

# CrossSum: Beyond English-Centric Cross-Lingual Abstractive Text Summarization for 1500+ Language Pairs

Abhik Bhattacharjee<sup>1</sup>, Tahmid Hasan<sup>1</sup>, Wasi Uddin Ahmad<sup>2</sup>,  
Yuan-Fang Li<sup>3</sup>, Yong-Bin Kang<sup>4</sup>, Rifat Shahriyar<sup>1</sup>

Bangladesh University of Engineering and Technology (BUET)<sup>1</sup>, University of California,  
Los Angeles<sup>2</sup>, Monash University<sup>3</sup>, Swinburne University of Technology<sup>4</sup>

{tahmidhasan, rifat}@cse.buet.ac.bd, abhik@ra.cse.buet.ac.bd

## Abstract

We present CrossSum, a large-scale cross-lingual abstractive summarization dataset comprising 1.7 million article-summary samples in 1500+ language pairs. We create CrossSum by aligning identical articles written in different languages via cross-lingual retrieval from a multilingual summarization dataset. We propose a multi-stage data sampling algorithm to effectively train a cross-lingual summarization model capable of summarizing an article in any target language. We also propose LaSE, a new metric for automatically evaluating model-generated summaries and showing a strong correlation with ROUGE. Performance on ROUGE and LaSE indicate that pretrained models fine-tuned on CrossSum consistently outperform baseline models, even when the source and target language pairs are linguistically distant. To the best of our knowledge, CrossSum is the largest cross-lingual summarization dataset and the first-ever that does not rely solely on English as the pivot language. We are releasing the dataset, alignment and training scripts, and the models to spur future research on cross-lingual abstractive summarization. The resources can be found at <https://github.com/csebuetnlp/CrossSum>.

## 1 Introduction

Cross-lingual summarization is the task of generating a summary in a target language given a source text in another language. The task is challenging as it combines summarization and translation in one task, both challenging tasks in their own right. Earlier approaches to cross-lingual summarization thus employed pipeline methods like translate-then-summarize (Leuski et al., 2003) or summarize-then-translate (Wan et al., 2010). Not only computationally expensive, having to use multiple models, these approaches also suffer from error-propagation (Zhu et al., 2019) from one model to another, degrading the overall performance.

---

**Input Article:** [...] 新型コロナウイルスに対し、様々な既存の治療法の効果を試す世界的規模の臨床試験の一環として、デキサメタゾンが試された。(Dexamethasone was tested as part of a global clinical trial to test the effectiveness of various existing therapies against the new coronavirus.) [...] その結果、人工呼吸器を必要とする重症患者の致死率が3割下がり。(As a result, the case fatality rate of critically ill patients who require a ventilator is reduced by 30%) [...] ボリス・ジョンソン英首相は「イギリス科学界の素晴らしい成果」を歓迎し。(British Prime Minister Boris Johnson welcomed "the great achievements of the British scientific community".) [...] 「しかもこれは、世界中で手に入る薬だ」("And this is a medicine available all over the world.") [...] きわめて安いステロイド剤だった (but a very cheap steroid that has been used for a long time.)

---

**Summary:** বিজ্ঞানীরা বলছেন ডেক্সামেথাসোন নামে সস্তা ও সহজলভ্য একটি ওষুধ করোনাভাইরাসে গুরুতর অসুস্থ রোগীদের জীবন রক্ষা করতে সাহায্য করবে। (Scientists say a cheap and readily available drug called dexamethasone will help save the lives of critically ill patients with coronavirus.)

---

Figure 1: A sample article-summary pair from CrossSum, the article is written in Japanese and the summary is in Bengali. We translate the texts in English for better understanding. Word and phrases of the article relevant to the summary are color-coded.

The success of sequence-to-sequence (seq2seq) models (Cho et al., 2014; Sutskever et al., 2014) and the advances in Transformer-based models (Vaswani et al., 2017; Rothe et al., 2020) have aided in the emergence of end-to-end methods that can produce cross-lingual summaries with one single model (Zhu et al., 2019). The availability of cross-lingual summarization datasets (Ladhak et al., 2020; Perez-Beltrachini and Lapata, 2021) has also helped this task gain popularity in recent times. However, these datasets cover only a few languages, have few samples for training and evaluation, or use English as the pivot language (i.e., the target language always remains English), thereby limiting the applicability to a great extent.

To democratize cross-lingual summarization beyond high-resource languages, in this work, we introduce **CrossSum**, a large-scale cross-lingual ab-

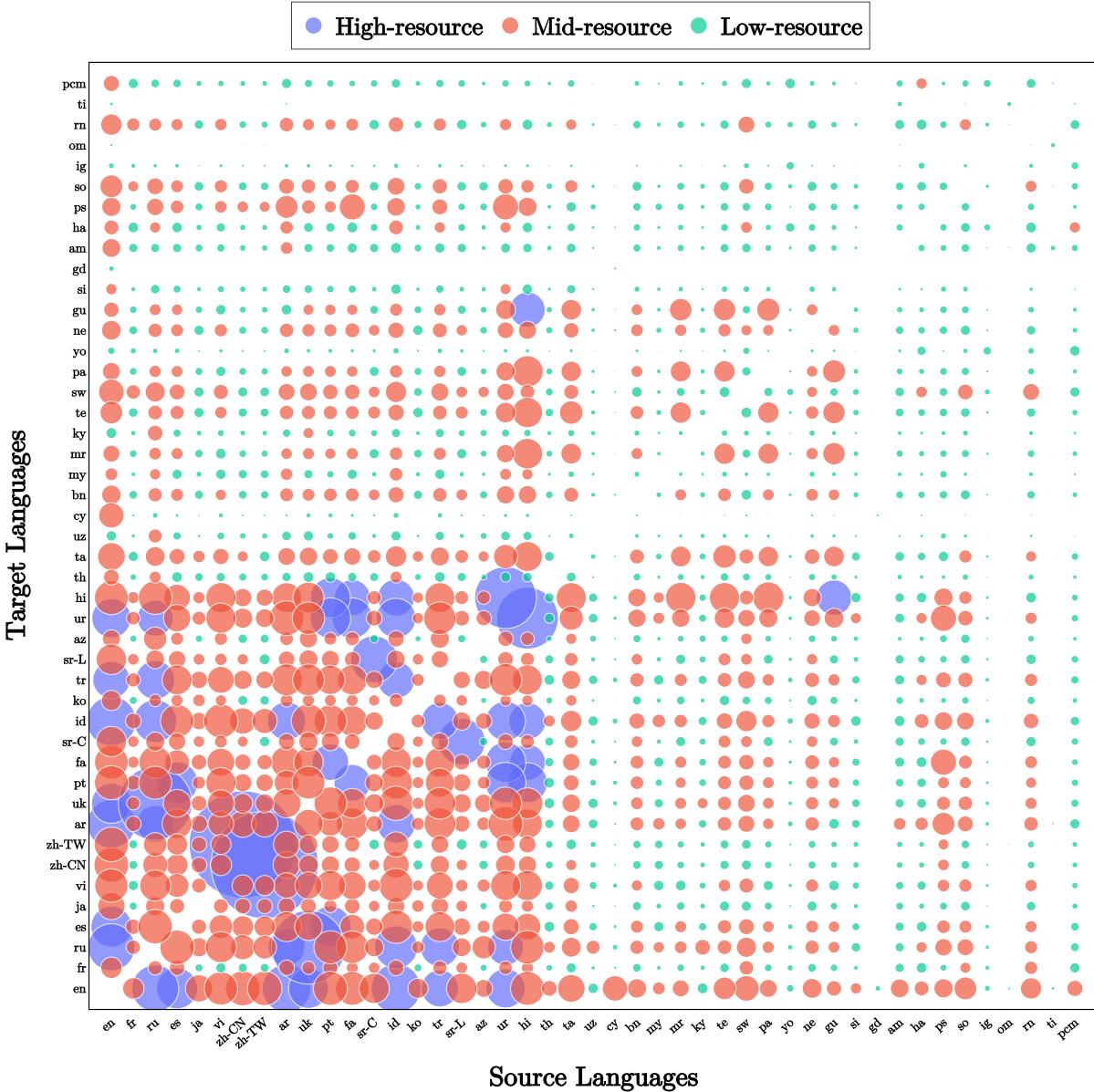


Figure 2: A bubble plot depicting the article-summary statistics of the CrossSum dataset. The radii of the bubbles are proportional to the number of article-summary pairs for the corresponding language pair. Languages in the axes are sorted by the number of their Wikipedia entries to show a sequential contrast from high- to low-resource languages. We consider a language pair as low-resource in CrossSum if the number of samples is below 500, mid-resource for 500 to less than 5000, and high-resource for pairs exceeding 5000.

