacking order of the two modules, i.e., whether we first translate and then (monolingually) summarize or vice versa.

Summarize and Translate. One variant of two-stage cross-lingual summarization is to have the model first take an English text as input and then generate a summary in English (we call this approach **EnSum-MT**): the English summary is then automatically translated into the target language using machine translation (Fig. 2a).[7] This model does not rely on

[7] Similarly to the creation of the multilingual portion of our dataset, we opt again for DeepL for all our languages (cf. Sect. 4).

any cross-lingual signal: it merely consists of two independent modules for translation and summarization and does not require any 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 [99], since errors of the first stage (i.e., summarization) get amplified in the second stage (i.e., translation) leading to overall performance degradation.

Translate and Summarize. An alternative two-stage summarization pipeline consists in training monolingual summarization models for each target language by translating English input documents to match the language of reference summaries (we call this method **TGTSum-MT**). During testing, input documents are then similarly first translated and then summarized using the corresponding monolingual summarization model.

While having a model observe more text in the target language is known to help its performance [39], the overall cost of translating texts is higher for TGTSum-MT than in EnSum-MT as it requires (1) to translate the entire input documents, as opposed to (shorter) generated summaries, and (2) to perform automatic translation not only of the test documents (for evaluation) but also of the train and development sets (for training monolingual summarization models and tune their hyperparameters).

5.2 Direct cross-lingual summarization

A third 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 pre-trained multilingual denoising autoencoders like mBART [57] and mT5 [92] and use the cross-lingual training data provided by our new X-SCITLDR dataset to fine-tune them and generate summaries in the target languages given abstracts in English, as depicted in Fig. 3. We follow Liu et al. [57] 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 [35, 66]. 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, [29, 32, 70, 71]). Gururangan et al. [32], 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 (Fig. 4a).
- **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.[8] To alleviate this problem, we investigate training our model on machine translation before fine-tuning it on the summarization task. For this, we automatically translate English abstracts into the target language and use these synthetic data as training data for fine-tuning on the task of automatically translating abstracts (Fig. 4b).

Knowledge distillation. Akin to other recent pre-trained language models, the demanding computational requirements of mBART hinder its deployment in real-world applications. To tackle this issue, there are works that aim to reduce the size of large summarization models [52, 83, 98]. In our work, we evaluate one of the knowledge distillation approaches proposed by Shleifer and Rush [83], dubbed 'shrink and fine-tune,' which takes a trained mBART as a teacher model and uses some of its parameters for the initialization of a smaller version of the model (called student) and finally fine-tunes the student model again on the target dataset. Clearly, training teacher and student takes more time than fine-tuning the teacher alone but this provides us with a smaller model which can be more easily deployed.

6 Experiments

Input documents. We follow Cachola et al. [10] 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 185.9 words for X-SCITLDR-PostEdit (EN/DE/IT/ZH) with an

[8] 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. [46]).

average compression ratio of 12.64% and 190.9 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 [54]. In the case of the post-edited portion of the X-SCITLDR dataset (X-SCITLDR-PostEdit, Sect. 4), 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 (*avg* and *max*) to account for these multiple references [10]. For *avg*, we compute ROUGE F1 scores with respect to the different references and take the average score, whereas for *max* we select the highest scoring one. The Japanese dataset does not contain multiple reference TLDRs: hence, we compute standard ROUGE F1 only. We test for statistical significance using sample level Wilcoxon signed-rank test with $\alpha = 0.05$ [23].

Hyperparameter tuning. To find the best hyperparameters for each model, we use the development data and run a grid search using ROUGE-1 avg as a reference metric. We run experiments with learning rate $\in \{1 \cdot 10^{-5}, 3 \cdot 10^{-5}\}$ 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 [58] 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 [45]), we freeze the parameters of the embedding and decoder layers during intermediate task and cross-lingual fine-tuning [60], 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. Our models are built using PyTorchLightning [25] and HuggingFace Transformers [91].

Research questions. We organize the presentation and discussion of our results in the remainder of this section using the following research questions:

Architecture

- **RQ1**: Which *pre-trained multilingual language model*, mBART or mT5, is best suitable for performing direct cross-lingual summarization on our dataset?

Table 5 Comparison between mBART and mT5 on our cross-lingual TLDR dataset: we report ROUGE-1,-2 and -L scores as well as how many summaries each model can generate in a second (# Sum/S). Best results per language and metric are in bold, and significant difference within the sub-table of each language is marked with †

Lang	Model	R1 (avg)	R2 (avg)	RL (avg)	# Sum /Sec
DE	mBART	**19.29**†	**5.46**	**16.02**	12.40
	mT5	17.99	4.09	15.03	6.08
IT	mBART	21.20	6.15	17.54	14.16
	mT5	**21.61**†	**6.43**†	**18.63**†	6.38
ZH	mBART	23.03	5.76	20.27	16.95
	mT5	**23.49**	**5.73**	**20.41**	6.16
JA	mBART	**30.94**†	4.66	20.34	11.97
	mT5	30.51	**4.93**	**20.55**	6.02

- **RQ2**: Which *stacking order* in the pipeline approach—i.e., first summarize and then translate or vice versa—performs better for two-stage cross-lingual summarization?
- **RQ3**: Which *architecture*—i.e., two-stage or direct cross-lingual summarization (Sects. 5.1 and 5.2)—is overall best suited for the CL-TLDR task?
- **RQ4**: How do the results compare for *different kinds of cross-lingual data*, i.e., different portions of our dataset (X-SCITLDR-PostEdit vs. -Human, Sect. 4)?
- **RQ5**: Does intermediate-stage fine-tuning help improve direct CL-TLDR summarization?
- **RQ6**: Can we reduce model size while keeping its summarization ability by knowledge distillation?

Analysis

- **RQ7**: How challenging is code-switching summary generation for direct cross-lingual summarization models?
- **RQ8**: How much data do we need to perform cross-lingual TLDR summarization? What is the performance in a zero-shot or few-shot setting?
- **RQ9**: What are the major kinds of errors that can be found by manual inspection of the summaries generated by direct cross-lingual models?

RQ1: mBART vs mT5. All our models, be it either a pipeline architecture or a direct cross-lingual summarizer, rely for summarization on a transformer-based multilingual pre-trained language model (cf. Sect. 5). Since different large-scale multilingual generative pre-training language models available to perform cross-lingual text generation, we first investigate *which* model to use: consequently, in this first set of experiments, we compare two popular models, namely mBART [51] and mT5 [92], which support all languages contained in our X-SCITLDR dataset and have been

[9] https://www.deepspeed.ai

Table 6 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. [10]. 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 †

Lang	Model	R1 (avg)	R2 (avg)	RL (avg)	R1 (max)	R2 (max)	RL (max)
DE	EnSum-MT	**19.29**	**5.46**	**16.02**	**30.74**	**13.37**	**26.61**
	TGTSum-MT	19.21	5.03	15.66	29.99	12.11	25.35
	CLSum	17.99	3.58	14.69	27.44	8.54	23.05
	CLSum+EnSum	18.06	3.61	14.75	27.36	8.47	23.04
	CLSum+MT	18.47	4.16	15.25	28.84	9.91	24.37
IT	EnSum-MT	20.76	6.88	17.46	31.53	14.96	27.51
	TGTSum-MT	21.07	6.82	17.63	31.21	14.24	26.93
	CLSum	21.20	6.15	17.54	30.98	12.77	26.25
	CLSum+EnSum	20.47	6.14	17.39	30.13	12.61	26.32
	CLSum+MT	**21.71**†	**7.04**	**18.11**†	**32.34**	**14.44**	**27.76**
ZH	EnSum-MT	**27.06**	**8.69**	**23.26**	**40.41**	**18.18**	**35.39**
	TGTSum-MT	26.39	8.00	22.40	38.68	16.56	33.49
	CLSum	23.03	5.76	20.27	34.11	11.77	30.12
	CLSum+EnSum	22.62	5.52	19.88	33.42	11.43	29.45
	CLSum+MT	23.28	5.97	20.27	35.15	12.54	30.72

shown to achieve state-of-the-performance on automated text summarization.

