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

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

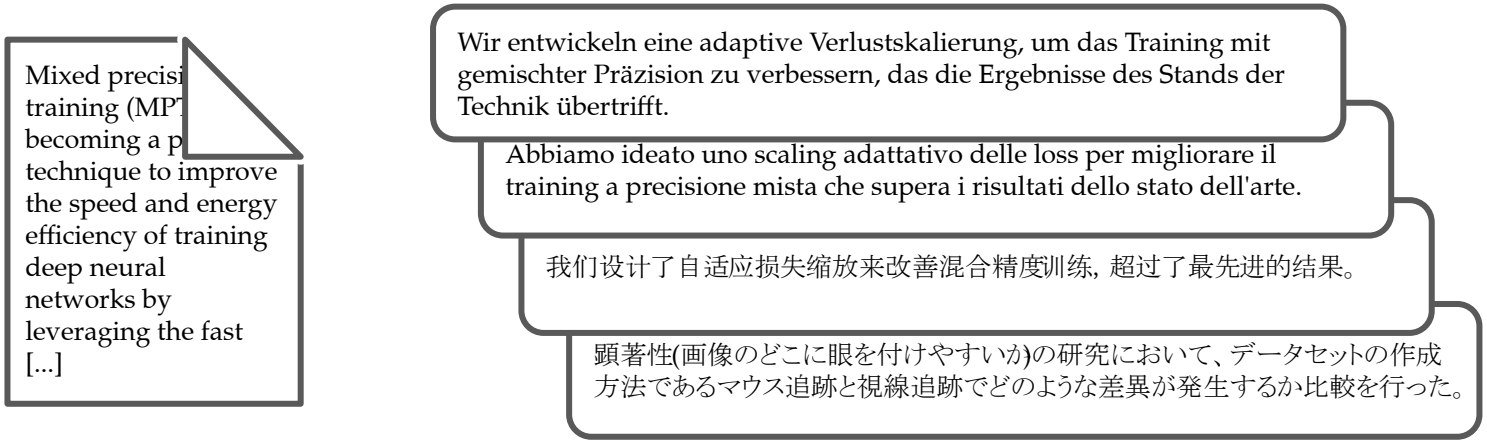
[1]Data and Web Science Group, University of Mannheim, Germany
[2]Web-based Information Systems and Services (WISS), Stuttgart Media University, Germany

UNIVERSITÄT MANNHEIM

HOCHSCHULE DER MEDIEN

X-SCITLDR

1. **New** dataset for extreme research papers summarization in four languages
2. Benchmark experiments with state-of-the-art summarization systems

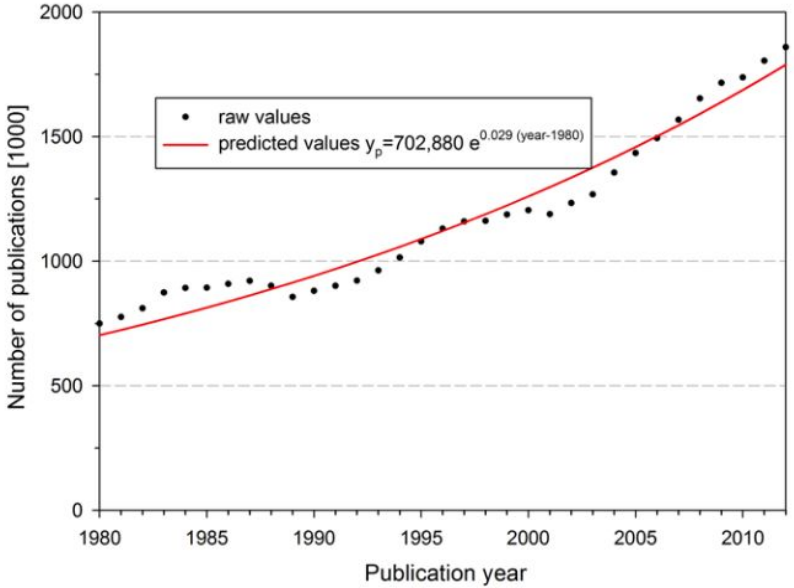


Papers in **English**

Summaries in
German, **Italian**, **Chinese**, **Japanese**

Background: Information Overload for Researchers

The number of publication is rapidly increasing *[Bornmann and Mutz 2015]*



Bornmann and Mutz (2015): **Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references.**