stractive summarization dataset containing 1.7 million article-summary samples in 1500+ language pairs by aligning identical articles written in different languages via cross-lingual retrieval from the multilingual XL-Sum (Hasan et al., 2021) dataset covering 45 languages. We design a multistage sampling algorithm for successful training of multilingual models that can generate a summary in any target language for a source article in any language (i.e., a many-to-many summarization model). We also propose LaSE, an automatic metric for evaluating cross-lingual summaries when reference

summaries in the target language may not be available (but available in another language), potentially opening new doors to evaluate low-resource language pairs. We also show a strong correlation between ROUGE and LaSE, validating the reliability of LaSE. For the very first time, we perform cross-lingual summarization on a broad and diverse set of languages without relying on English as the standalone pivot language, consistently outperforming several many-to-one and one-to-many models, as well as summarize-then-translate baselines.

To the best of our knowledge, CrossSum is the

first publicly available cross-lingual summarization dataset for a large number of language pairs. We are releasing the dataset, alignment and training scripts, and models hoping that these resources will encourage the community to push the boundaries of cross-lingual abstractive summarization beyond the English and other high-resource languages.

## 2 The CrossSum Dataset

The idea of curating a cross-lingual summarization dataset is to pair the source text of an article A with the summary of another identical article B written in a different language and vice-versa, with the availability of a multilingual dataset where different languages have identical contents. Language-agnostic sentence representations (Artetxe and Schwenk, 2019a; Feng et al., 2022) have achieved state-of-the-art results in cross-lingual text mining (Zweigenbaum et al., 2017; Artetxe and Schwenk, 2019b), and therefore, provide a way to search identical contents across languages.

Two contemporary works have compiled large-scale multilingual summarization datasets, namely XL-Sum (Hasan et al., 2021) (1.35M samples in 45 languages) and MassiveSumm (Varab and Schluter, 2021) (28.8M samples in 92 languages). Though substantially larger than the other, MassiveSumm is not publicly available. Since public availability is crucial for promoting open research, we opted for the other alternative, XL-Sum, which is distributed under a non-commercial research license. XL-Sum has another benefit: all articles are crawled from a single source, BBC News. We observed that BBC publishes similar news contents in different languages and follow similar summarization strategies; hence it would increase the quality and quantity of the mined article-summary pairs.

For simplicity, we perform the similarity search over summaries only. To ensure maximum quality, we set two strong prerequisites for a summary  $S_A$  in language A to be paired with another summary  $S_B$  of language B:

1.  $S_B$  must be the nearest neighbor of  $S_A$  among all summaries in B, and vice-versa.
2. The similarity between  $S_A$  and  $S_B$  must be above the threshold,  $\tau$ .

To measure similarity, we use the inner products of Language-agnostic BERT Sentence Representation (LaBSE) (Feng et al., 2022) (a unit vector for an input text sequence). We set the minimum similarity threshold as the average threshold

$(\tau = 0.7437)$  of all languages that maximized respective  $F_1$  score for LaBSE in the BUCC mining tasks (Zweigenbaum et al., 2017).<sup>1</sup>

**Induced Pairs** We noticed that many summaries, despite being nearest neighbors, were filtered out because of the threshold, although interestingly, both were matched with the exact same summary in a different language. To accommodate these pairs into CrossSum, we introduce ‘*induced pairs*.’ Formally, two summaries  $S_A, S_B$  in languages A, B are induced pairs if they are nearest neighbors of one another in A, B, their similarity score is below  $\tau$ , and both are matched with  $S_C$  in language C as valid pairs  $(S_A, S_B), (S_B, S_C)$  (or through a chain of matched pairs in other languages).

We observed that induced pairs are prevalent if their languages are distant or low-resource. LaBSE uses contrastive learning (Guo et al., 2018; Yang et al., 2019) to rank parallel sentences over non-parallels. Since parallel pairs are mostly found for high-resource and linguistically close languages, we hypothesize that LaBSE fails to assign high similarity to sentences from languages that are not. We thus try to incorporate the induced pairs into CrossSum through a simple graph-based algorithm:

We represent all summaries as vertices in a graph and draw edges between two vertices if the summaries are matched as valid pairs. Then we find the connected components in the graph and draw edges (i.e., induced pairs) between all vertices in a component. Again to ensure quality, before computing the induced pairs, we use the max-flow min-cut theorem (Dantzig and Fulkerson, 1955) considering the similarity scores as edge weights to limit the size of each component to 50 vertices (since ideally a component should have at most 45 vertices, one summary from each language) and set the minimum threshold to  $\tau' = (\tau - 0.10)$ .

We finally assembled the original matched pairs and induced pairs to create the CrossSum dataset. Figure 2 shows the article-summary statistics for all language pairs in CrossSum.

**Implicit Leakage** We initially made the train-dev-test splits respecting the original XL-Sum split: and performed an initial assessment of the dataset using the splits by training a many-to-one model

---

<sup>1</sup>Around 90%  $F_1$  score is achieved using LaBSE in the BUCC tasks, hence it is expected that not all alignment will be correct in CrossSum. Since Hasan et al. (2021) reported summaries around this percentage to be good-quality in XL-Sum, we went ahead with this threshold.

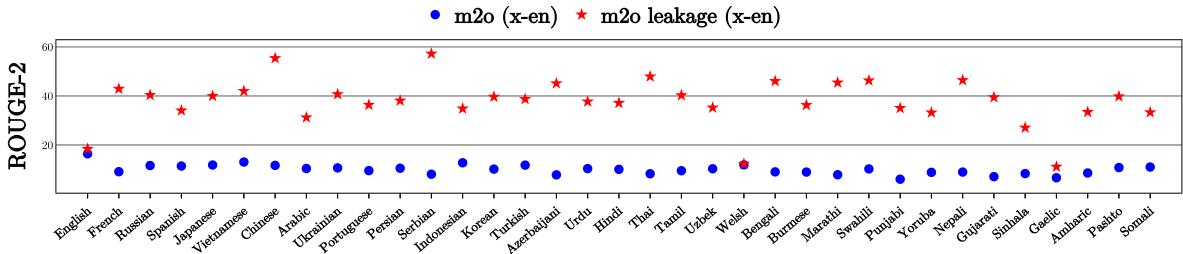


Figure 3: Training on the dataset respecting the original XL-Sum splits causes absurdly high ROUGE scores (marked red) in many-to-one models due to implicit data leakage. Therefore, we split taking the issue into account and consequently, models trained on the new set (marked blue) does not exhibit any unusual spike in ROUGE-2.

(articles written in any source language being summarized into one target language) in a supervised fashion. Upon evaluating the model, we found very high ROUGE-2 scores (up to 60) for many language pairs, even reaching as high as 80 for some (Figure 3). For contrast, Hasan et al. (2021) reported ROUGE-2 in the 10-20 range in the multilingual summarization task.

We inspected the model outputs and found that many summaries were exactly the same as the references. Through closer inspection, we found that all the articles, the summaries of which are exact copies of references, had their identical counterparts in some other language occurring in the training set. During training, the models were successfully able to align the representations of identical articles (albeit written in different languages) and were able to generate the exact same output by memorizing from the training sample. While models should undoubtedly be credited for being able to make these cross-lingual mappings, this is not ideal for benchmarking purposes as this creates unusually high ROUGE scores. We denote this phenomenon as ‘*implicit leakage*’ and make a new dataset split to avoid this. Before proceeding, we deduplicate the XL-Sum<sup>2</sup> dataset using semantic similarity, considering two summaries  $S_A, S'_A$  in language A to be duplicates if their LaBSE representations have similarity above 0.95. We take advantage of the component graph mentioned previously to handle the leakage and assign all article-summary pairs originating from a single component in the training (dev/test) set of CrossSum, creating an even 80%-10%-10% split for all language pairs. Since identical articles no longer appear in the train set of one language and dev/test set of another, the leakage is not observed anymore (Fig-

ure 3). We further validated this by inspecting the model outputs and found no exact copies.