We compare these two models along two different dimensions, namely overall summarization performance using ROUGE scores, as well as efficiency, here measured as number of summaries produced per second. Comparison of ROUGE scores and summary generation speed for both models on the direct cross-lingual summarization setup (Sect. 5.2) is shown in Table 5. We observe that the two models perform on par across different languages, e.g., mBART shows better performance on German, whereas mT5 slightly outperforms it on Italian and Chinese (possibly because of a larger exposure during pre-training to these languages).[10] mT5, however, has twice more parameters, which leads to more expensive computational costs, as highlighted by the model being able to generate half the summaries per second when compared with mBART. Given the comparable performance, yet this major difference in text generation speed, we opt for mBART in the remainder of our experiments.

RQ2: summarize and translate versus translate and summarize. We next conduct experiments to compare two ways to implement a pipeline architecture for cross-lingual summarization using a summarization and translation module in different orders, namely a 'summarize and translate' (EnSum-MT) or 'translate and summarize' (TGTSum-MT) approach (Sect. 5.1).

Table 7 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

Model	Rouge-1	Rouge-2	Rouge-L
EnSum-MT	24.38	4.42	16.54
TGTSum-MT	32.24	5.65	20.46
CLSum	30.94	4.66	20.34
CLSum+EnSum	**32.30**	**5.66**	**20.89**
CLSum+MT	**32.30**	5.47	20.85

Results on X-SCITLDR-PostEdit (German, Italian and Chinese) and X-SCITLDR-Human (Japanese) portions of our dataset are shown in Tables 6 and 7, respectively. We notice that performance varies greatly across the two different portions of the dataset. On the automatically translated, post-edited portion of the data (Table 6) we observe no major difference in ROUGE scores between the two different architectures. On the contrary, on the manually generated, expert-authored portion (Table 7) TGTSum-MT outperforms EnSum-MT by a large margin. This is because EnSum-MT relies on English monolingual summarization using English-only SCITLDR as training data for fine-tuning. On the contrary, TGTSum-MT uses target language-specific models, which have been fine-tuned on the respective languages using multilingual summaries from X-SCITLDR. Using language-specific fine-tuning can, in turn, better account

[10] ROUGE-2 scores are (much) lower than other ROUGE metrics since it computes matches of consecutive bigrams, which is harder and thus less frequent than matching unigrams (evaluated by ROUGE-1) or non-consecutive longest common sequences (evaluated by ROUGE-L), and which is in line with previous works [56, 75]

Table 8 Example of gold-standard summaries and automatically generated versions

a) gold standard

Abstract: Convolution acts as a local feature extractor in convolutional neural networks (CNNs). However, the convolution operation is not applicable when the input data is supported on an irregular graph such as with social networks, citation networks, or knowledge graphs. This paper proposes the topology adaptive graph convolutional network (TAGCN), a novel graph convolutional network that generalizes CNN architectures to graph-structured data and provides a systematic way to design a set of fixed-size learnable filters to perform convolutions on graphs. [...]
TLDR: The paper introduces Topology Adaptive GCN to generalize convolutional networks to graph-structured data.
German TLDR: Die Arbeit führt Topology Adaptive GCN ein, um Convolutional Networks auf graphstrukturierte Daten zu verallgemeinern.

b) automatically generated summaries

EnSum-MT: In diesem Beitrag wird das topologieadaptive graphische Faltungsnetzwerk (TAGCN) vorgeschlagen, das CNN-Architekturen auf graphisch strukturierte Daten verallgemeinert und einen systematischen Weg zur Entwicklung einer Reihe von lernfähigen Filtern fester Größe zur Durchführung von Faltungen auf Graphen bietet.
CLSum: Wir schlagen das topologie adaptive graph convolutional network (TAGCN) vor, ein neuartiges graphisches Convolutional Network, das CNN-Architekturen auf graphenstrukturierte Daten verallgemeinert.

Table 9 Average summary length (number of tokens) of pipeline-based versus cross-lingual models

Language	EnSum-MT	TGTSum-MT	CLSum
DE	23.48	27.04	22.94
IT	24.17	25.12	22.73
ZH	25.90	28.17	19.76
JA	30.50	55.71	56.76

for different styles across different subsets of our data (cf. different summary length and compression ratio of JA vs. DE/IT/ZH in Table 4), thus allowing the generation of summaries that are best aligned with the test data.

RQ3: two-stage versus direct cross-lingual summarization. We next compare our two main architectures, namely the pipeline models (Sect. 5.1, EnSum-MT/TGTSum) with a multilingual encoder–decoder architecture based on mBART (Sect. 5.2, CLSum/+EnSum/+MT), again on the basis of the results on the two main portions of our dataset, i.e., post-edited translations and human-generated summaries, from Tables 6 and 7, respectively.

Similarly to the case of RQ2, we again see major performance differences across the two dataset portions. While MT-based summarization (EnSum-MT/TGTSum) is superior or comparable when evaluated on translated/post-edited TLDRs in German, Italian and Chinese (Table 6), the direct cross-lingual summarization model (CLSum) improves results on native Japanese summaries by a large margin when compared to the 'summarize and translate' EnSum-MT pipeline (Table 7). The differences in performance figures between X-SCITLDR-PostEdit and X-SCITLDR-Human 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 EnSum-MT 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. As a result of this, models that have been trained on target language-specific data like 'translate and summarize' (TGTSum-MT) and direct cross-lingual summarization (CLSum) are better adapted to style variations across different portions and languages of our cross-lingual dataset and thus achieve better results.

RQ4: post-edited versus human-generated cross-lingual summaries. To better understand the behavior of the system in light of the different performance on post-edited versus human-generated data, we manually inspected the output of the two systems. Table 8 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 use 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 [85]. Table 9

Table 10 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

presents the average summary lengths in different languages for our MT-based and cross-lingual models: the numbers show that the summaries of CLSum are indeed shorter than those generated from the pipeline models for those languages that are found in the X-SCITLDR-PostEdit portion of our dataset (DE/IT/ZH). Japanese summaries generated using models that went through language-specific training (namely TGTSum-MT and CLSum) are instead longer: this is because the reference summaries used for training are comparably longer than those in SCITLDR (cf. Table 4, column 8 and 9) and thus allow for language- and style-specific adaptation.

Within the post-edited portion of our dataset, EnSum-MT performs significantly better than the cross-lingual models in German and Chinese; however, there is generally no significant difference with cross-lingual models in Italian, where CLSum+MT is even able to achieve statistically significant improvements on average Rouge-1 and Rouge-L. To better understand such different behavior across languages, we computed for each language the word-level Jaccard coefficients between the automatically translated summaries and their post-edited versions in X-SCITLDR-PostEdit.

$$\mathrm{Jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

The Jaccard coefficient takes two sets, i.e., in our case words from automatically translated English TLDR summaries and their human-corrected version, and computes the ratio of overlapping elements to all elements in both sets, thus making it possible to quantify the rate of human correction as the ratio of words that are fixed during the process of post-editing. As Table 10 shows, the Italian post-edited translations contain more edits than the other two languages. This, in turn, seems to disadvantage the two-stage EnSum-MT pipeline, whose output aligns more with the ‘vanilla’ automatic translations.

We notice also differences in absolute numbers between German, Italian and Chinese, which could be due to the distribution of training data used to train the multilingual transformer [57], with mBART being trained on more Chinese than Italian data. However, German performs worst among the three languages, despite mBART being trained on more German data than Chinese or Italian. 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 preprocessing [28], 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 8). 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 [48].

RQ5: The potential benefits of intermediate fine-tuning. In the next set of experiments, we investigate whether intermediate-stage training, which aims at learning the summarization and translation tasks from additional data, can improve the performance of a vanilla cross-lingual model. 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 6) 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 a new language (i.e., English and the target language). On the contrary, CLSum+EnSum improves on all ROUGE metrics for Japanese (Table 7). This is because, as previously mentioned, the Japanese data have a different style from those from SCITLDR consequently, English TLDRs provide an additional training signal that helps to improve results for the summarization task.

Including MT-based pre-training, i.e., fine-tuning mBART on texts that have been automatically translated 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 Ladhak et al. [46] 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.