- Citation Recommendation
- Automatic Sentence Classification
- **Scholarly Document Summarization**

Background: Scholarly Document Summarization

Objective

Create a short summary of the research paper that covers its main points

Existing Datasets

SCITLDR *[Cachola et al. 2020]*, CSPubSum/CSPubSumExt *[Collins et al. 2017]*, ScisummNet *[Yasunaga et al. 2019]*

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.

Figure 1: A data sample in SCITLDR

Background: Scholarly Document Summarization

Objective

Create a short summary of the research paper that covers its main points

Existing Datasets

SCITLDR *[Cachola et al. 2020]*, CSPubSum/CSPubSumExt *[Collins et al. 2017]*, ScisummNet *[Yasunaga et al. 2019]*

Problem

Existing datasets are in English and cannot be used by non-English speaking researchers

Our Work

X-SCITLDR: the first cross-lingual scholarly document summarization dataset for four languages

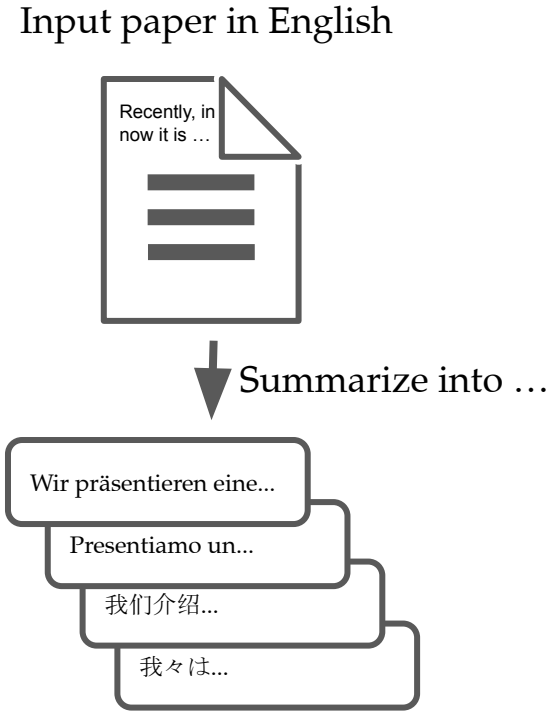
X-SCITLDR

A cross-lingual extreme summarization dataset of scholarly documents.

Goal: Make summarization systems available for non-English speakers

Composed of two main sources,

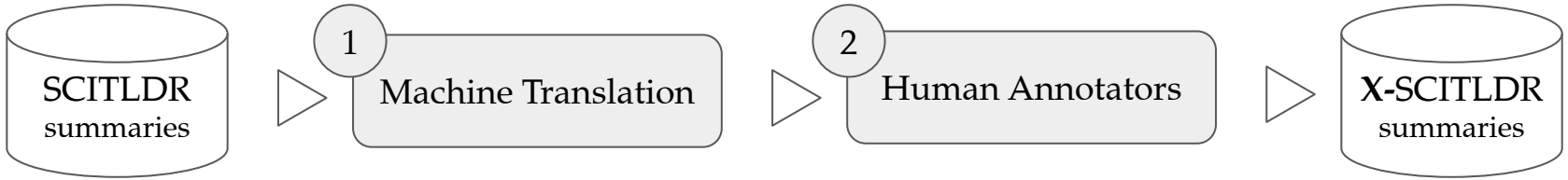
- **X-SCITLDR-PostEdit** (German, Italian and Chinese)
  - Post-edit translation of SCITLDR *[Cachola et al. 2020]*
- **X-SCITLDR-Human** (Japanese)
  - Human-written summaries from an online platform



X-SCITLDR: PostEdit

- SCITLDR *[Cachola et al. 2020]*
  - English extreme summarization dataset for research papers
  - Written by authors and reviewers on the OpenReview platform
  - One-sentence extreme summaries with high compression ratio

- Translation process
  1. Automatically translate summaries in SCITLDR by DeepL
  2. Manual correction by graduate students in CS (Post-editing *[Green et al. 2013]*)



X-SCITLDR: PostEdit - Wrong sense

Original Summary	The **paper** presents a multi-view framework for improving sentence representation in NLP tasks [...]
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 [...]