### 3 Training & Evaluation Methodologies

In this section, we discuss the multistage sampling strategy for training cross-lingual text generation models and our proposed metric for evaluating model-generated summaries.

#### 3.1 Multistage Language Sampling

From Figure 2, we can see that CrossSum is heavily imbalanced in terms of samples for different language pairs, and thus training directly without upsampling low-resource languages may result in their degraded performance. Conneau et al. (2020) used a probability smoothing technique for upsampling in multilingual pretraining and sampled all data points of a batch from one language. However, if we did the same for the language pairs in CrossSum, many batches would have duplicate samples since many pairs do not have enough examples, and at the same time, many would not be sampled during training for lack of enough training steps (due to a limitation of computational resources from our side). To address this, we adapt their algorithm to introduce a multistage upsampling method and ensure either the source or the target texts of a batch are sampled from the same language.

Let  $L_1, L_2, \dots, L_n$  be the languages of a cross-lingual source-target dataset. Let  $c_{ij}$  be the number of training samples where the source is from  $L_i$  and target from  $L_j$ . We compute the probabilities of the source languages by

$$p_i = \frac{\sum_{k=1}^n c_{ik}}{\sum_{j=1}^n \sum_{k=1}^n c_{jk}} \forall i \in \{1, 2, \dots, n\}$$

We then use an exponent smoothing factor  $\alpha$  and normalize the probabilities

$$q_i = \frac{p_i^\alpha}{\sum_{j=1}^n p_j^\alpha} \forall i \in \{1, 2, \dots, n\}$$

<sup>2</sup>XL-Sum is deduplicated using lexical overlap methods only. But due to the risk of implicit leakage, which is not lexical, we further perform semantic deduplication.

Given the source language  $L_i$ , we now compute the probabilities of its target languages.

$$p_{j|i} = \frac{c_{ij}}{\sum_{k=1}^n c_{ik}} \forall j \in \{1, 2, \dots, n\}$$

We again smooth  $p_{j|i}$  by a factor of  $\beta$  and obtain the normalized probabilities

$$q_{j|i} = \frac{p_{j|i}^\beta}{\sum_{k=1}^n p_{k|i}^\beta} \forall j \in \{1, 2, \dots, n\}$$

We analogously compute  $p_j$  and  $p_{i|j}$  and, using them, describe the training algorithm with multi-stage sampling in Algorithm 1.

Note that the proposed algorithm can be applied to any cross-lingual seq2seq task where both the source and target languages are imbalanced.

### 3.2 Evaluating Summaries Across Languages

A sufficient number of reference samples are essential for the reliable evaluation of model-generated summaries. However, for many CrossSum language pairs, especially low-resource ones, even the training sets are very small, let alone their test sets. Being able to evaluate using reference summaries written in a different language would allow evaluation in a broad range of languages, especially for which there are inadequate references in the target language. Embedding-based similarity metrics (Zhang et al., 2020; Zhao et al., 2019) have gained popularity in the last few years. We draw inspiration from them and design a similarity metric that does not rely on the lexical overlap between the generated and reference texts. As a result, this new metric can effectively measure similarity across languages in a language-independent manner. We consider three essential factors for our metric:

**1. Meaning Similarity:** The generated summary and the reference summary should convey the same meaning irrespective of their language. Just like our alignment procedure from Section 2, we use LaBSE to compute the meaning similarity between the generated ( $s_{gen}$ ) and reference summary ( $s_{ref}$ ):

$$\text{MS}(s_{gen}, s_{ref}) = \text{emb}(s_{gen}) \cdot \text{emb}(s_{ref})^T,$$

where,  $\text{emb}(s)$  denotes the embedding vector output of LaBSE for input text  $s$ .

**2. Language Confidence:** The metric should identify, with high confidence, that the summary is indeed being generated in the target language. As such, we use the *fastText* language-ID classifier (Joulin et al., 2017) to obtain the language probability distribution of the generated summary and

---

**Algorithm 1:** A pseudocode of the multi-stage sampling algorithm.

---

```

Input:  $D_{ij} \forall i, j \in \{1, 2, \dots, n\}$ : training data with source/target languages  $L_i/L_j$ ;
 $c_{ij} \leftarrow |D_{ij}| \forall i, j \in \{1, 2, \dots, n\}$ ;
 $m$ : number of mini-batches.

1 Compute  $p_i, p_j, p_{j|i}, p_{i|j}$  using  $c_{ij}$ 
2 while (Model Not Covered) do
3    $batch \leftarrow \emptyset$ 
4   Sample  $r \sim Unif(0, 1)$ 
5   if  $r > 0.5$  then
6     Sample  $L_i \sim p_i$ 
7     for  $i \leftarrow 1$  to  $m$  do
8       Sample  $L_j \sim p_{j|i}$ 
9       Create mini-batch  $mb$  from  $D_{ij}$ 
10       $batch \leftarrow batch \cup \{mb\}$ 
11    end
12  end
13  else
14    Sample  $L_j \sim p_j$ 
15    for  $j \leftarrow 1$  to  $m$  do
16      Sample  $L_i \sim p_{i|j}$ 
17      Create mini-batch  $mb$  from  $D_{ij}$ 
18       $batch \leftarrow batch \cup \{mb\}$ 
19    end
20  end
21  Optimize model using  $batch$ 
22 end
```

---

define the Language Confidence (LC) as:

$$\text{LC}(s_{gen}, s_{ref}) = \begin{cases} 1, & \text{if } L_{ref} = \text{argmax } P(L_{gen}) \\ P(L_{gen} = L_{ref}), & \text{otherwise} \end{cases}$$

**3. Length Penalty:** Generated summaries should not be unnecessarily long, and the metric should penalize long summaries. While model-based metrics may indicate how similar a generated summary is to its reference and its language, it is not clear how they can be used to determine its brevity. As such, we adapt the BLEU (Papineni et al., 2002) brevity penalty to measure the length penalty of generated summaries:

$$\text{LP}(s_{gen}, s_{ref}) = \begin{cases} 1, & \text{if } |s_{gen}| \leq |s_{ref}| + c \\ \exp(1 - \frac{|s_{gen}|}{|s_{ref}| + c}), & \text{otherwise} \end{cases}$$

The languages of  $s_{gen}$  and  $s_{ref}$  may not be the same, and identical texts may vary in length across languages. Hence, we used a length offset  $c$

to avoid penalizing generated summaries slightly longer than the references. By examining the standard deviation of mean summary lengths of the languages, we set  $c = 6$ .

We finally define our metric, **Language-agnostic Summary Evaluation (LaSE)** score as follows.

$$\begin{aligned} \text{LaSE}(s_{gen}, s_{ref}) &= \text{MS}(s_{gen}, s_{ref}) \\ &\times \text{LC}(s_{gen}, s_{ref}) \times \text{LP}(s_{gen}, s_{ref}) \end{aligned} \quad (1)$$

## 4 Experiments & Benchmarks

We aim to train one model to generate summaries in any target language for an input article from another language by providing explicit cross-lingual supervision. Fine-tuning pretrained language models (Devlin et al., 2019; Xue et al., 2021) have shown state-of-the-art results on monolingual and multilingual abstractive text summarization (Rothe et al., 2020; Hasan et al., 2021). Many pretrained multilingual generative models are currently available, some prominent ones being mBART (Liu et al., 2020), CRISS (Tran et al., 2020), mT5 (Xue et al., 2021). Though CRISS is pretrained with a cross-lingual objective, which better suits our use case, in contrast to the multilingual objective of mBART and mT5, we choose mT5 for fine-tuning because of its broad coverage of 101 languages with support for 41 languages from CrossSum.

We compare our proposed multistage many-to-many (m2m) model with the standard unistage m2m model as well as many-to-one (m2o) and one-to-many (o2m) models, standards for cross-lingual summarization. We train four different m2o and o2m models using four highly spoken and typologically diverse pivot languages: English, Hindi, Arabic, and Russian. As another baseline, we use a summarize-then-translate pipeline. First, we fine-tune mT5 on our proposed split of the in-language data to obtain a multilingual summarization model. Then we use the M2M-100 model (Fan et al., 2021) (418M parameters variant) to translate the summaries into the target language.