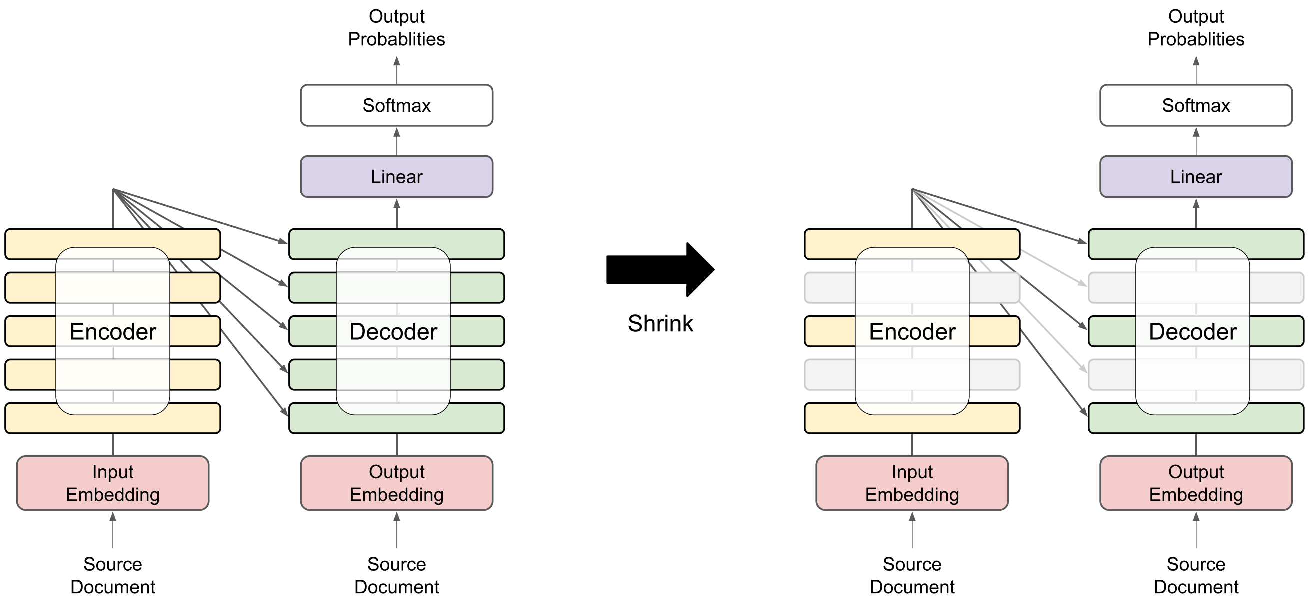


Fig. 5 Overview of the 'shrink and fine-tune' procedure: **a)** we fine-tune the teacher model in its original size; **b)** we shrink its size by selecting subsets of layers from the encoder and/or decoder and copying them into the student model; **c)** we fine-tune the student

RQ6: The effects of knowledge distillation. We now investigate the effect of knowledge distillation for cross-lingual summarization models on our dataset. In our experiments we use the 'shrink and fine-tune' (SFT) distillation introduced in Shleifer and Rush [83], as it has been shown to perform well on summarization while being conceptually simple. Using this method, we initialize a smaller model by taking the parameters from a fine-tuned full-sized model before fine-tuning it again on our dataset. Specifically, this method extracts the smaller student model by taking the maximally spaced layers of a fine-tuned teacher (with ties being broken arbitrarily), copying the selected layers from teacher to student and re-fine-tuning the student model (see Fig. 5 for a schematic overview). For example, when creating a distilled BART model with three encoder and three decoder layers (referred to as a 3-3 student model) from a full 12-12 BART teacher, we copy layers 0, 6 and 11 of both encoder and decoder to the student before re-fine-tuning it. Consequently, to understand the effect of layer selection, we conduct experiments with several combinations of layers of the teacher model to be used to initialize the student model and analyze their performance on the cross-lingual summarization task.

The results of layer selection for SFT knowledge distillation are presented in Table 11, where we denote with N-M the number of layers copied from the teacher encoder and decoder to the student. To highlight the reduction in computational costs provided by the student, i.e., distilled, models, we complement ROUGE scores with the number of parameters and the number of summaries that a model can generate in a second in our infrastructure. To obtain the latter, we take the best-performing CLSum model for each target language and generate summaries on a single NVIDIA RTX A6000 GPU with batch size 32, number of beams equal to 3 and 1.0 as the repetition penalty rate. The number of summaries per second is only comparable within a language as the generating speed is highly dependent on the output sequence length which varies for each language and data source. Reducing the number of layers in half (i.e., 6–6) does not reduce the parameters by half since mBART contains, in addition to the encoder and decoder, an embedding layer and a final prediction head. As shown in the comparison of 12–3 and 3–12, removing layers in the decoder results in greater parameter reduction and faster inference since one decoder layer contains more parameters than an encoder layer, due to the additional encoder–decoder attention parameters which are not present in the encoder layers. By removing the same number of layers from the encoder and decoder, as in the 12–3 versus 3–12 and 12-6 versus 6–12 setups, we observe that removing layers from the encoder has a stronger negative impact on the ROUGE score and provides lower inference time speedup, in line with previous findings by Shleifer and Rush [83]. From this we can draw a useful practical tip for future work: when distilling summarization models, it is more cost- and performance-efficient to reduce the layers in the decoder.

The impact of distillation on the ROUGE score is highly dependent on the target language. For German and Japanese, the performance drop is minor even when one-fourth of the layers is removed. In contrast, in Italian and Chinese ROUGE scores can drop up to 6.35 points which reflects a large degradation of the summary quality.

Table 11 Comparison between the original models and the distilled students: scores in parentheses report the difference with the original model. For each language, we order the models by the number of parameters. The second column shows how many layers are transferred from the encoder and decoder: e.g., 12–3 indicates that twelve layers from the encoder and three layers from the decoder were copied into the student, while 12–12 represents the original model

Lang	Layers	# Params	# Summaries per Second		ROUGE-1 (avg)	
DE	12-12	610M	12.40		17.99	
	3-3	346M	18.87	(+52.18%)	16.40	(-1.59)
	6-6	434M	16.29	(+31.37%)	16.94	(-1.05)
	12-3	459M	17.10	(+37.90%)	16.55	(-1.44)
	3-12	497M	13.63	(+9.92%)	16.07	(-1.92)
	12-6	510M	15.02	(+21.13%)	18.28	(+0.29)
	6-12	535M	12.48	(+0.65%)	17.63	(-0.36)
IT	12-12	610M	14.16		21.20	
	3-3	346M	26.23	(+85.24%)	14.85	(-6.35)
	6-6	434M	18.53	(+30.86%)	19.05	(-2.15)
	12-3	459M	21.64	(+52.82%)	19.20	(-2.00)
	3-12	497M	14.50	(+2.40%)	17.54	(-3.66)
	12-6	510M	17.34	(+22.46%)	20.61	(-0.59)
	6-12	535M	14.33	(+1.20%)	19.84	(-1.36)
ZH	12-12	610M	16.95		23.03	
	3-3	346M	28.53	(+68.22%)	18.13	(-4.90)
	6-6	434M	23.72	(+39.86%)	21.14	(-1.89)
	12-3	459M	25.90	(+52.71%)	21.61	(-1.42)
	3-12	497M	18.70	(+10.26%)	19.06	(-3.97)
	12-6	510M	20.10	(+18.51%)	21.92	(-1.11)
	6-12	535M	19.05	(+12.32%)	21.41	(-1.62)
JA	12-12	610M	11.97		30.94	
	3-3	346M	17.54	(+46.41%)	31.17	(+0.23)
	6-6	434M	14.81	(+23.62%)	29.87	(-1.07)
	12-3	459M	16.00	(+33.56%)	32.14	(+1.20)
	3-12	497M	12.27	(+2.42%)	28.85	(-2.09)
	12-6	510M	14.93	(+24.62%)	30.97	(+0.03)
	6-12	535M	12.66	(+5.68%)	30.82	(-0.12)