Table 1: Example of a post-editing correction

X-SCITLDR: PostEdit - English-preserving translation

Original Summary	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.

Table 2: Example of a post-editing correction

X-SCITLDR: Human

- arXivTimes (https://arxivtimes.herokuapp.com/)
  - Online platform actively updated by volunteers
  - Each post has a paper link and a human-written extreme summary
  - Extreme Summaries are written in Japanese

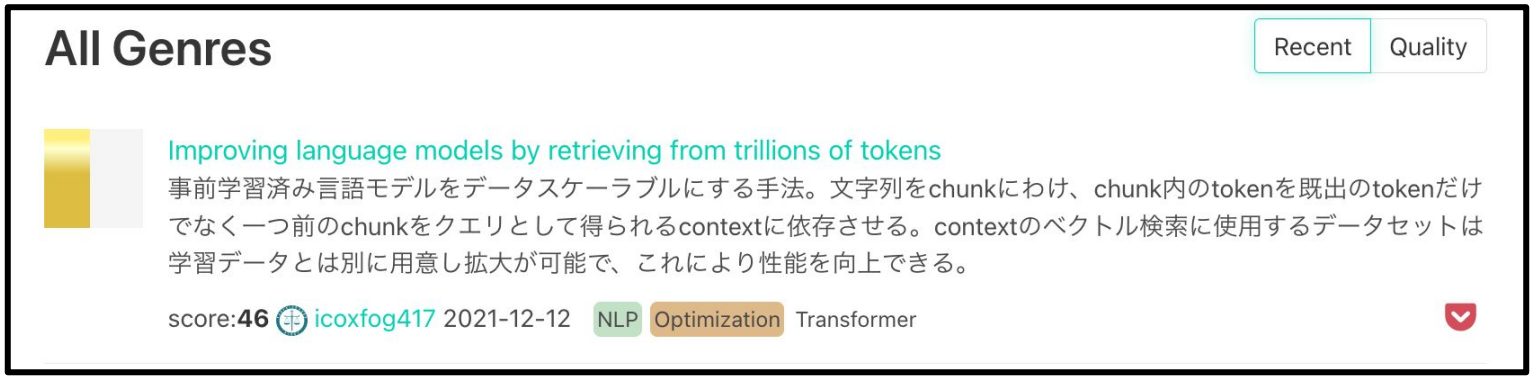


Figure 2: A screenshot of arXivTimes

X-SCITLDR: Dataset Statistics

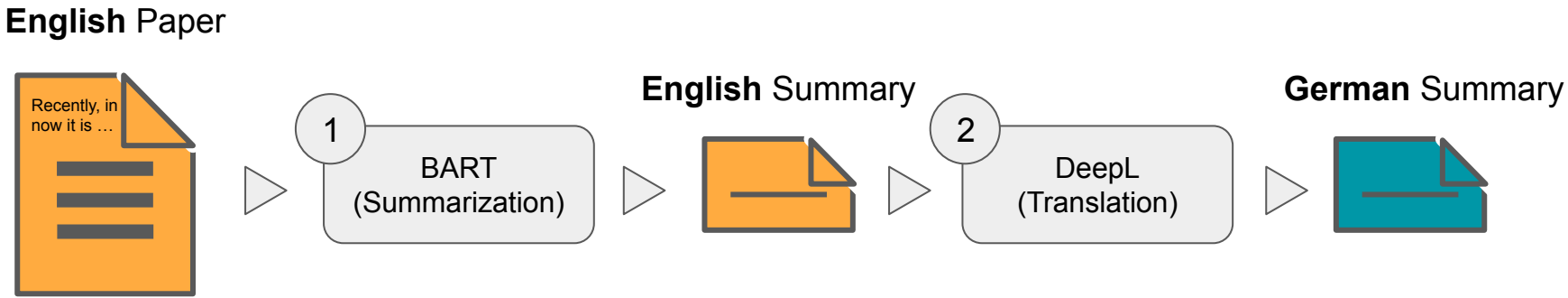
	Documents				Summaries			
	# documents (train/dev/test)	# words	vocabulary size	average # words per doc	# words	vocabulary size	average # words per summary	compression ratio (%)
EN	1,992/619/618	370,244	20,819	5,000	47,574	6,725	23.88	244.57
DE					43,929	13,808	22.05	264.87
IT					48,050	7,127	24.12	242.14
ZH					47,711	7,953	23.95	243.86
JA	1,606/199/199	306,815	14,769	10,000	121,989	6,706	75.91	131.73