We fine-tune mT5-base with the multistage algorithm with batch size 256, mini-batch size 32 on CrossSum (together with the in-language samples) with  $\alpha = 0.5, \beta = 0.75$ . The unistage m2m model is sampled with  $\alpha = 0.25$ , and each batch is packed with 8 mini-batches, each sample of which being taken from one language pair. m2o and o2m models are also trained in the same manner. All models are trained for 25k steps on 8 Nvidia Tesla

Target Lang.	ROUGE-2 vs. LaSE-in-lang.	LaSE-in-lang vs. LaSE-out-lang.
	Pearson/Spearman	Pearson/Spearman
English	0.923/0.821	0.931/0.929
Hindi	0.967/1.000	0.940/0.600
Arabic	0.963/1.000	0.924/1.000
Russian	0.477/0.489	0.024/0.257

Table 1: Correlation analysis of ROUGE-2 and LaSE for different target languages.

P100 GPUs for 3 days. We discard a language pair from training if it has fewer than 30 training samples to prevent too many duplicates in a mini-batch. We limit the input to 512 and output to 84 tokens and use language-specific BOS (beginning of sequence) tokens (Wu et al., 2016) for guiding the decoder to generate summaries in the intended target language during inference and use a length penalty of 0.6. We show the evaluation results using ROUGE-2 and LaSE in Figure 4 and 5. Results indicate that the m2m model trained with our proposed algorithm consistently outperforms the unistage sampling model, the m2o and o2m models, and the summarize-then-translate pipeline.

## 5 Analysis & Discussions

**Zero-shot/few-shot cross-lingual transfer:** Experiments are done in Section 4 in fully supervised fashion. However, for many low-resource language pairs, samples are not available. Hence, it is attractive to be able to perform zero-shot cross-lingual transfer without relying on any labeled examples. To this end, we fine-tune mT5 with the in-language (both source and target are in the same language) samples only in a multilingual fashion and, during inference, vary the target language. Unfortunately, the model fails at generating cross-lingual summaries and performs in-language summarization instead. We also fine-tune m2o models in a zero-shot setting (with only the in-language samples of the target language). Here, the model can generate non-trivial summaries but still lags behind fully supervised models (results in the Appendix). We do not perform any few-shot experiments and leave them as potential future directions.

**How reliable is LaSE?** To validate the reliability of LaSE, we show its correlation with ROUGE-2. To further posit that LaSE is language-agnostic and can be effectively evaluated with references in a different language from the target, we swap the

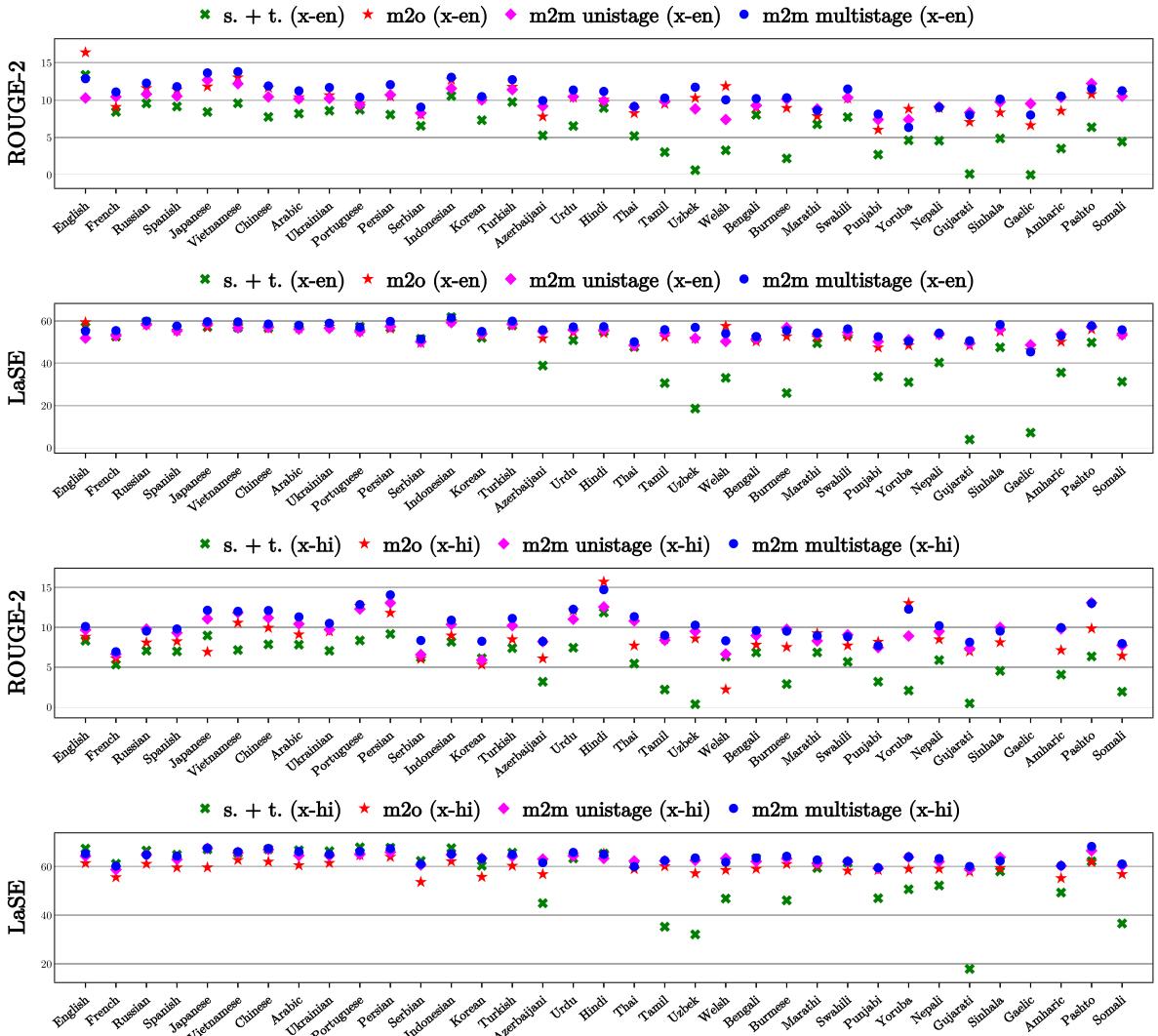


Figure 4: ROUGE-2 and LaSE scores for English and Hindi as target pivots as the source languages vary. Scores indicate that our many-to-many (m2m) model with multistage sampling significantly outperforms the one-to-many models, summarize-then-translate and unistage m2m baselines models on most languages. The comparisons with other pivots are shown in the appendix due to space restrictions.

reference texts with the references in the language of the source text and show the correlation between the two variants of LaSE. We present the Pearson and Spearman correlation coefficients in Table 1. Since we were concerned that data scarcity would question the reliability of evaluation, we only take those language pairs into account that have at least 500 test samples. Results show that there is a high correlation between ROUGE-2 and LaSE for English, Hindi, and Arabic, and moderate for Russian. On the other hand, we find a strong correlation even when the references are swapped for the three above-mentioned languages. However, for Russian, we observed little to no correlation. We wish to investigate this discrepancy in the future and find ways to mitigate this.

