A Robust Abstractive System for Cross-Lingual Summarization

Jessica Ouyang and **Boya Song** and **Kathleen McKeown**
Deptartment of Computer Science
Columbia University
New York, NY 10027
{ouyangj, kathy}@cs.columbia.edu, bs3065@columbia.edu

Abstract

We present a robust neural abstractive summarization system for cross-lingual summarization. We construct summarization corpora for documents automatically translated from three low-resource languages, Somali, Swahili, and Tagalog, using machine translation and the New York Times summarization corpus. We train three language-specific abstractive summarizers and evaluate on documents originally written in the source languages, as well as on a fourth, unseen language: Arabic. Our systems achieve significantly higher fluency than a standard copy-attention summarizer on automatically translated input documents, as well as comparable content selection.

1 Introduction

Cross-lingual summarization is a little-explored task combining the difficulties of automatic summarization with those of machine translation. The goal is to summarize in one language a document available only in another language. Wan et al. (2010) describe two approaches: summarize then translate, and translate then summarize. They argue that summarize-then-translate is preferable to avoid both the computational expense of translating more sentences and sentence extraction errors caused by incorrect translations.

However, summarize-then-translate can only be used when the source language is high-resource (Wan et al. used English as the source, for example); if the source language is one of the thousands of low-resource languages in the world, there are no summarization corpora available. Language-independent techniques, such as TextRank (Mihalcea), might be used, but there may be serious difficulties in their application, such as morphologically rich languages that render token-based similarity measures useless. In such a case, translate-then-summarize is the only possible approach.

We address this scenario through the development of a neural abstractive summarization system that fluently summarizes potentially disfluent, automatically-translated documents by generating short, simple phrases to replace awkward input phrases resulting from difficult to translate source documents. Our novel combination of existing building block systems results in a summarization solution that can be easily applied to new low-resource languages. We use machine translation on the New York Times annotated corpus of document/summary pairs to create summarization corpora for documents automatically translated from three low-resource languages, Somali, Swahili, and Tagalog. We use these corpora to train cross-lingual summarizers for these source languages, with English as the target. We also evaluate our systems on a fourth source language, Arabic. Our experiments show that our abstractive summarizers produce more fluent English summaries from automatically-translated documents, and that this improvement generalizes across source languages.

Our main contributions are as follows:

- We create summarization corpora for automatically translated Somali, Swahili, and Tagalog documents: noisy English input documents paired with clean English reference summaries.
- We present a method for producing cross-lingual summarization systems for low resource languages where no summarization corpora currently exist, providing a potential summarization solution for thousands of such languages.
- Our novel approach of training on noisy input with clean references outperforms a standard copy-attention abstractive summarizer on real-world Somali, Swahili, and Tagalog documents.
- Our evaluation on Arabic documents demonstrates that our robust abstractive summarizers

Proceedings of NAACL-HLT 2019, pages 2025–2031
Minneapolis, Minnesota, June 2 - June 7, 2019.

in the editor: why did president clinton continue to praise a program on welfare-to-work that failed in half of those assigned? in his comments, he praised the consultation of the community of kansas city, but was advised by gary j. stangler, director of the department of social service of missouri, which half of the participants failed. where are these people helping each other when the government cut them? back to the pantry of food. bad news, mr. president. the charity of the community will not help everyone who will come to us for help. glenn classic valley park, mo.

Figure 1: A synthetic noisy English article from the Tagalog NYT training set.

generalize to unseen languages.

2 Related Work

Cross-Lingual Summarization. Orăsan and Chiorean (2008) extractively summarized Romanian news articles and automatically translated the summaries into English. Their experiments showed that the poor quality of the translations turned reasonable Romanian summaries into barely legible English ones.

The most extensively investigated source-target language pair is English-to-Chinese. Wan et al. (2010) used a predicted translation quality score as a feature in extracting sentences for their summaries. Wan (2011) translated the English sentences into Chinese and represented sentences in the extraction stage by both the original English and the Chinese translation. Yao et al. (2015) scored aligned phrases from the original English documents and the Chinese translations to perform sentence extraction and compression based on both salience and translation quality. Zhang et al. (2016) parsed the original English documents into predicate-argument structures that were aligned with their Chinese translations and generated the summary from these structures. Finally, Wan et al. (2018) experimented with extracting and ranking multiple candidate summaries.