Table 3: Overview statistics of X-SCITLDR

Short summaries in four languages!

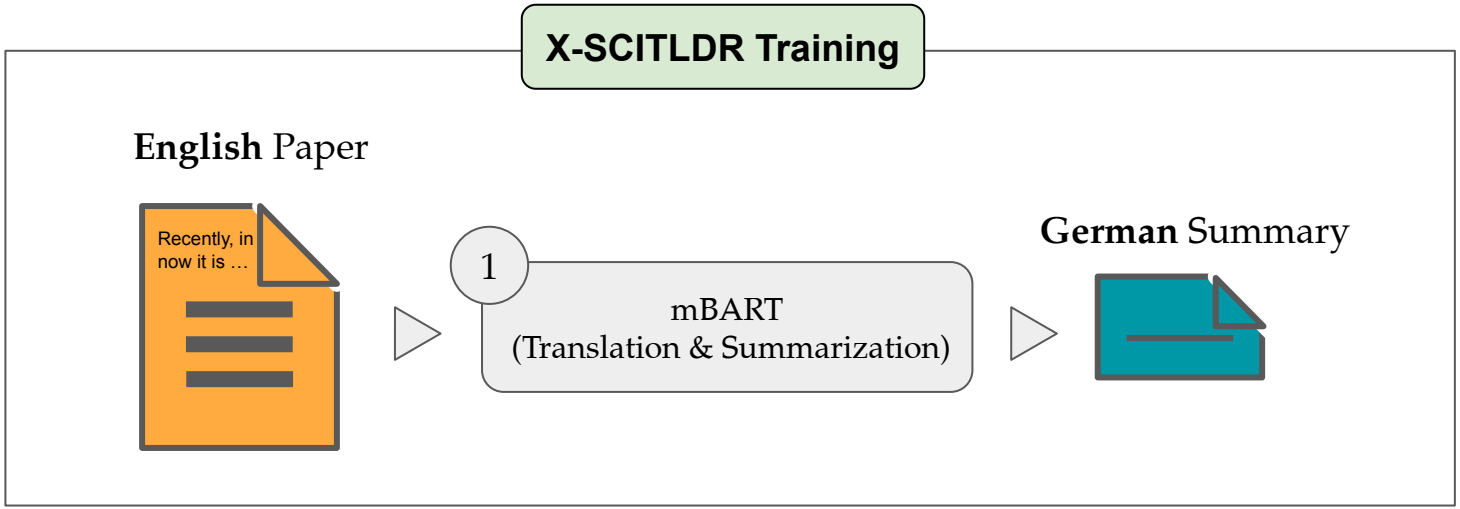
Benchmark: EnSum-MT

- Two-staged summarize-then-translate approach
- We use BART *[Lewis et al. 2020]* trained on SCITLDR and DeepL for translation