Table 12 CLSum model coverage of ‘copied‘ words inclusion in generated summaries on the test split of the post-edited portion of our dataset

Language	Coverage
DE	28.95
IT	33.67
ZH	21.07

RQ7: The ability to retain English domain terminology (‘code-switching’). Much domain terminology, especially in technical fields, is in English. As a matter of fact, one of the two main types of correction performed by the human annotators in order to fix the automatically generated translations was to create ’English-preserving translations’, which was often done to include the original English terms (see examples in Table 3). To better understand how much our direct cross-lingual models are able to generate these code-switched summaries,[11] we collect words ‘copied’ from the original English summaries by extracting overlapping words between original English summaries and post-edited versions, and compute the coverage of such words within summaries generated by direct cross-lingual summarization models (CLSum) on X-SCITLDR-PostEdit. Concretely, we compute coverage as follows:

$$\text{Cov} = \sum_{i=1}^{N} \frac{|E_i \cap T_i \cap H_i|}{|E_i \cap T_i|}$$

[11] The term ‘code-switching’ or ‘code-mixing’ is used to refer to texts that alternate between multiple (natural) languages [21]. In our case, code-switching does not cover common English words but rather technical terms in computer science. However, for the sake of illustration, we use ‘code-switching’ to express this domain-specific phenomenon as well.

Table 13 Performance in zero-shot settings. No intermediate fine-tuning (CLSum) versus 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	4.83	0.97	4.41
	CLSum+EnSum	5.87	1.29	5.35
	CLSum+MT	16.11	3.48	12.38
ZH	CLSum	0.64	0.06	0.61
	CLSum+EnSum	0.79	0.10	0.76
	CLSum+MT	17.88	3.60	13.95
JA	CLSum	2.34	0.59	2.06
	CLSum+EnSum	2.37	0.68	2.17
	CLSum+MT	29.43	4.29	18.27

where N, E_i, T_i, H_i are the number of summaries in the dataset, the set of words (i.e., unigrams) from the original English reference summary, its post-edited version in the target language and a generated summary, respectively. We present the coverage of code-switching words in Table 12. Overall, the remarkably low scores indicate the difficulty of including those words in summaries, especially in Chinese. With respect to Italian and German, Chinese is the language most typologically distant from English, which is the most common language that takes part in code-switching in our dataset. Generally, manual inspection of the output reveals how our CL summarization models have only limited ability to generate (often, English) domain-specific terminology, a point to which we come back later in the qualitative evaluation.

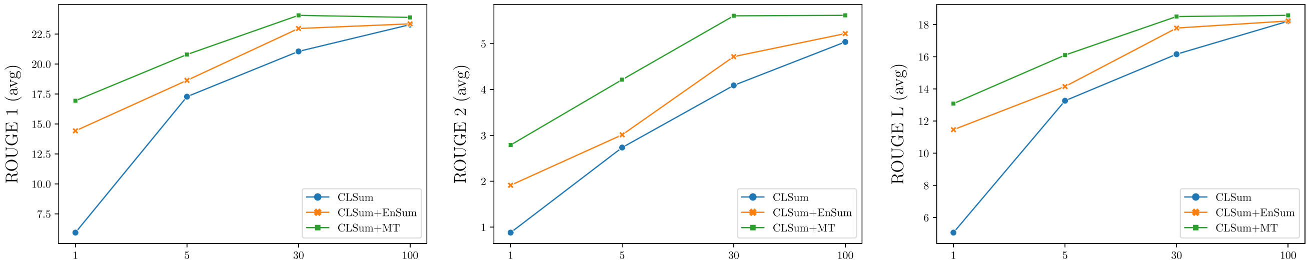
RQ8: Zero- and few-shot experiments. To better understand the contribution of intermediate fine-tuning and to analyze performance in the absence of multilingual summarization training data (i.e., in zero-shot settings), we present experiments in which we compare: a) using mBART with no fine-tuning (CLSum); b) fine-tuning mBART on English SCITLDR data only and evaluating 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 constructing cross-lingual 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 13. 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. [10] 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 [37]). 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 thus provide a strong baseline.

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 6 shows few-shot results averaged across all four target languages (detailed per-language results are given in Table 14) 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 a small amount of 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).

RQ9: Common errors of direct cross-lingual summarization models. To shed light on the main errors found in the summaries generated by direct cross-lingual summarization models, we conclude our empirical analysis by presenting a list of limitations revealed through manual quality evaluation. Such inspection of the output of the CLSum model shows that, while using pre-trained language models ensures that the output is generally fluent, the generated summaries still have limitations along two main dimensions:

- **Domain specificity and technical terminology**: summaries fail to include important 'keywords' that define fine-grained aspects of the employed models, developed methodology or experimental evaluation they perform. For instance, in Table 15a) both abstract and reference summary contain the expression 'catastrophic forgetting'



(a) Rouge-1 (avg)

(b) Rouge-2 (avg)

(c) Rouge-L (avg)

Fig. 6 Few-shot results without (CLSum) and with intermediate task (+EnSum) and cross-lingual fine-tuning (+MT) for different sizes of the training data in the target language (i.e., different number of *shots*): 1%, 5%, 30% and 100% of the X-SCITLDR training set of each target language

Table 14 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

		Rouge-1 (avg)				Rouge-2 (avg)				Rouge-L (avg)			
Lang	Model	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	11.55	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

in English and German, respectively, which plays an important role to explain the corresponding paper. Such notion, however, is missing in the generated summary.

- **Overly generic summaries**: summaries are correct, but are too generic in that they do not cover enough specific details of the scientific paper. For instance, in the automatically generated Italian summary from Table 15b), there is no mention of 'confidence thresholding' or 'defense against adversarial examples.'

Notice how both of these errors are critical for summarization in the scholarly domain, since information seeking in this domain is heavily focused on domain-specific information and technical lingo that is specific to certain research communities.

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 two-stage 'summarize and translate' (or vice-versa) approaches and a direct cross-lingual approach with two different underlying models. We additionally explored the potential benefits of intermediate task and cross-lingual fine-tuning and analyzed the performance in zero- and few-shot scenarios as well as the model's behavior on 'code-switched' texts. Furthermore, we conducted extensive experiments with a knowledge distillation approach aimed at reducing model size in a performance-preserving way. For future work, we plan to extend X-SCITLDR to include papers from research communities other than computer science or other STEM-fields, specifically those that use languages other than English for professional communication (e.g., humanities in German-speaking countries). From a methodological perspective, we plan to investigate how to apply additional techniques designed for cross-lingual text generation such as training with multiple decoders [99], automatically complementing our multilingual TLDRs with visual summaries [93], as well as devising new methods to include background knowledge such as, in our case, technical terminology and domain adaptation capabilities [96], into multilingual pre-trained models.

Our work crucially builds upon recent advances in multilingual pre-trained models [57] and cross-lingual summa-

Table 15 Two main categories of errors found in summaries generated by a multilingual pre-trained language model like BART: (a) domain specificity and technical terminology; (b) overly generic summaries

a) missing keywords

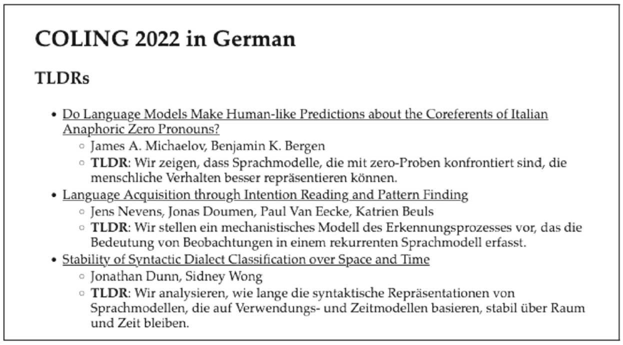
Abstract:	Incremental class learning involves sequentially learning classes in bursts of examples from the same class. This violates the assumptions that underlie methods for training standard deep neural networks, and will cause them to suffer from catastrophic forgetting. Arguably, the best method for incremental class learning is iCaRL, but it requires storing training examples for each class, making it challenging to scale. Here, we propose FearNet for incremental class learning. [...]
TLDR:	FearNet ist ein speichereffizientes neuronales Netz, das von der Gedächtnisbildung im Säugetiergehirn inspiriert ist und inkrementelles Klassenlernen ohne katastrophales Vergessen ermöglicht.
CLSum:	Wir schlagen FearNet vor, ein neuronales Lernsystem, das neue Erinnerungen aus einem Netzwerk für jüngste Erinnerungen inspiriert durch das medial prefrontale Kortex und das basolaterale amygdala verwendet.