Abstractive Summarization. Rush et al. (2015) presented the first neural abstractive summarization model, a convolutional neural network encoder and feed-forward network decoder with attention, which learned to generate news headlines from the lead sentences of their articles; Chopra et al. (2016) extended their work using a recurrent network for the decoder. Nallapati et al. (2016) improved on the RNN encoder-decoder with attention model by adding linguistically-motivated part of speech and named entity type embeddings, as well as a pointer-network (Vinyals et al., 2015) to allow copying of rare or out-of-vocabulary words from the input document. In this work, we use See et al.'s (2017) definition of the pointer-generator network, which adds a coverage vector and coverage penalty to prevent repetition in generated words.

The New York Times annotated corpus (Sandhaus, 2008) was first used for neural abstractive summarization by Paulus et al. (2018), who used attention over the decoder's previous predictions to both prevent repetition and to allow for coherent longer summaries. Celikyilmaz et al. (2018) also used the New York Times corpus, training multiple, collaborating encoders to encode long documents one paragraph at a time.

3 Data

We use the New York Times (hereafter *NYT*) summarization corpus (Sandhaus, 2008), consisting of 650k articles and their human-written abstractive summaries. We follow the train/test/validation split and preprocessing steps used by Paulus et al. (2018), with one exception: we do not anonymize named entities. We first translate 112k articles from the NYT corpus into each of our three low-resource languages, Somali, Swahili, and Tagalog, using neural machine translation. Of the 112k articles, 100k are taken from the training set, 6k from validation, and 6k from test. We then translate the articles back into noisy English, again using neural machine translation. Figure 1 shows an example noisy English article.

We pair each noisy English article with the clean English reference summary corresponding to the clean English article that generated it. Thus our abstractive summarization model learns to take a "bad" English input document with translation errors and disfluencies and produce a "good" English summary. For simplicity, we refer to the corpus created by translating into Somali and back as the *Somali NYT corpus*, and similarly with Swahili and Tagalog, but all three corpora are in (noisy) English, not Somali, Swahili, or Tagalog.

4 Models

4.1 Machine Translation

We use neural machine translation systems built on the Marian framework (Junczys-Dowmunt

	BLEU	
Language	from English	to English
Somali	21.8	29.4
Swahili	44.5	37.8
Tagalog	37.2	36.2

Table 1: Neural machine translation performance.

Model	ROUGE-1	ROUGE-2	ROUGE-L
NYT-base	**48.26**	29.30	36.81
Paulus	47.03	30.51	**43.27**
Celikyilmaz	48.08	**31.19**	42.33

Table 2: Baseline summarizer performance.

et al., 2018) to translate the NYT corpus into Somali, Swahili, and Tagalog, and back to English. The systems were developed at the University of Edinburgh and were trained on a mix of clean, human-curated parallel data (about 23k sentences for Somali and Swahili and 51k for Tagalog); noisy, web-crawled parallel data (Somali only, about 354k sentences); and synthetic, backtranslated parallel data created from monolingual sources including news articles, the Common Crawl, and Wikipedia (250-600k sentences). Table 1 shows the performance of the machine translation systems for each of the three languages on held-out test sets of 500 sentences taken from the clean, human-curated parallel data.

4.2 Abstractive Summarization

For our abstractive summarizers (hereafter *abstractors*), we implemented See et al.'s (2017) pointer-generator network in PyTorch (Paszke et al., 2017). We pre-train for 12 epochs on the unmodified NYT corpus to obtain a baseline system. Table 2 shows the performance of this baseline on the unmodified NYT test set; our baseline underperforms the more complex systems of Paulus et al. (2018) and Celikyilmaz et al. (2018), but we are more interested in the improvements our fluency-focused approach makes over this baseline than in the baseline's performance compared to state-of-the-art systems. We use each of the three noisy English corpora to train the baseline system for another 8 epochs, producing three language-specific abstractors. We also train a fourth, mixed-language abstractor using 100k articles randomly selected from the Somali, Swahili, and Tagalog training sets, evenly split among the three.