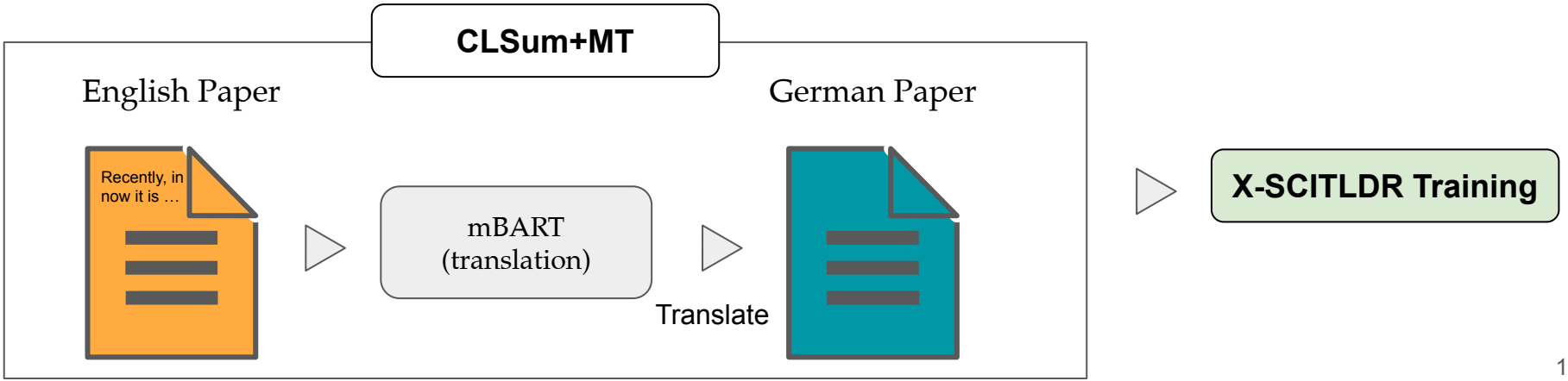
Benchmark: CLSum

- One-step cross-lingual summarization
- We use mBART *[Liu et al. 2020]* trained on our new X-SCITLDR



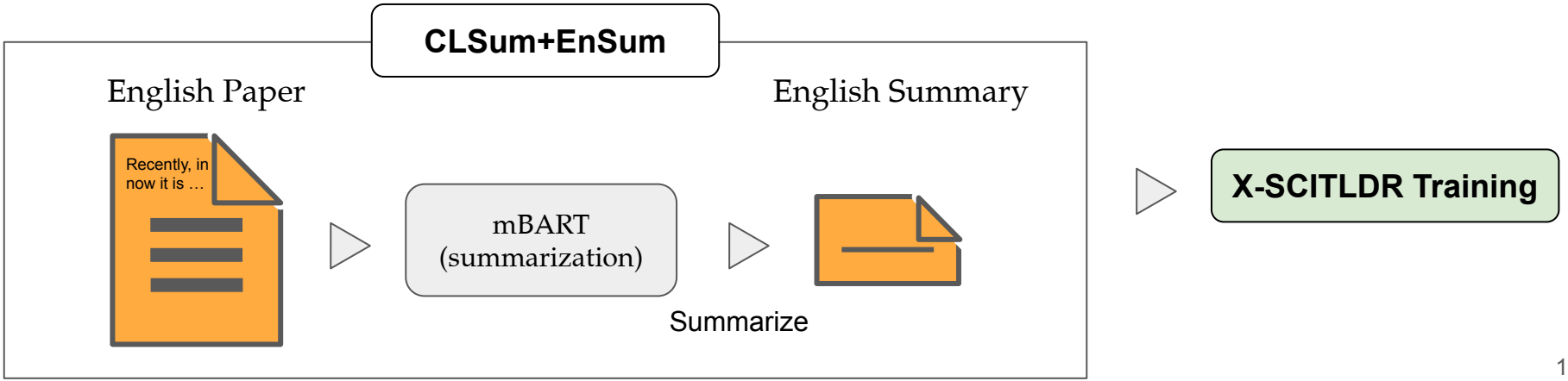
Benchmark: CLSum+MT

- One-step cross-lingual summarization
- We use mBART *[Liu et al. 2020]* trained on our new X-SCITLDR
- Two variants with intermediate fine-tuning *[Glavaš et al. 2020]*
  - CLSum+MT



Benchmark: CLSum+EnSum

- One-step cross-lingual summarization
- We use mBART *[Liu et al. 2020]* trained on our new X-SCITLDR
- Two variants with intermediate fine-tuning *[Glavaš et al. 2020]*
  - CLSum+MT
  - CLSum+EnSum



Research Questions

We aim to answer the following three questions,

- RQ1: Summarize-then-translate or direct cross-lingual summarization?
- RQ2: Does intermediate fine-tuning help?
- RQ3: How much training data do we need?