## 6 Related Works

Pipeline-based methods were popular at the beginning stages of cross-lingual summarization research (Leuski et al., 2003; Orasan and Chioorean, 2008; Wan et al., 2010), breaking it into two sequential summarization and translation tasks. The lack of large datasets was a major obstruction towards end-to-end methods. End-to-end methods that performed cross-lingual summarization with a single model gained popularity with the emergence of neural models. Using a synthetic dataset, Zhu et al. (2019); Cao et al. (2020) performed cross-lingual summarization with a dual Transformer architecture in a multitask framework, while Bai et al. (2021) propose a single encoder-decoder for better transfer across tasks. Until recently, cross-lingual

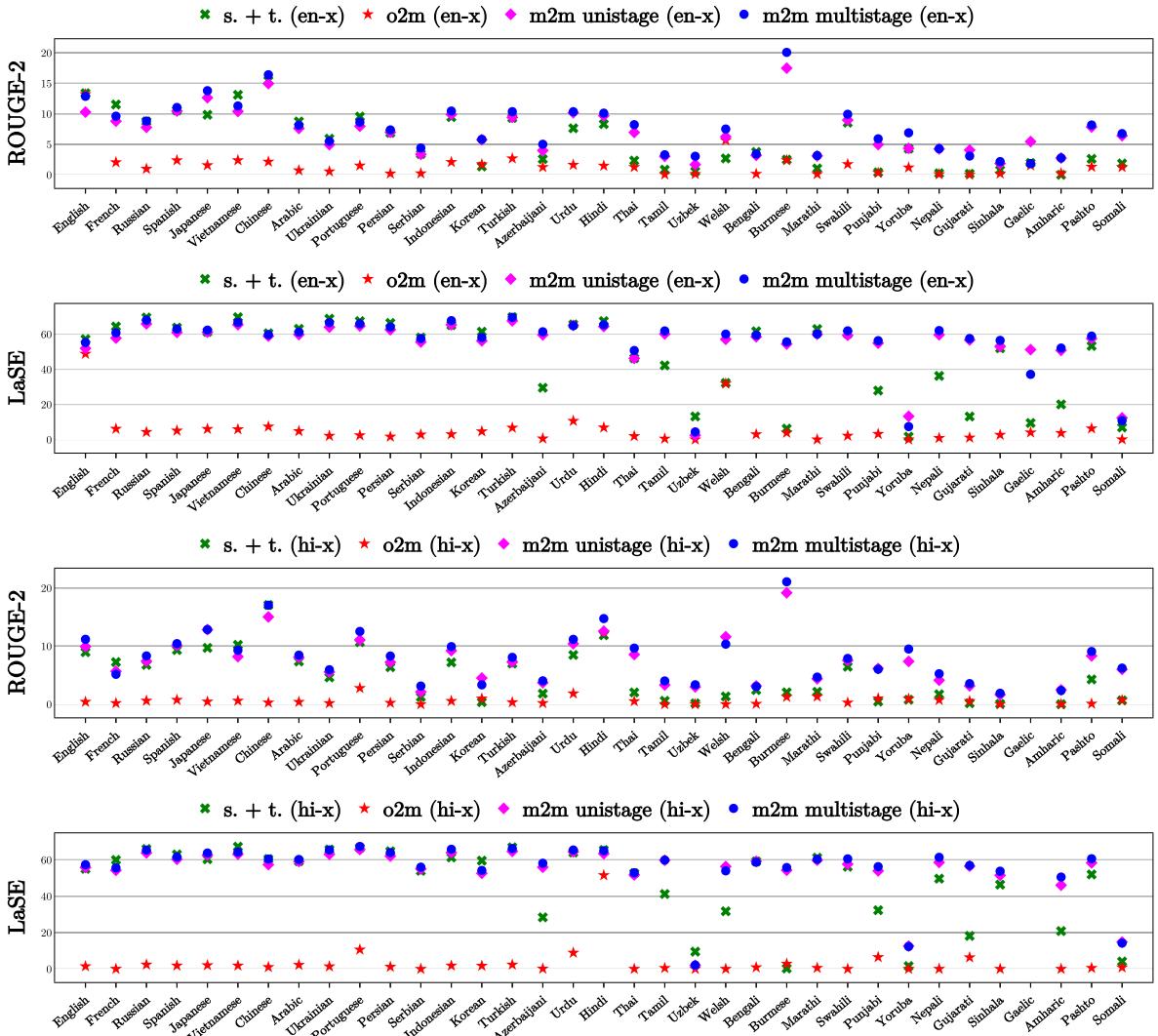


Figure 5: ROUGE-2 and LaSE scores for English and Hindi as source pivots as the target languages vary. Scores indicate that our many-to-many (m2m) model with multistage sampling significantly outperforms the one-to-many models, summarize-then-translate and unistage m2m baselines models on most languages. The comparisons with other pivots are shown in the appendix due to space restrictions.

summarization was limited to English-Chinese pair only due to the lack of benchmark datasets. To promote the task beyond them, Ladhak et al. (2020) introduced Wikilingua, a large-scale many-to-one dataset with English as the pivot language, while Perez-Beltrachini and Lapata (2021) introduced XWikis, containing 4 European languages in 12 many-to-many directions.

## 7 Conclusion & Future Works

In this paper, we present CrossSum, a large-scale, non-English-centric cross-lingual abstractive summarization dataset containing 1.7 million samples across 1500+ language pairs. CrossSum provides the first publicly available cross-lingual summarization dataset and benchmarks for many of these

pairs. We also make the alignment scripts available for the researchers, which will help produce better alignments. Furthermore, we introduced a new multistage sampling algorithm that can be generalized to any cross-lingual generation task and a new language-agnostic metric for evaluating cross-lingual summaries when references in the target languages may not be available. Additionally, we demonstrate that training one multilingual model can help better cross-lingual summarization than baselines. Moreover, CrossSum can also be helpful in zero-shot cross-lingual settings.

In the future, we will investigate the use of our dataset for other summarization tasks, e.g., multi-document (Fabbri et al., 2019) and multi-modal summarization (Zhu et al., 2018).

## Acknowledgements

This work was performed using the OzSTAR national facility at the Swinburne University of Technology. The OzSTAR program receives funding in part from the Astronomy National Collaborative Research Infrastructure Strategy (NCRIS) allocation provided by the Australian Government.

## References

- Mikel Artetxe and Holger Schwenk. 2019a. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy. Association for Computational Linguistics.
- Mikel Artetxe and Holger Schwenk. 2019b. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.
- Yu Bai, Yang Gao, and Heyan Huang. 2021. Cross-lingual abstractive summarization with limited parallel resources. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6910–6924, Online. Association for Computational Linguistics.
- Yue Cao, Hui Liu, and Xiaojun Wan. 2020. Jointly learning to align and summarize for neural cross-lingual summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6220–6231, Online. Association for Computational Linguistics.
- Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- George Bernard Dantzig and Delbert Ray Fulkerson. 1955. On the max flow min cut theorem of networks. Technical report, RAND CORP SANTA MONICA CA.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
- Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.
- Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.
- Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.
- Mandy Guo, Qinlan Shen, Yinfei Yang, Heming Ge, Daniel Cer, Gustavo Hernandez Abrego, Keith Stevens, Noah Constant, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Effective parallel corpus mining using bilingual sentence embeddings. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 165–176, Belgium, Brussels. Association for Computational Linguistics.
- Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.
- Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–

- 4048, Online. Association for Computational Linguistics.
- Anton Leuski, Chin-Yew Lin, Liang Zhou, Ulrich Germann, Franz Josef Och, and Eduard Hovy. 2003. Cross-lingual c\* st\* rd: English access to hindi information. *ACM Transactions on Asian Language Information Processing (TALIP)*, 2(3):245–269.
- Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.
- Constantin Orasan and Oana Andreea Chiorean. 2008. Evaluation of a cross-lingual romanian-english multi-document summariser. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08)*, Marrakech, Morocco. European Language Resources Association (ELRA). <Http://www.lrec-conf.org/proceedings/lrec2008/>.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Laura Perez-Beltrachini and Mirella Lapata. 2021. Models and datasets for cross-lingual summarisation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9408–9423, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280.
- Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS 2014)*, pages 3104–3112, Montreal, Canada.
- Chau Tran, Yuqing Tang, Xian Li, and Jiatao Gu. 2020. Cross-lingual retrieval for iterative self-supervised training. In *Advances in Neural Information Processing Systems*, volume 33, pages 2207–2219. Curran Associates, Inc.
- Daniel Varab and Natalie Schluter. 2021. MassiveSumm: a very large-scale, very multilingual, news summarisation dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10150–10161, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017)*, page 6000–6010, Long Beach, California, USA.
- Xiaojun Wan, Huiying Li, and Jianguo Xiao. 2010. Cross-language document summarization based on machine translation quality prediction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 917–926, Uppsala, Sweden. Association for Computational Linguistics.
- Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv:1609.08144*.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.
- Yinfei Yang, Gustavo Hernandez Abrego, Steve Yuan, Mandy Guo, Qinlan Shen, Daniel Cer, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5370–5378. International Joint Conferences on Artificial Intelligence Organization.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BertScore: Evaluating text generation with bert. In *International Conference on Learning Representations*.
- Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.
- Junnan Zhu, Haoran Li, Tianshang Liu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2018. Msmo: Multimodal summarization with multimodal output. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 4154–4164.