b) too generic summary

Abstract:	Current machine learning algorithms can be easily fooled by adversarial examples. One possible solution path is to make models that use confidence thresholding to avoid making mistakes. Such models refuse to make a prediction when they are not confident of their answer. We propose to evaluate such models in terms of tradeoff curves with the goal of high success rate on clean examples and low failure rate on adversarial examples. [...]
TLDR:	Presentiamo una metrica e un attacco ottimale per valutare i modelli che si difendono dagli adversarial example usando confidence thresholding.
CLSum:	Proponiamo una famiglia di attacchi MaxConfidence per i modelli di apprendimento automatico.

rization [46] and investigates how these methodologies can be applied for multilingual scholarly document processing. In future, we propose to explore the direction of keyword-oriented summarization systems [22, 50] to enforce our models to include domain-specific terminology and have better overall focus.

The application of NLP techniques for mining scientific papers has been primarily focused on the English language: with 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., [73] and [6], *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 [43]) accessibility to scholarly resources.



COLING 2022 in German

TLDRs

- Do Language Models Make Human-like Predictions about the Coreferents of Italian Anaphoric Zero Pronouns?
  - James A. Michaelov, Benjamin K. Bergen
  - **TLDR**: Wir zeigen, dass Sprachmodelle, die mit zero-Proben konfrontiert sind, die menschliche Verhalten besser repräsentieren können.
- Language Acquisition through Intention Reading and Pattern Finding
  - Jens Nevens, Jonas Doumen, Paul Van Eecke, Katrien Beuls
  - **TLDR**: Wir stellen ein mechanistisches Modell des Erkennungsprozesses vor, das die Bedeutung von Beobachtungen in einem rekurrenten Sprachmodell erfasst.
- Stability of Syntactic Dialect Classification over Space and Time
  - Jonathan Dunn, Sidney Wong
  - **TLDR**: Wir analysieren, wie lange die syntaktische Repräsentationen von Sprachmodellen, die auf Verwendungs- und Zeitmodellen basieren, stabil über Raum und Zeit bleiben.

Fig. 7 A screenshot of a web page of papers from COLING 2022 with one sentence summaries in German at https://sotaro.io/info/2022_coling/2022.coling.de

8 Limitations

In this section, we follow recent proposals from major conferences in the Natural Language Processing community,[12] and present the limitations of our work. We hope to help researchers who plan to conduct further studies on cross-lingual summarization systems by hinting at possible extensions of this work.

Our work's first and foremost limitation is the X-SciTLDR dataset itself, as it only contains paper in the computer science domain. We are currently evaluating models trained with X-SciTLDR on social science papers in German for our ongoing VADIS project [44]. This, however, still covers a tiny fraction of all scholarly documents that could be processed by cross-lingual summarization systems.

Another limitation of this work is in the evaluation of model-generated summaries. Following prior research, we used ROUGE-1/2/L to compute performance scores on the summarization task. However, while ROUGE is the most widely used metric to evaluate summarization, several works from the literature show its problems and limitations, which call for more rigorous means of evaluation for summarization systems. One of the most reliable ways of evaluating is to hire human annotators to assess generated summaries. However, our experiments involve summaries in the scholarly domain for multiple languages, making manual evaluation highly expensive. Since the main objective of this work is constructing a multilingual gold standard and benchmarking existing models, we did not perform human evaluations on

[12] https://aclrollingreview.org/responsibleNLPresearch/

generated summaries. We plan to further investigate summary quality evaluation in detail in our future work.

Another limitation is that our analysis for research question 9 ("Common errors of direct cross-lingual summarization models") is qualitative rather than quantitative. However, we think such a level of analysis can shed light on future lines of research by identifying the weaknesses of the current summarization systems through manual inspections. We accordingly include our observations to facilitate further studies on cross-lingual extreme summarization.

Finally, this paper focuses on abstractive summarization using a multilingual pre-trained generative model as the underlying architecture. Much in line with other work in NLP, state-of-the-art models for summarization rely on transfer learning and the pre-trained model paradigm. We leave the exploration of cross-lingual extractive summarization, e.g., using multilingual embeddings, for future work.

Acknowledgements We thank Ines Rehbein and the anonymous reviewers for their helpful comments and also the support by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant INST 35/1597-1 FUGG.

Funding Open Access funding enabled and organized by Projekt DEAL. 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.

Data and code availability The X-SCITLDR corpus and the code used in our CL-TLDR experiments are available under an open license at https://github.com/sobamchan/xscitldr. We additionally provide a website(https://sotaro.io/tldrs) with automatically generated summaries for major natural language conferences (2021-ongoing) in English and the four other languages (German, Italian, Chinese and Japanese) that we cover in our study (see Fig. 7 for a screenshot).

Declarations

Conflict of interest The authors declare they have no financial interests.

References

1. Abu-Jbara, A., Radev, D.R.: Coherent citation-based summarization of scientific papers. In: Lin D, Matsumoto Y, Mihalcea R (eds) The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. In: Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA. The Association for Computer Linguistics, pp 500–509, (2011) https://aclanthology.org/P11-1051/
2. AbuRa'ed, A., Chiruzzo, L., Saggion, H., et al.: Lastus/taln @ clscisumm-17: Cross-document sentence matching and scientific text summarization systems. In: Jaidka K, Chandrasekaran MK, Kan M (eds) Proceedings of the Computational Linguistics Scientific Summarization Shared Task (CL-SciSumm 2017) organized as a part of the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2017) and co-located with the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2017), Tokyo, Japan, August 11, 2017, CEUR Workshop Proceedings, vol 2002. CEUR-WS.org, pp 55–66, (2017)http://ceur-ws.org/Vol-2002/talnclscisumm2017.pdf
3. AbuRa'ed, A., Bravo, À., Chiruzzo, L., et al.: Lastus/taln+inco @ cl-scisumm 2018 - using regression and convolutions for cross-document semantic linking and summarization of scholarly literature. In: Mayr P, Chandrasekaran MK, Jaidka K (eds) Proceedings of the 3rd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2018) co-located with the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2018), Ann Arbor, USA, July 12, 2018, CEUR Workshop Proceedings, vol 2132. CEUR-WS.org, pp 150–163, (2018) http://ceur-ws.org/Vol-2132/paper15.pdf
4. Accuosto, P., Saggion, H.: Mining arguments in scientific abstracts with discourse-level embeddings. Data Knowl Eng **129**(101), 840 (2020). https://doi.org/10.1016/j.datak.2020.101840
5. Accuosto, P., Neves, M., Saggion, H.: Argumentation mining in scientific literature: From computational linguistics to biomedicine. In: Frommholz I, Mayr P, Cabanac G, et al (eds) Proceedings of the 11th International Workshop on Bibliometric-enhanced Information Retrieval co-located with 43rd European Conference on Information Retrieval (ECIR 2021), Lucca, Italy (online only), April 1st, 2021, CEUR Workshop Proceedings, vol 2847. CEUR-WS.org, pp 20–36, (2021) http://ceur-ws.org/Vol-2847/paper-03.pdf
6. Adelani, D.I., Abbott, J., Neubig, G., et al.: Masakhaner: Named entity recognition for african languages. (2021) arxiv:2103.11811
7. Beltagy, I., Cohan, A., Feigenblat, G., et al.: Overview of the second workshop on scholarly document processing. In: Proceedings of the Second Workshop on Scholarly Document Processing. Association for Computational Linguistics, Online, pp 159–165, (2021) https://aclanthology.org/2021.sdp-1.22
8. Bhagavatula, C., Feldman, S., Power, R., et al.: Content-based citation 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, pp 238–251, (2018) https://doi.org/10.18653/v1/N18-1022, https://aclanthology.org/N18-1022
9. Bornmann, L., Mutz, R.: Growth rates of modern science: a bibliometric analysis based on the number of publications and cited references. J Assoc Inf Sci Technol **66**(11), 2215–2222 (2015). https://doi.org/10.1002/asi.23329
10. Cachola, I., Lo, K., Cohan, A., et al.: TLDR: Extreme summarization of scientific documents. In: Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics, Online, pp 4766–4777, (2020) https://doi.org/10.18653/v1/2020.findings-emnlp.428, https://aclanthology.org/2020.findings-emnlp.428
11. Cafarella, M.J., Anderson, M.R., Beltagy, I., et al.: Infrastructure for rapid open knowledge network development. AI Mag. **43**(1), 59–68 (2022). https://doi.org/10.1609/aimag.v43i1.19126