Benchmark: Results and Analysis for RQ1, RQ2

Lang	Model	R1 (avg)	R2 (avg)	RL (avg)
German	EnSum-MT	**19.29**	**5.46**	**16.02**
	CLSum	17.99	3.58	14.69
	CLSum+EnSum	18.06	3.61	14.75
	CLSum+MT	18.47	4.16	15.25
Italian	EnSum-MT	20.76	6.88	17.46
	CLSum	21.20	6.15	17.54
	CLSum+EnSum	20.47	6.14	17.39
	CLSum+MT	**21.71**†	**7.04**	**18.11**†
Chinese	EnSum-MT	**27.06**	**8.69**	**23.26**
	CLSum	23.03	5.76	20.27
	CLSum+EnSum	22.62	5.52	19.88
	CLSum+MT	23.28	5.97	20.27

Table 4: Results on the X-SCITLDR-PostEdit (German, Italian and Chinese)

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

Table 5: Results on the X-SCITLDR-Human (Japanese)

- EnSum-MT (Summarize-then-translate) is better for German and Chinese

Benchmark: Results and Analysis for RQ1, RQ2

Lang	Model	R1 (avg)	R2 (avg)	RL (avg)
German	EnSum-MT	**19.29**	**5.46**	**16.02**
	CLSum	17.99	3.58	14.69
	CLSum+EnSum	18.06	3.61	14.75
	CLSum+MT	18.47	4.16	15.25
Italian	EnSum-MT	20.76	6.88	17.46
	CLSum	21.20	6.15	17.54
	CLSum+EnSum	20.47	6.14	17.39
	CLSum+MT	**21.71**†	**7.04**	**18.11**†
Chinese	EnSum-MT	**27.06**	**8.69**	**23.26**
	CLSum	23.03	5.76	20.27
	CLSum+EnSum	22.62	5.52	19.88
	CLSum+MT	23.28	5.97	20.27

Table 4: Results on the X-SCITLDR-PostEdit (German, Italian and Chinese)

Japanese

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

Table 5: Results on the X-SCITLDR-Human (Japanese)

- EnSum-MT (Summarize-then-translate) is better for German (DE) and Chinese (ZH)
- Doesn't work well if summaries have different style from original English dataset

Benchmark: Results and Analysis for RQ1, RQ2

Lang	Model	R1 (avg)	R2 (avg)	RL (avg)
German	EnSum-MT	19.29	5.46	16.02
	CLSum	17.99	3.58	14.69
	CLSum+EnSum	18.06	3.61	14.75
	CLSum+MT	18.47	4.16	15.25
Italian	EnSum-MT	20.76	6.88	17.46
	CLSum	21.20	6.15	17.54
	CLSum+EnSum	20.47	6.14	17.39
	CLSum+MT	**21.71**†	**7.04**	**18.11**†
Chinese	EnSum-MT	27.06	8.69	23.26
	CLSum	23.03	5.76	20.27
	CLSum+EnSum	22.62	5.52	19.88
	CLSum+MT	23.28	5.97	20.27

Table 4: Results on the X-SCITLDR-PostEdit

Lang	Model	Rouge-1	Rouge-2	Rouge-L
Japanese	EnSum-MT	24.38	4.42	16.54
	CLSum	30.94	4.66*	20.34
	CLSum+EnSum	32.30	5.66	20.89
	CLSum+MT	**32.30**	5.47	20.85

Table 5: Results on the X-SCITLDR-Human

- EnSum-MT (Summarize-then-translate) is better for German (DE) and Chinese (ZH)
- Doesn't work well if summaries have different style from original English dataset
- Machine translation-based intermediate fine-tuning (CLSum+MT) helps in all languages