Junnan Zhu, Qian Wang, Yining Wang, Yu Zhou, Jiajun Zhang, Shaonan Wang, and Chengqing Zong. 2019. **NCLS: Neural cross-lingual summarization**. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3054–3064, Hong Kong, China. Association for Computational Linguistics.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2017. Overview of the second bucc shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 60–67.

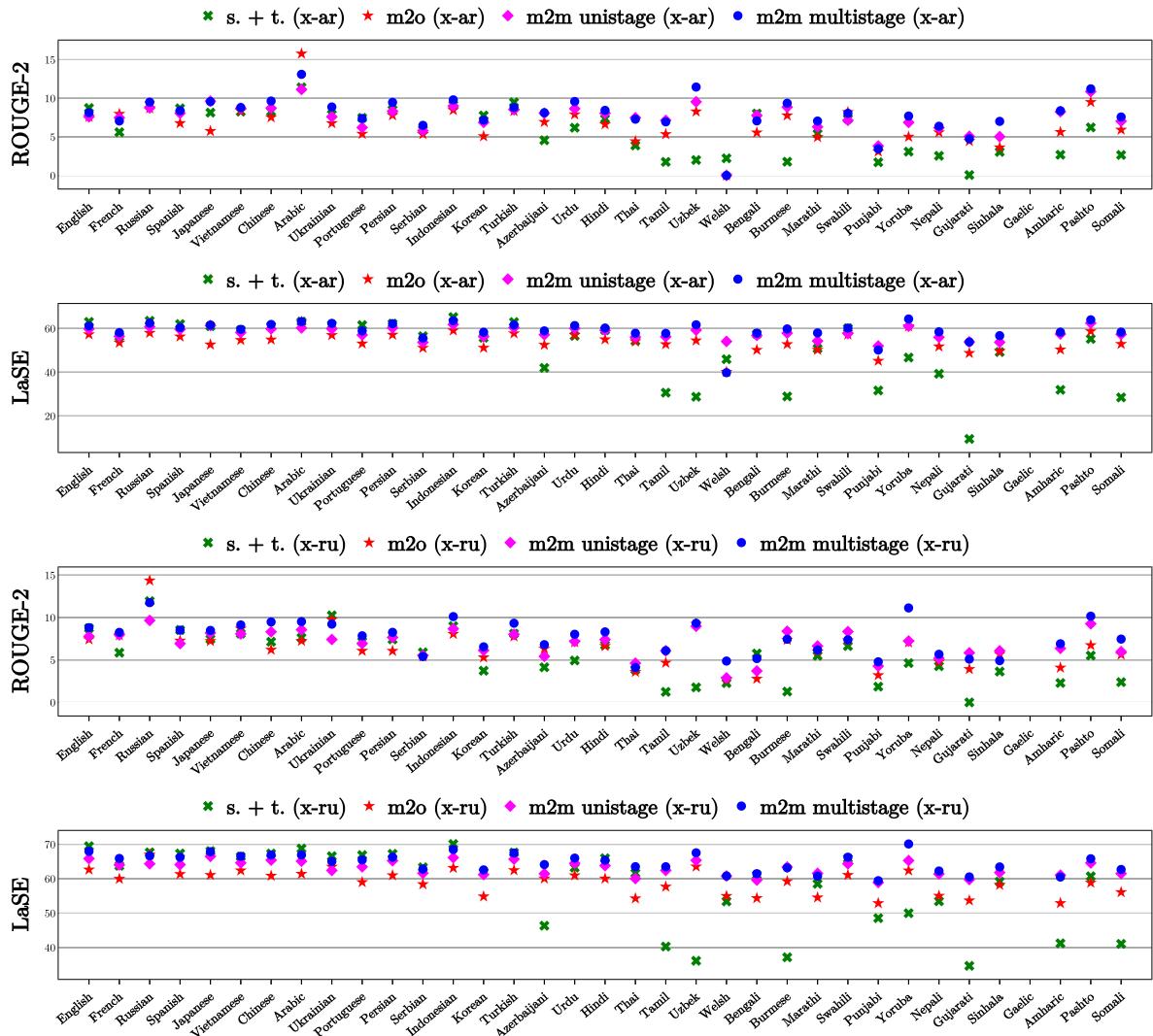


Figure 6: ROUGE-2 and LaSE scores for Arabic and Russian as target pivots as the source languages vary. Scores indicate that our many-to-many (m2m) model with multistage sampling significantly outperforms the one-to-many models, summarize-then-translate and unistage m2m baselines models on most languages. The comparisons with other pivots are shown in the appendix due to space restrictions.

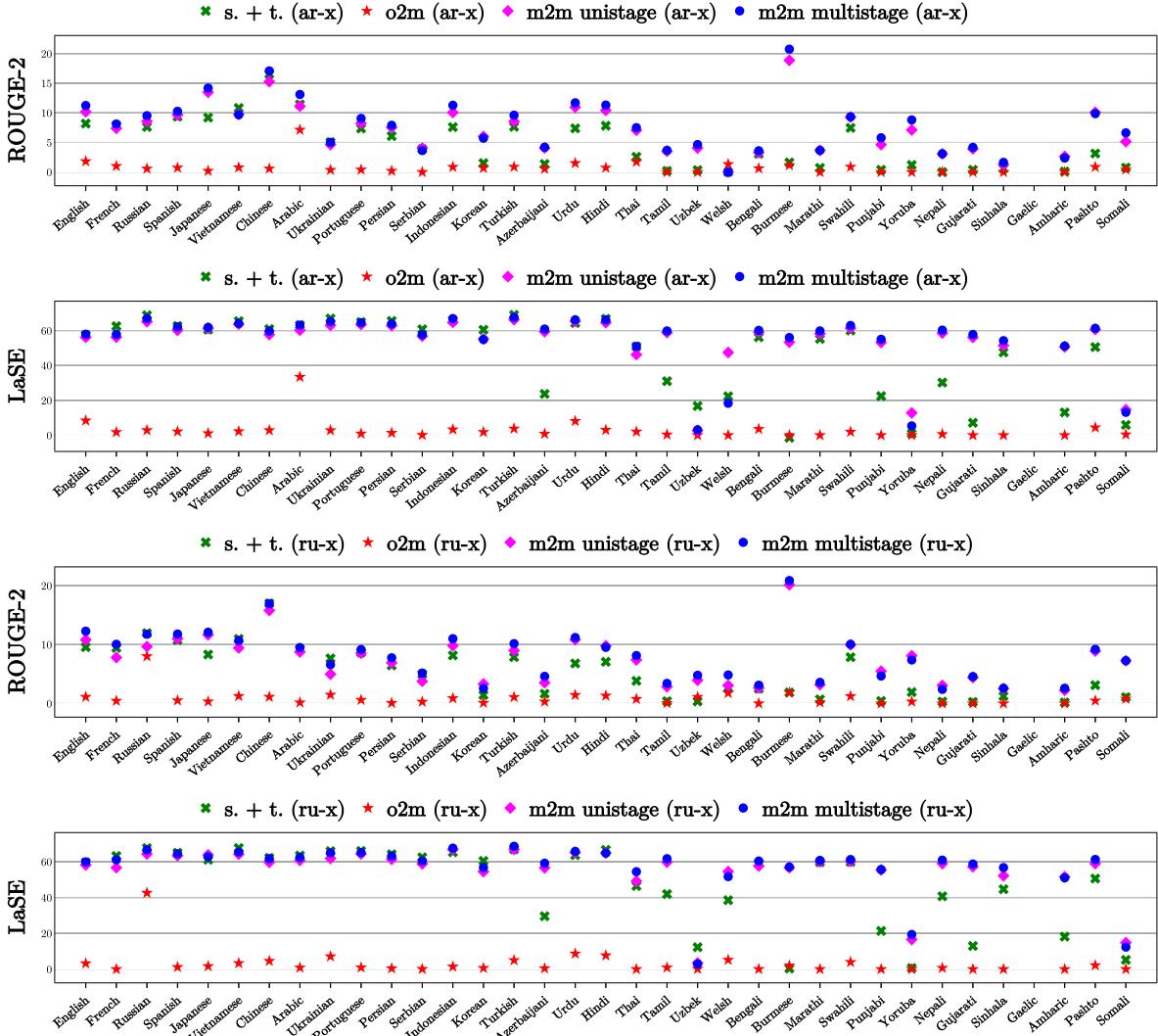


Figure 7: ROUGE-2 and LaSE scores for Arabic and Russian as source pivots as the target languages vary. Scores indicate that our many-to-many (m2m) model with multistage sampling significantly outperforms the one-to-many models, summarize-then-translate and unistage m2m baselines models on most languages. The comparisons with other pivots are shown in the appendix due to space restrictions.