Model	ROUGE-1	ROUGE-2	ROUGE-L
NYT-base	32.94	10.36	22.51
Abs-so*	37.72	15.39	26.56
Abs-mix*	**38.07**	**15.76**	**26.82**

(a) Performance on Somali NYT.

Model	ROUGE-1	ROUGE-2	ROUGE-L
NYT-base	35.28	12.96	25.64
Abs-sw*	39.24	17.01	29.88
Abs-mix*	**39.96**	**17.56**	**30.24**

(b) Performance on Swahili NYT.

Model	ROUGE-1	ROUGE-2	ROUGE-L
NYT-base	37.17	14.67	27.27
Abs-tl*	**40.96**	18.72	31.06
Abs-mix*	40.87	**18.91**	**31.14**

(c) Performance on Tagalog NYT.

Table 3: Abs-so, -sw, and -tl are the Somali, Swahili, and Tagalog systems, respectively. * indicates significant improvement over NYT-base ($p < 1.16 \times 10^{-19}$).

	Perplexity		
Model	Somali NYT	Swahili NYT	Tagalog NYT
NYT-base	4986	4428	4707
Abs-so	**3357**	3429	3528
Abs-sw	3384	**3247**	**3312**
Abs-tl	3501	3476	3457
Abs-mix	3464	3285	3402

Table 4: Language model perplexity of generated summaries on noisy Somali, Swahili, and Tagalog NYT.

5 Evaluation

5.1 Noisy NYT Evaluation.

Table 3 shows the performance of our abstractors on the Somali, Swahili, and Tagalog NYT test sets. Differences among the language-specific systems are not statistically significant, and the more general mixed model achieved the best scores[1]. However, we found that abstractors trained solely on one language and tested on another significantly ($p < 0.05$) underperformed the mixed model, which was trained on all three languages, suggesting that training on some same-language data is still important.

We also trained a bigram language model on the entire set of NYT reference summaries and

[1]These results are shown in Appendix A, along with all combinations of the language-specific models on the three languages.

Document: mange kimambi 'i pray for the parliamentary seat for kinondoni constituency for ticket of ccm. not special seats' kinondoni without drugs is possible i pray for the parliamentary seat for kinondoni constituency on the ticket of ccm. yes, it's not a special seats, khuini kinondoni, what will i do for kinondoni? tension is many i get but we must remember no good that is available easily. kinondoni without drugs is possible. as a friend, fan or patriotism i urge you to grant your contribution to the situation and propert. you can use western union or money to go to mange john kimambi. account of crdb bank is on blog. reduce my profile in my blog understand why i have decided to vie for kinondoni constituency. you will understand more.

NYT-base: mange kimambi, who pray for parliamentary seat for kinondoni constituency for ticket of ccm in 0 , is on blog, and not special seats' kinondoni without drugs.

Abs-mix: mange kimambi, who pray for parliamentary seat for kinondoni constituency for ticket of ccm, comments on his plans to vie for 'kinondoni' without drugs.

Figure 2: An automatically translated Swahili weblog entry and its baseline and mixed abstractor summaries.

Somali Weblogs		
Model	Content	Fluency
NYT-base	1.66	1.62
Abs-so	1.92	1.90
Abs-sw	1.94	1.88
Abs-tl	1.86	1.82
Abs-mix	**2.08**	**2.04**

Swahili Weblogs		
Model	Content	Fluency
NYT-base	1.88	1.76
Abs-so	2.14	1.90
Abs-sw	2.22	**2.08**
Abs-tl	2.18	1.86
Abs-mix	**2.36**	**2.08**

Tagalog Weblogs		
Model	Content	Fluency
NYT-base	1.72	1.76
Abs-so	1.76	1.88
Abs-sw	1.94	1.92
Abs-tl	1.80	2.08
Abs-mix	**2.08**	**2.16**

Table 5: Average human-rated content and fluency scores on Somali, Swahili, and Tagalog weblog entries.

calculated the average perplexity of our abstractors' output as a proxy for fluency (Table 4). We see that Somali is the most difficult overall, but all three language-specific systems and the mixed model produce more fluent English across source languages than does the base model.