12. Chandrasekaran, M.K., Yasunaga, M., Radev, D., et al.: Overview and results: Cl-scisumm shared task 2019. (2019) arxiv:1907.09854
13. Cohan, A., Goharian, N.: Scientific document summarization via citation contextualization and scientific discourse. Int. J. Digit. Libr. **19**(2–3), 287–303 (2018). https://doi.org/10.1007/s00799-017-0216-8
14. Cohan, A., Dernoncourt, F., Kim, D.S., et al,: 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). Association for Computational Linguistics, New Orleans, Louisiana, pp 615–621, (2018a) https://doi.org/10.18653/v1/N18-2097, https://aclanthology.org/N18-2097
15. Cohan, A., Dernoncourt, F., Kim, D.S., et al.: 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). Association for Computational Linguistics, New Orleans, Louisiana, pp 615–621, (2018b) https://doi.org/10.18653/v1/N18-2097, https://aclanthology.org/N18-2097
16. Cohan, A., Feigenblat, G., Freitag, D., et al.: Overview of the third workshop on scholarly document processing. In: Proceedings of the Third Workshop on Scholarly Document Processing. Association for Computational Linguistics, Gyeongju, Republic of Korea, pp 1–6, (2022) https://aclanthology.org/2022.sdp-1.1
17. Collins, E., Augenstein, I., Riedel, S.: A supervised approach to extractive summarisation of scientific papers. In: Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017). Association for Computational Linguistics, Vancouver, Canada, pp 195–205, (2017). https://doi.org/10.18653/v1/K17-1021, https://aclanthology.org/K17-1021
18. Conneau, A., Lample, G., Rinott, R., et al.: XNLI: Evaluating cross-lingual sentence representations, (2018). arxiv:1809.05053, iSBN: 1809.05053 Publication Title: arXiv [cs.CL]
19. Conroy, J.M., Davis, S.T.: Section mixture models for scientific document summarization. Int. J. Digit. Libr. **19**(2–3), 305–322 (2018). https://doi.org/10.1007/s00799-017-0218-6
20. Daniele, F.: Performance of an automatic translator in translating medical abstracts. Heliyon **5**(10), e02687 (2019)
21. Doğruöz, A.S., Sitaram, S., Bullock, B.E., et al.: A survey of code-switching: Linguistic and social perspectives for language technologies. 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). Association for Computational Linguistics, Online, pp 1654–1666, (2021). https://doi.org/10.18653/v1/2021.acl-long.131, https://aclanthology.org/2021.acl-long.131
22. Dou, Z.Y., Liu, P., Hayashi, H., et al.: GSum: a general framework for guided neural abstractive summarization. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021, pp 4830–4842, (2021). https://aclanthology.org/2021.naacl-main.384/
23. Dror, R., Baumer, G., Shlomov, S., et al.: The hitchhiker's guide to testing statistical significance in natural language processing. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Melbourne, Australia, pp 1383–1392, (2018). https://doi.org/10.18653/v1/P18-1128, https://aclanthology.org/P18-1128
24. Esteva, A., Kale, A., Paulus, R., et al.: Covid-19 information retrieval with deep-learning based semantic search, question answering, and abstractive summarization. npj Digit. Med. **4**(1), 68 (2021). https://doi.org/10.1038/s41746-021-00437-0
25. Falcon, W., Borovec, J., Wälchli, A., et al.: Pytorchlightning/pytorch-lightning: 0.7.6 release. (2020). https://doi.org/10.5281/zenodo.3828935,
26. Fatima, M., Strube, M.: A novel wikipedia based dataset for monolingual and cross-lingual summarization. In: Proceedings of the Third Workshop on New Frontiers in Summarization. Association for Computational Linguistics, Online and in Dominican Republic, pp 39–50, (2021). https://doi.org/10.18653/v1/2021.newsum-1.5, https://aclanthology.org/2021.newsum-1.5
27. Fok, R., Head, A., Bragg, J., et al.: Scim: Intelligent faceted highlights for interactive, multi-pass skimming of scientific papers. CoRR abs/2205.04561. (2022). https://doi.org/10.48550/arXiv.2205.04561, arxiv:2205.04561
28. Frefel, D.: Summarization corpora of wikipedia articles. In: Calzolari N, Béchet F, Blache P, et al (eds) Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020. European Language Resources Association, Marseille, France, pp 6651–6655, (2020). https://aclanthology.org/2020.lrec-1.821/
29. Glavaš, G., Karan, M., Vulić, I.: XHate-999: analyzing and detecting abusive language across domains and languages. In: Proceedings of the 28th International Conference on Computational Linguistics. International Committee on Computational Linguistics, Barcelona, Spain (Online), pp 6350–6365, (2020). https://doi.org/10.18653/v1/2020.coling-main.559, https://aclanthology.org/2020.coling-main.559
30. Green, S., Heer, J., Manning, C.D.: The efficacy of human post-editing for language translation. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, CHI '13, p 439-448, (2013). https://doi.org/10.1145/2470654.2470718,
31. Guo, Y., Korhonen, A., Poibeau, T.: A weakly-supervised approach to argumentative zoning of scientific documents. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Edinburgh, Scotland, UK., pp 273–283, (2011). https://aclanthology.org/D11-1025
32. Gururangan, S., Marasović, A., Swayamdipta, S., et al.: Don't Stop Pretraining: Adapt Language Models to Domains and Tasks, (2020). arxiv:2004.10964, iSBN: 2004.10964 Publication Title: arXiv [cs.CL]
33. Hahn, U., Mani, I.: The challenges of automatic summarization. Computer **33**(11), 29–36 (2000). https://doi.org/10.1109/2.881692
34. He, Q., Kifer, D., Pei, J., et al.: Citation recommendation without author supervision. In: King I, Nejdl W, Li H (eds) Proceedings of the Forth International Conference on Web Search and Web Data Mining, WSDM 2011, Hong Kong, China, February 9-12, 2011. ACM, pp 755–764, (2011). https://doi.org/10.1145/1935826.1935926,
35. Hermann, K.M., Kocisky, T., Grefenstette, E., et al.: Teaching machines to read and comprehend. In: Cortes C, Lawrence N, Lee D, et al (eds) Advances in Neural Information Processing Systems, vol 28. Curran Associates, Inc., Montréal, Canada, pp 1693–1701, (2015). https://proceedings.neurips.cc/paper/2015/file/afdec7005cc9f14302cd0474fd0f3c96-Paper.pdf
36. Hinton, G.E., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. CoRR abs/1503.02531. (2015). arxiv:1503.02531,
37. Hu, J., Ruder, S., Siddhant, A., et al.: XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In: III HD, Singh A (eds) Proceedings of the 37th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 119. PMLR, Online, pp 4411–4421, (2020). https://proceedings.mlr.press/v119/hu20b.html