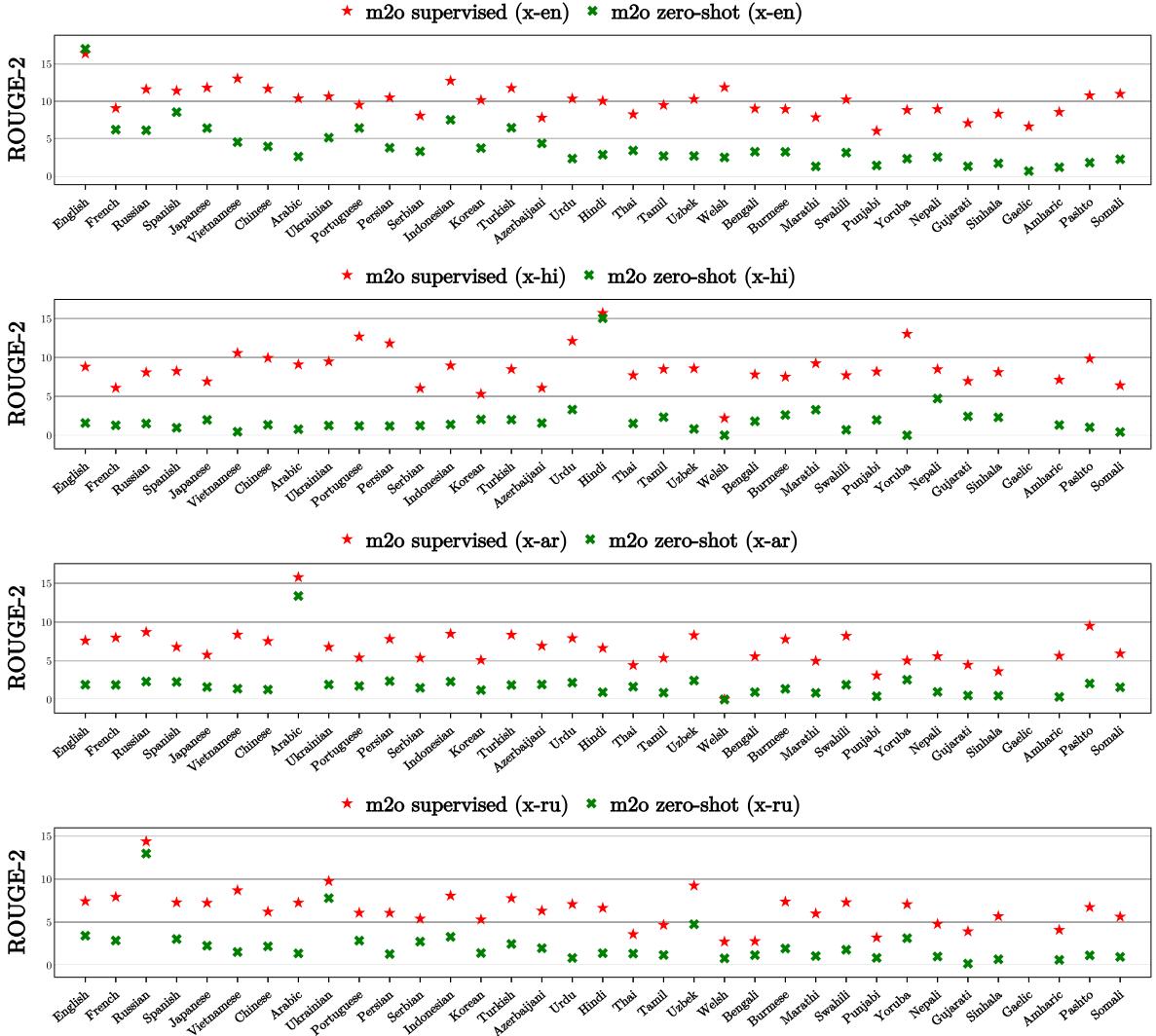


Figure 8: Zero-shot ROUGE-2 scores for the pivot languages as the target languages vary. The zero-shot models are trained with only the in-language samples of the pivot. Though the results are clearly behind the fully supervised model, the model is able to generate non-trivial summaries for many language pairs.

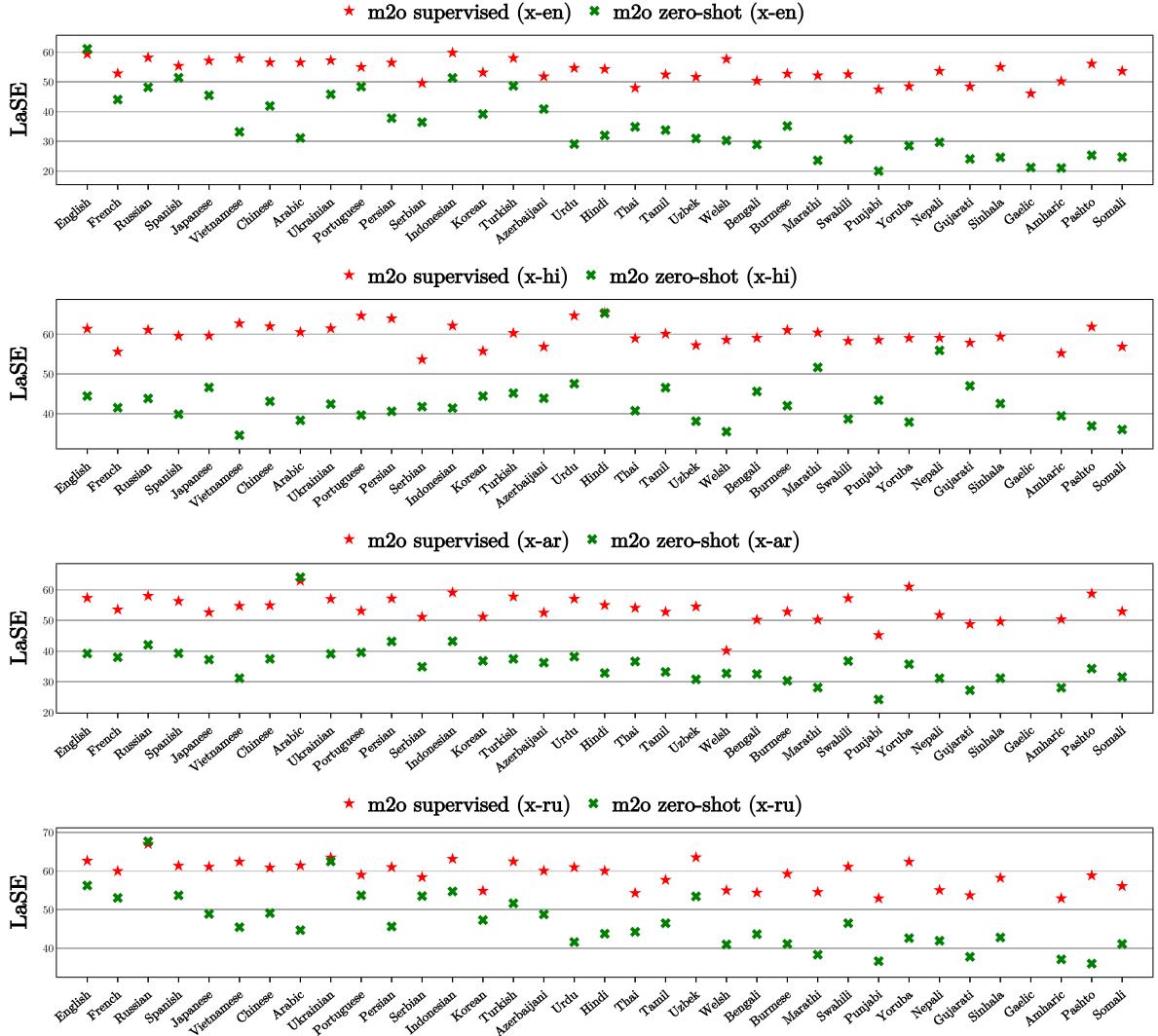


Figure 9: Zero-shot LaSE scores for the pivot languages as the target languages vary. The zero-shot models are trained with only the in-language samples of the pivot. Though the results are clearly behind the fully supervised model, the model is able to generate non-trivial summaries for many language pairs.