5.2 Weblog Evaluation.

We perform a human evaluation on 20 Somali, 20 Swahili, and 20 Tagalog weblog entries that we automatically translate into English using the same neural machine translation systems we used to create our noisy NYT corpora. Unlike our NYT data, which we translated from English into the low-resource languages, these weblogs are real-world Somali, Swahili, and Tagalog documents – this evaluation demonstrates the performance of our system in a real use-case. Figure 2 shows a Swahili weblog entry and its summaries[2]. This example shows the advantage of our approach: unlike a machine translation system, which must translate every part of its input, our abstractor is able to delete most of the long, rambling, and disfluent blog entry, instead summing it up fluently with the generated phrase "comments on his plans" and the repurposed phrase "to vie for".

We use five human judges, all native English speakers and none of whom are the authors. The judges were shown a translated document and a summary and asked to rate the content and fluency of the summary on a scale of 1–3 (Table 5). Our human judges rated our abstractors higher in both fluency and content, and we see again that while the language-specific systems are more fluent on their own languages than are the language-specific systems for the other languages, the mixed model still performs the best. We also see that, while our improvement in content is more modest, our improvement in fluency – the goal of this work – is significant. The judges achieved substantial agreement (Fleiss's $\kappa = 0.72$).

5.3 DUC 2004 Arabic Evaluation.

Finally, we evaluate our system on a new language: Arabic. We use the DUC 2004 Task 3 test set, which consists of real-world Arabic news articles translated into English, each paired with four human-written summaries.

Table 6 shows the performance of our abstractors on the Arabic data, demonstrating their ability to generalize and improve the fluency of input documents automatically translated from a previously unseen language, yielding a significant improvement in ROUGE. Compared to the 28 DUC 2004 systems, our performance would have ranked 1st on summarizing the machine-translated documents; despite our use of these lower-quality, *au-*

[2] All four abstractors produced very similar summaries.

Document: washington 10-23 (afp) was signed by benjamin netanyahu and yasser arafat on friday at the white house agreed on the israeli military withdrawal from the west bank in return for palestinian additional security guarantees.
NYT-base: washington 10-23, signed by benjamin netanyahu and yasser arafat, agrees on israeli military withdrawal from west bank in return for palestinian additional security guarantees.
Abs-mix: benjamin netanyahu and yasser arafat agree on israeli military withdrawal from west bank in return for palestinian security guarantees.

Figure 3: An Arabic article, automatically translated into English, and its baseline and mixed model summaries.

Model	ROUGE-1	ROUGE-2	ROUGE-L
NYT-base	26.56	5.86	15.76
Abs-so*	28.64	6.66	19.62
Abs-sw* †	28.08	6.39	18.36
Abs-tl* †	**29.43**	**7.02**	**19.89**
Abs-mix*	28.79	6.74	19.79

Table 6: DUC 2004 with ISI translations. * indicates significant improvement over NYT-base ($p < 2.09 \times 10^{-6}$); † indicates significant difference between systems ($p < 0.05$).

tomatically-translated documents, we performed extremely well even in comparison with the DUC 2004 systems on high-quality, *human*-translated documents: we would have ranked 1st, 4th, and 5th on ROUGE-1, -2, and -L, respectively. Figure 3 compares the baseline system and our abstractors on the Arabic data[3].

6 Discussion

We find that the NYT-base model tends to copy heavily from the beginning of its input documents. Since it was trained entirely on clean English news articles, it is understandable that it tries to copy the *lead* sentence, but in both examples, it copies errors: the confusing run-on sentence "not special seats' kinondoni without drugs is possible" (shown in yellow in Figure 2) and the phrase "signed by" (shown in green in Figure 3), whose subject is missing. In contrast, our abstractors are able to correctly identify the important information in the input documents and produce fluent summaries presenting this information. In Figure 3, Abs-mix deletes the unnecessary "washington 10-23" and produces the verb "agree" in the plural form, agreeing with its plural subject. More dramatically, in Figure 2, Abs-mix identifies "kinondoni without drugs" as Mange Kimambi's campaign platform and succinctly summarizes this using both the purely generated phrase "comments on his plans" and the repurposed – but still fluent and correct – "to vie for."

[3]All four abstractor summaries were identical.

The main limitation of our approach is that it assumes the existence of a machine translation system for the source language. Although our abstractors are able to handle errorful, disfluent translations, for extremely low-resource languages, there may be no translations of any kind available; in such a case, another approach, such as cross-lingual word embeddings, is necessary.