38. Huang, W., Wu, Z., Liang, C., et al.: A neural probabilistic model for context based citation recommendation. In: Bonet B, Koenig S (eds) Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA. AAAI Press, pp 2404–2410, (2015). http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9737
39. Hung, C.C., Lauscher, A., Vulić, I., et al.: Multi2WOZ: A robust multilingual dataset and conversational pretraining for task-oriented dialog. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Seattle, United States, pp 3687–3703, (2022). https://doi.org/10.18653/v1/2022.naacl-main.270, https://aclanthology.org/2022.naacl-main.270
40. Jaidka, K., Yasunaga, M., Chandrasekaran, M.K., et al.: The clscisumm shared task 2018: Results and key insights. (2019). arxiv:1909.00764
41. Jiao, X., Yin, Y., Shang, L., et al.: TinyBERT: Distilling BERT for natural language understanding. In: Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics, Online, pp 4163–4174, (2020). https://doi.org/10.18653/v1/2020.findings-emnlp.372, https://aclanthology.org/2020.findings-emnlp.372
42. Jin, D., Szolovits, P.: Hierarchical neural networks for sequential sentence classification in medical scientific abstracts. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Brussels, Belgium, pp 3100–3109, (2018). https://doi.org/10.18653/v1/D18-1349, https://aclanthology.org/D18-1349
43. Joshi, P., Santy, S., Budhiraja, A., et al.: The state and fate of linguistic diversity and inclusion in the NLP world. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, pp 6282–6293, (2020). https://doi.org/10.18653/v1/2020.acl-main.560, https://aclanthology.org/2020.acl-main.560
44. Kartal, Y.S., Takeshita, S., Tsereteli, T., et al.: Towards automated survey variable search and summarization in social science publications. CoRR abs/2209.06804. (2022). https://doi.org/10.48550/arXiv.2209.06804,
45. Kemker, R., McClure, M., Abitino, A., et al.: Measuring catastrophic forgetting in neural networks. In: McIlraith SA, Weinberger KQ (eds) Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th 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. AAAI Press, New Orleans, Louisiana, USA, pp 3390–3398, (2018). https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16410
46. Ladhak, F., Durmus, E., Cardie, C., et al.: WikiLingua: a new benchmark dataset for cross-lingual abstractive summarization. In: Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics, Online, pp 4034–4048, (2020). https://doi.org/10.18653/v1/2020.findings-emnlp.360, https://aclanthology.org/2020.findings-emnlp.360
47. Lahav, D., Saad-Falcon, J., Kuehl, B., et al.: A search engine for discovery of scientific challenges and directions. In: Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022. AAAI Press, pp 11,982–11,990, (2022) https://ojs.aaai.org/index.php/AAAI/article/view/21456
48. Lauscher, A., Ravishankar, V., Vulić, I., et al.: From Zero to Hero: On the Limitations of Zero-Shot Cross-Lingual Transfer with Multilingual Transformers, (2020). arxiv:2005.00633, iSBN: 2005.00633 Publication Title: arXiv [cs.CL]
49. Lewis, M., Liu, Y., Goyal, N., et al.: BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, pp 7871–7880, (2020). https://doi.org/10.18653/v1/2020.acl-main.703, https://aclanthology.org/2020.acl-main.703
50. Li, C., Xu, W., Li, S., et al.: Guiding generation for abstractive text summarization based on key information guide network. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). Association for Computational Linguistics, New Orleans, Louisiana, pp 55–60, (2018). https://doi.org/10.18653/v1/N18-2009, https://aclanthology.org/N18-2009
51. Li, H., Zhu, J., Zhang, J., et al.: Multimodal sentence summarization via multimodal selective encoding. In: Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020. International Committee on Computational Linguistics, Barcelona, Spain (Online), pp 5655–5667, (2020). https://doi.org/10.18653/v1/2020.coling-main.496, https://aclanthology.org/2020.coling-main.496
52. Li, Z., Wang, Z., Tan, M., et al.: DQ-BART: Efficient sequence-to-sequence model via joint distillation and quantization. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics, Dublin, Ireland, pp 203–211, (2022) https://aclanthology.org/2022.acl-short.22
53. Liang, Y., Duan, N., Gong, Y., et al.: XGLUE: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Online, pp 6008–6018, (2020). https://doi.org/10.18653/v1/2020.emnlp-main.484, https://aclanthology.org/2020.emnlp-main.484
54. Lin, C.Y., Hovy, E.: Automatic evaluation of summaries using n-gram co-occurrence statistics. In: Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, Edomnton, Canada, pp 150–157, (2003). https://aclanthology.org/N03-1020
55. Lin, H., Ng, V.: Abstractive summarization: a survey of the state of the art. In: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019. AAAI Press, pp 9815–9822, (2019). https://doi.org/10.1609/aaai.v33i01.33019815,
56. Liu, Y., Lapata, M.: Text summarization with pretrained encoders. 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). Association for Computational Linguistics, Hong Kong, China, pp 3730–3740, (2019). https://doi.org/10.18653/v1/D19-1387, https://aclanthology.org/D19-1387
57. Liu, Y., Gu, J., Goyal, N., et al.: Multilingual denoising pre-training for neural machine translation. Trans Assoc Comput Linguist **8**, 726–742 (2020). https://doi.org/10.1162/tacl_a_00343
58. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. (2019). arxiv:1711.05101
59. Lu, Y., Dong, Y., Charlin, L.: Multi-XScience: a large-scale dataset for extreme multi-document summarization of scientific articles.

In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Online, pp 8068–8074, (2020) https://doi.org/10.18653/v1/2020.emnlp-main.648, https://aclanthology.org/2020.emnlp-main.648