Language	am	ar	az	bn	my	zh-CN	zh-TW	en	fr	gu	ha	hi	ig	id	ja	rn	ko	ky	mr	np	on	ps	fa	pm	pt	pa	ru	gd	sr	si	so	es	sw	ta	te	th	ti	tr	uk	ur	uz	vi	cy	yo								
am	—	667	100	272	95	179	167	1426	358	173	221	377	26	494	264	423	244	92	221	301	21	192	431	209	307	189	347	0	357	365	62	309	351	378	390	329	124	131	435	345	409	41	285	1	67							
ar	—	667	—	787	804	652	2968	2843	9653	989	475	747	3665	86	6884	1188	876	707	299	559	181	211	188	155	221	194	242	1	252	817	91	678	190	2238	4	289	283	124	367	704	539	515	245	140	2	149	1383	966	199	725	30	42
az	100	787	—	277	84	371	334	1317	208	192	126	748	28	111	231	188	155	221	194	242	1	252	817	91	678	190	2238	4	289	283	124	367	704	539	515	245	140	2	149	1383	966	199	725	30	42							
bn	272	804	277	—	139	318	284	1549	317	559	231	1396	35	1076	342	298	352	154	586	668	2	300	790	135	764	580	838	0	562	564	151	412	701	471	919	793	245	6	860	688	1382	98	527	37	61							
my	95	652	84	139	—	356	314	685	90	96	74	528	12	761	144	100	112	58	89	152	1	234	426	39	230	86	535	0	115	123	87	79	431	86	185	147	71	4	449	350	591	62	447	4	12							
zh-CN	179	2968	371	318	356	—	4710	4975	348	201	159	1379	38	2851	1017	240	139	240	275	14	559	1111	149	1371	250	2572	2	504	530	166	323	2002	412	511	353	269	11	1511	1691	1651	1761	1858	33	39								
zh-TW	167	2843	334	284	314	4710	—	4884	331	174	121	135	35	2588	953	209	382	131	213	252	16	501	967	141	1271	226	2286	1	453	494	150	302	1873	383	465	335	250	12	1294	1464	1444	158	1663	31	38							
en	14569653	1317	1549	685	4975	4884	—	1889	978	913	4728	144	10040	3040	1878	1673	490	1181	1614	38	1522	4680	1074	4744	1330	9080	128	3760	3809	532	2141	6910	2701	3156	2121	1020	58	5676	6362	6320	450	4574	2655	229								
fr	358	989	208	317	90	348	331	1889	—	242	477	616	106	1018	274	735	264	124	241	323	4	196	602	439	921	247	849	2	555	569	98	502	990	872	425	380	185	10	829	721	76	438	40	159								
id	173	475	192	559	96	201	174	978	242	510	34	710	228	183	268	106	299	561	1	246	522	101	529	221	582	0	331	345	125	261	300	1762	2066	164	5	631	508	1619	80	450	21	54										
gu	221	747	126	231	74	159	913	477	147	—	460	202	901	157	485	135	61	159	239	5	229	487	529	375	157	525	1	258	258	49	391	463	568	299	260	87	9	519	400	526	59	352	30	362								
hi	377	3665	748	1396	528	1379	1213	4728	616	5170	460	—	65	6527	623	489	520	234	383	11357	4	1519	5351	192	6563	4052	4622	1	809	807	449	747	2931	893	371	31762	378	7	3694	335	15666	352	3738	77	79							
ig	26	86	28	35	12	38	35	144	106	34	202	65	—	113	24	107	32	16	51	36	3	11	49	255	61	39	79	0	51	51	13	77	91	51	52	54	18	5	91	83	61	15	65	6	296							
ja	494	6084	1111	1076	761	2851	10040	1018	701	591	627	113	—	1274	994	274	735	264	124	241	323	4	196	602	439	921	247	849	2	555	569	98	502	990	872	425	380	185	10	829	721	76	438	40	159							
m	424	1188	231	342	144	1017	953	3040	274	228	157	623	24	161	510	34	710	228	183	268	106	299	561	1	246	522	101	529	221	582	0	331	345	125	261	300	1762	2066	164	5	631	508	1619	80	450	21	54					
ro	244	707	155	352	112	412	382	1673	264	268	135	520	32	774	654	283	—	99	319	445	1	150	596	130	587	264	649	0	522	543	81	234	163	24	541	452	197	5	680	616	532	54	530	12	45							
ky	92	299	221	154	58	139	131	490	124	106	61	234	16	347	140	106	99	—	107	167	4	102	252	59	251	118	1013	1	206	211	45	279	150	206	174	109	3	346	508	207	270	113	201	12	23							
mr	221	559	194	586	89	213	181	241	2091	159	3831	51	171	545	321	745	231	319	317	630	1	232	608	138	524	179	210	603	1	419	426	129	210	603	168	575	197	473	16	48												
up	301	854	242	668	152	275	1614	323	561	239	1357	34	104	424	369	445	167	630	—	1	303	916	134	706	545	849	2	553	538	164	420	687	513	994	741	217	7	39														
om	21	9	1	2	1	14	16	38	4	1	5	4	3	8	2	18	1	4	—	2	3	11	4	6	8	0	2	3	0	6	7	5	2	1	103	5	10	1	4	2	0	7										
ps	192	2161	252	300	234	559	501	1522	196	246	229	1519	11	1430	266	228	150	102	232	303	2	—	2815	94	594	249	1246	0	235	242	156	304	766	314	441	314	92	8	1049	818	2833	156	657	7	32							
fa	431	4186	817	790	426	1111	967	4680	602	522	487	5351	49	3892	1014	684	598	562	608	916	308	2815	—	186	551	242	432	88	1028	1012	2812	797	364	679	3367	3131	3190	66	74													
pm	209	436	91	135	39	141	1074	439	101	529	192	117	367	152	398	136	510	398	130	59	138	11	94	186	—	227	112	322	0	234	246	28	314	232	162	82	287	280	232	18	170	9	462									
pt	307	259	678	764	230	1371	1271	4744	921	529	375	6563	61	4409	706	526	587	251	524	706	4	594	5512	227	—	579	452	7	1371	1341	231	602	1112	983	1042	820	468	3	3483	442	6759	1863	754	110	97							
pa	189	547	190	580	86	250	226	1330	247	221	1057	4052	39	725	269	206	264	118	179	545	6	249	541	112	579	—	629	0	410	404	128	283	585	357	1726	892	200	10	643	570	1515	73	431	16	44							
sr	0	1	4	0	0	2	1	128	0	1	1	0	1	0	1	0	1	0	0	2	0	0	0	7	0	—	2	3	1	3	1	1	0	6	5	2	1	3	36	2												
sr	357	1109	289	562	115	504	453	3760	555	331	258	809	51	1387	500	443	522	206	419	513	325	1028	234	1371	410	1495	2	—	9041	—	137	382	1260	568	775	699	347	10	1229	1498	1009	112	639	45	79							
si	365	1145	283	564	123	530	494	3809	569	345	258	807	51	1379	571	450	543	211	436	538	3	242	1023	246	1341	404	1460	3	9041	—	127	382	1260	568	775	699	347	10	1229	400	1434	1092	32	68								
si	62	315	124	151	87	166	166	309	149	449	13	470	109	78	1	45	129	164	0	156	276	28	231	74	162	892	892	0	691	697	1384	2278	—	306	11	893	832	1748	107	644	21	61										
so	309	1049	367	412	79	323	302	2141	502	261	391	747	77	1312	387	584	234	145	306	6	304	812	219	602	283	1166	1																									