7 Conclusion

We have presented a robust abstractive summarization system for the task of cross-lingual summarization, taking advantage of an abstractive system's ability to delete difficult to translate phrases and generate new text to use instead. Our straightforward method allows us to produce summarization systems for low resource languages where no summarization corpora are currently available, providing a potential summarization solution for thousands of such languages. Our experiments demonstrate that, by using our novel approach of training on noisy English documents and clean English reference summaries, the model learns to produce fluent summaries from disfluent inputs. Further, we have shown that, while training a system for a specific source language gives strong performance, the abstractive fluency of these systems generalize to other source languages.

Acknowledgements

This research is based upon work supported in part by the National Science Foundation (NSF), under Grant No. IIS-1422863, and the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via contract # FA8650-17-C-9117. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of NSF, ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Asli Celikyilmaz, Antoine Bosselut, Xiaodong He, and Yejin Choi. 2018. Deep communicating agents for abstractive summarization. In *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Sumit Chopra, Michael Auli, and Alexander M Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*.

Rada Mihalcea. Language independent extractive summarization. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics*.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Ça glar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*.

Constantin Orăsan and Oana Andreea Chiorean. 2008. Evaluation of a cross-lingual romanian-english multi-document summariser. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *Proceedings of the 6th International Conference on Learning Representations*.

Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Evan Sandhaus. 2008. The New York Times annotated corpus. *Linguistic Data Consortium*.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Advances in Neural Information Processing Systems*.

Xiaojun Wan. 2011. Using bilingual information for cross-language document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

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*.

Xiaojun Wan, Fuli Luo, Xue Sun, Songfang Huang, and Jin-ge Yao. 2018. Cross-language document summarization via extraction and ranking of multiple summaries. *Knowledge and Information Systems*.

Jin-ge Yao, Xiaojun Wan, and Jianguo Xiao. 2015. Phrase-based compressive cross-language summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Jiajun Zhang, Yu Zhou, and Chengqing Zong. 2016. Abstractive cross-language summarization via translation model enhanced predicate argument structure fusing. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(10).

A Appendix: Language-Specific Abstractor Performance on Noisy NY

Table 7 expands Table 3, showing the performance of each of our four abstractors on the Somali, Swahili, and Tagalog NYT test sets. As discussed in Section 5.1, the differences among the three language-specific abstractors are not statistically significant on any of the three languages. However, the differences between the mixed model and the two language-specific models *not* trained on a given test language are significant ($p < 0.05$). That is, the difference between Abs-mix and Abs-sw and the difference between Abs-mix and Abs-tl are significant on the Somali test set, the difference between Abs-mix and Abs-so and the difference between Abs-mix and Abs-tl are significant on the Swahili test set, and the difference between Abs-mix and Abs-so and the difference between Abs-mix and Abs-sw are significant on the Tagalog test set.

Model	ROUGE-1	ROUGE-2	ROUGE-L
NYT-base	32.94	10.36	22.51
Abs-so*	37.72	15.39	26.56
Abs-sw* †	37.26	14.94	25.92
Abs-tl* †	36.89	14.41	25.53
Abs-mix*	**38.07**	**15.76**	**26.82**

(a) Performance on Somali NYT.

Model	ROUGE-1	ROUGE-2	ROUGE-L
NYT-base	35.28	12.96	25.64
Abs-so* †	38.42	16.34	29.06
Abs-sw*	39.24	17.01	29.88
Abs-tl* †	38.24	16.02	28.79
Abs-mix*	**39.96**	**17.56**	**30.24**

(b) Performance on Swahili NYT.

Model	ROUGE-1	ROUGE-2	ROUGE-L
NYT-base	37.17	14.67	27.26
Abs-so* †	38.97	17.01	29.16
Abs-sw* †	39.14	17.28	29.43
Abs-tl*	**40.96**	18.72	31.06
Abs-mix*	40.87	**18.91**	**31.14**

(c) Performance on Tagalog NYT.

Table 7: Abs-so, -sw, and -tl are the Somali, Swahili, and Tagalog language-specific systems, respectively. * indicates significant improvement over NYT-base ($p < 1.16 \times 10^{-19}$); † indicates significant difference between the mixed model and language-specific abstractors ($p < 0.05$).