60. Maurya, K.K., Desarkar, M.S., Kano, Y., et al.: ZmBART: an unsupervised cross-lingual transfer framework for language generation. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, Online, pp 2804–2818, https://doi.org/10.18653/v1/2021.findings-acl.248, (2021) https://aclanthology.org/2021.findings-acl.248
61. Medić, Z., Snajder, J.: Improved local citation recommendation based on context enhanced with global information. In: Proceedings of the First Workshop on Scholarly Document Processing. Association for Computational Linguistics, Online, pp 97–103, (2020) https://doi.org/10.18653/v1/2020.sdp-1.11, https://aclanthology.org/2020.sdp-1.11
62. Meng, R., Thaker, K., Zhang, L., et al.: Bringing structure into summaries: a faceted summarization dataset for long scientific documents. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Association for Computational Linguistics, Online, pp 1080–1089, (2021) https://doi.org/10.18653/v1/2021.acl-short.137, https://aclanthology.org/2021.acl-short.137
63. Mihalcea, R., Tarau, P.: TextRank: bringing order into text. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Barcelona, Spain, pp 404–411, (2004) https://aclanthology.org/W04-3252
64. Mishra, R., Bian, J., Fiszman, M., et al.: Text summarization in the biomedical domain: a systematic review of recent research. Journal of Biomedical Informatics 52:457–467. (2014) https://doi.org/10.1016/j.jbi.2014.06.009, https://www.sciencedirect.com/science/article/pii/S1532046414001476, special Section: Methods in Clinical Research Informatics
65. Moosavi, N.S., Fan, A., Shwartz, V., et al.: (eds) Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing, Association for Computational Linguistics, Online, (2020) https://aclanthology.org/2020.sustainlp-1.0
66. Narayan, S., Cohen, S.B., Lapata, M.: Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Brussels, Belgium, pp 1797–1807,(2018) https://doi.org/10.18653/v1/D18-1206, https://aclanthology.org/D18-1206
67. Nenkova, A., McKeown, K.R.: Automatic summarization. Found. Trends Inf. Retr. **5**(2–3), 103–233 (2011). https://doi.org/10.1561/1500000015
68. Ouyang, J., Song, B., McKeown, K.: 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). Association for Computational Linguistics, Minneapolis, Minnesota, pp 2025–2031, (2019) https://doi.org/10.18653/v1/N19-1204, https://aclanthology.org/N19-1204
69. Perez-Beltrachini, L., Lapata, M.: Models and datasets for cross-lingual summarisation. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp 9408–9423, (2021) https://doi.org/10.18653/v1/2021.emnlp-main.742, https://aclanthology.org/2021.emnlp-main.742
70. Phang, J., Févry, T., Bowman, S.R.: Sentence encoders on stilts: supplementary training on intermediate labeled-data tasks. (2019) arxiv:1811.01088
71. Pruksachatkun, Y., Phang, J., Liu, H., et al.: Intermediate-task transfer learning with pretrained language models: When and why does it work? In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, pp 5231–5247, (2020) https://doi.org/10.18653/v1/2020.acl-main.467, https://aclanthology.org/2020.acl-main.467
72. Ronzano, F., Freire, A., Sáez-Trumper, D., et al.: Making sense of massive amounts of scientific publications: the scientific knowledge miner project. In: Cabanac G, Chandrasekaran MK, Frommholz I, et al (eds) Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL) co-located with the Joint Conference on Digital Libraries 2016 (JCDL 2016), Newark, NJ, USA, June 23, 2016, CEUR Workshop Proceedings, vol 1610. CEUR-WS.org, pp 36–41, (2016) http://ceur-ws.org/Vol-1610/paper5.pdf
73. Ruder, S., Constant, N., Botha, J., et al.: XTREME-R: towards more challenging and nuanced multilingual evaluation. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp 10,215–10,245, (2021a) https://doi.org/10.18653/v1/2021.emnlp-main.802, https://aclanthology.org/2021.emnlp-main.802
74. Ruder, S., Constant, N., Botha, J., et al.: XTREME-R: towards more challenging and nuanced multilingual evaluation, arxiv:2104.07412, iSBN: 2104.07412 Publication Title: arXiv [cs.CL] (2021b)
75. Rush, A.M., Chopra, S., Weston, J.: A neural attention model for abstractive sentence summarization. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Lisbon, Portugal, pp 379–389, (2015)https://doi.org/10.18653/v1/D15-1044, https://aclanthology.org/D15-1044
76. Saggion, H., Ronzano, F.: Scholarly data mining: making sense of scientific literature. In: 2017 ACM/IEEE Joint Conference on Digital Libraries, JCDL 2017, Toronto, ON, Canada, June 19-23, 2017. IEEE Computer Society, Toronto, ON, Canada, pp 346–347, (2017)https://doi.org/10.1109/JCDL.2017.7991622,
77. Saggion, H., AbuRa'ed, A., Ronzano, F,: Trainable citation-enhanced summarization of scientific articles. In: Cabanac G, Chandrasekaran MK, Frommholz I, et al (eds) Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL) co-located with the Joint Conference on Digital Libraries 2016 (JCDL 2016), Newark, NJ, USA, June 23, 2016, CEUR Workshop Proceedings, vol 1610. CEUR-WS.org, pp 175–186, (2016) http://ceur-ws.org/Vol-1610/paper20.pdf
78. Sandhaus, E.: The New York Times Annotated Corpus. (2008) https://doi.org/10.35111/77BA-9X74, https://catalog.ldc.upenn.edu/LDC2008T19, artwork Size: 3250585 KB Pages: 3250585 KB Type: dataset
79. Sanh, V., Debut, L., Chaumond, J., et al.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, arxiv:1910.01108, iSBN: 1910.01108 Publication Title: arXiv [cs.CL] (2019)
80. Scialom, T., Dray, P.A., Lamprier, S., et al.: MLSUM: the multilingual summarization corpus. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Online, pp 8051–8067, (2020) https://doi.org/10.18653/v1/2020.emnlp-main.647, https://aclanthology.org/2020.emnlp-main.647

81. See, A., Liu, P.J., Manning, C.D.: 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). Association for Computational Linguistics, Vancouver, Canada, pp 1073–1083, (2017) https://doi.org/10.18653/v1/P17-1099, https://aclanthology.org/P17-1099
82. Shleifer, S., Rush, A.M.: Pre-trained Summarization Distillation. https://doi.org/10.48550/arXiv.2010.13002, (2020a) , arXiv:2010.13002 [cs]
83. Shleifer, S., Rush, A.M.: Pre-trained summarization distillation. CoRR abs/2010.13002. arxiv:2010.13002, (2020b)
84. Sotudeh, S., Deilamsalehy, H., Dernoncourt, F., et al.: TLDR9+: a large scale resource for extreme summarization of social media posts. In: Proceedings of the Third Workshop on New Frontiers in Summarization. Association for Computational Linguistics, Online and in Dominican Republic, pp 142–151, https://doi.org/10.18653/v1/2021.newsum-1.15, https://aclanthology.org/2021.newsum-1.15 (2021)
85. Sun, S., Shapira, O., Dagan, I., et al.: How to compare summarizers without target length? pitfalls, solutions and re-examination of the neural summarization literature. In: Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation. Association for Computational Linguistics, Minneapolis, Minnesota, pp 21–29, https://doi.org/10.18653/v1/W19-2303, https://aclanthology.org/W19-2303 (2019)
86. Takeshita, S., Green, T., Friedrich, N., et al.:X-SCITLDR: cross-lingual extreme summarization of scholarly documents. In: Aizawa A, Mandl T, Carevic Z, et al (eds) JCDL '22: The ACM/IEEE Joint Conference on Digital Libraries in 2022, Cologne, Germany, June 20 - 24, 2022. ACM, pp 1–12, (2022) https://doi.org/10.1145/3529372.3530938,
87. Varab, D., Schluter, N.: MassiveSumm: a very large-scale, very multilingual, news summarisation dataset. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp 10,150–10,161, https://doi.org/10.18653/v1/2021.emnlp-main.797, https://aclanthology.org/2021.emnlp-main.797 (2021)
88. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol 30. Curran Associates, Inc., Long Beach, CA, USA, pp 5998–6008, (2017) https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
89. Verspoor, K., Cohen, K.B., Dredze, M., et al.: (eds) Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020, Association for Computational Linguistics, Online, (2020) https://aclanthology.org/2020.nlpcovid19-acl.0
90. Wang, J., Meng, F., Zheng, D., et al.: A Survey on Cross-Lingual Summarization. Trans. Assoc. Comput. Linguist. **10**, 1304–1323 (2022). https://doi.org/10.1162/tacl_a_00520
91. Wolf, T., Debut, L., Sanh, V., et al.: Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020. Association for Computational Linguistics, Online, pp 38–45, (2020). https://www.aclweb.org/anthology/2020.emnlp-demos.6
92. Xue, L., Constant, N., Roberts, A., et al.: 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. Association for Computational Linguistics, Online, pp 483–498, https://doi.org/10.18653/v1/2021.naacl-main.41, https://aclanthology.org/2021.naacl-main.41 (2021)
93. Yamamoto, S., Lauscher, A., Ponzetto, S.P., et al.: Visual summary identification from scientific publications via self-supervised learning. Front. Res. Metrics Anal. **6**, 719004 (2021). https://doi.org/10.3389/frma.2021.719004
94. Yang, S.T., Lee, P., Kazakova, L., et al.: Identifying the central figure of a scientific paper. In: 2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019. IEEE, Sydney, Australia, pp 1063–1070, https://doi.org/10.1109/ICDAR.2019.00173, (2019)
95. Yasunaga, M., Kasai, J., Zhang, R., et al.: ScisummNet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. Proc. Conf. AAAI Artif. Intell. 33:7386–7393. https://doi.org/10.1609/aaai.v33i01.33017386, https://wvvw.aaai.org/ojs/index.php/AAAI/article/view/4727, publisher: Association for the Advancement of Artificial Intelligence (AAAI) (2019)
96. Yu, T., Liu, Z., Fung, P.: AdaptSum: towards low-resource domain adaptation for abstractive summarization. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Online, pp 5892–5904, https://doi.org/10.18653/v1/2021.naacl-main.471, https://aclanthology.org/2021.naacl-main.471 (2021)
97. Zerva, C., Nghiem, M., Nguyen, N.T.H., et al.: Cited text span identification for scientific summarisation using pre-trained encoders. Scientometrics **125**(3), 3109–3137 (2020). https://doi.org/10.1007/s11192-020-03455-z
98. Zhang, S., Zhang, X., Bao, H., et al.: Attention temperature matters in abstractive summarization distillation. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Dublin, Ireland, pp 127–141, (2022) https://aclanthology.org/2022.acl-long.11
99. Zhu, J., Wang, Q., Wang, Y., et al.: 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). Association for Computational Linguistics, Hong Kong, China, pp 3054–3064, https://doi.org/10.18653/v1/D19-1302, https://aclanthology.org/D19-1302 (2019)