Benchmark: Results and Analysis for Q3

Lang	Model	R1 (avg)	R2 (avg)	RL (avg)
German	CLSum	2.67	0.46	2.58
	CLSum+EnSum	3.46	0.70	3.32
	CLSum+MT	14.42	2.04	10.75
Italian	CLSum	4.83	0.97	4.41
	CLSum+EnSum	5.87	1.29	5.35
	CLSum+MT	16.11	3.48	12.38
Chinese	CLSum	0.64	0.06	0.61
	CLSum+EnSum	0.79	0.10	0.76
	CLSum+MT	17.88	3.60	13.95
Japanese	CLSum	2.34	0.59	2.06
	CLSum+EnSum	2.37	0.68	2.17
	CLSum+MT	29.43	4.29	18.27

Table 6: Zero-shot performance

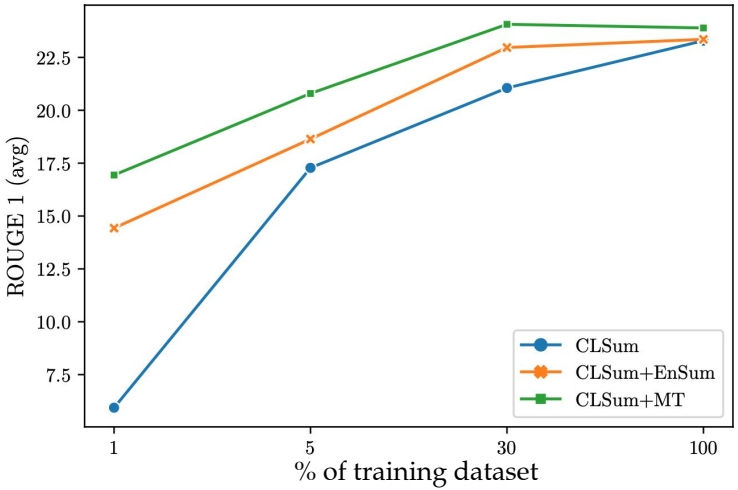


Figure 5: Relationship between R1 and dataset (%)

- Low zero-shot CLSum+EnSum results indicate cross-lingual difficulty of our task
- Even 1% of training data substantially improve
- Intermediate fine-tuning models can learn with less data

Conclusion

- **X-SCITLDR** - English papers to extreme summaries in four languages
- Benchmarking with State-of-the-Art models

huggingface.co/datasets/umanlp/xscitldr

We would like to thank to SIGIR student travel grant.

This work was funded by:

German Research Foundation (DFG) under the VADIS (PO 1900/5-1; EC 477/7-1) and JOIN-T2 (PO 1900/1-2)

Conclusion

- **X-SCITLDR** - English papers to extreme summaries in four languages
- Benchmarking with State-of-the-Art models

huggingface.co/datasets/umanlp/xscitldr

We would like to thank to SIGIR student travel grant.

This work was funded by:

German Research Foundation (DFG) under the VADIS (PO 1900/5-1; EC 477/7-1) and JOIN-T2 (PO 1900/1-2)

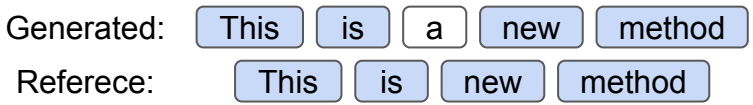


Figure 4: Example of ROUGE-1 (uni-gram based).

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**
	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
	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**
	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

Table 5: Results on the X-SCITLDR-PostEdit portion of our cross-lingual TLDR dataset (ROUGE-1,-2 and -L): English to German, Italian or Chinese TLDR-like summarization using post-edited, automatically-translated summaries of the English data from Cachola et al. [5]. Best results per language and metric are bolded. Statistically significant improvements of the cross-lingual models (CLSum/+EnSum/+MT) with respect to the 'summarize and translate' pipeline (EnSum-MT) are marked with †.

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

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

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

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

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

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