Global Voices: Crossing Borders in Automatic News Summarization

Khanh Nguyen[⊙] and **Hal Daumé III**[⊙♡]
University of Maryland, College Park[⊙], Microsoft Research, New York[♡]
{kxnguyen,hal}@umiacs.umd.edu

Abstract

We construct *Global Voices*, a multilingual dataset for evaluating cross-lingual summarization methods. We extract social-network descriptions of Global Voices news articles to cheaply collect evaluation data for into-English and from-English summarization in 15 languages. Especially, for the into-English summarization task, we crowd-source a high-quality evaluation dataset based on guidelines that emphasize accuracy, coverage, and understandability. To ensure the quality of this dataset, we collect human ratings to filter out bad summaries, and conduct a survey on humans, which shows that the remaining summaries are preferred over the social-network summaries. We study the effect of translation quality in cross-lingual summarization, comparing a translate-then-summarize approach with several baselines. Our results highlight the limitations of the ROUGE metric that are overlooked in monolingual summarization.

1 Introduction

Cross-lingual summarization is an important but highly unexplored task. The ability to summarize information written or spoken in any language at a large scale would empower humans with much more knowledge about the diverse world. Despite the fast development of automatic summarization (Allahyari et al., 2017; Dong, 2018; Gambhir and Gupta, 2017), present technology mostly focuses on monolingual summarization. There is currently lacking a standard, high-quality multilingual dataset for evaluating cross-lingual summarization methods. Two main challenges present in constructing such a dataset. First, the cost of crowd-sourcing human-written summaries is high. It generally takes a long time for a human to summarize a document, as they not only have to read and understand information in the article, but also have to make complex decisions in sieving and paraphrasing the information. Second, it is difficult to design summarization guidelines for humans, as the task is generally not well-defined: the selection of what content is "important" in a summary is based on subjective and common-sense rules that vary among individuals and are difficult to be expressed precisely in words.

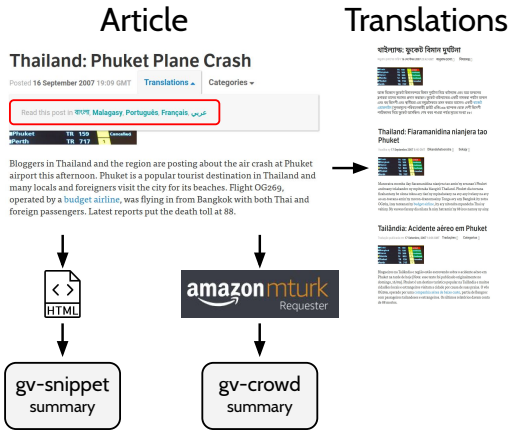

Figure 1: Data construction pipeline. We collect two types of summary: (a) the social network description of the article (`gv-snippet`) and (b) the 50-word summary written by Mechanical Turk workers following our guidelines (`gv-crowd`).

Even in monolingual summarization, there were limited attempts in constructing summarization datasets via crowd-sourcing (Over et al., 2007; Dang and Owczarzak, 2008, 2010). These datasets are mostly used for evaluation due to their small sizes. To construct large-scale training datasets, researchers mine news sources that naturally provide human-written summaries (Hermann et al., 2015; Sandhaus, 2008), or construct artificial summaries from document titles (Rush et al., 2015). Summaries collected in this way may be not best for evaluation because they are generated under unknown guidelines (or there may be no guide-

Proceedings of the 2nd Workshop on New Frontiers in Summarization, pages 90–97
Hong Kong, China, November 4, 2019.

lines at all). Previous work on cross-lingual summarization performs evaluation with human judgments (Orsan and Chiorean, 2008), or with automatic metrics and noisy source articles generated by automatic translation systems (Wan et al., 2010; Ouyang et al., 2019). The former approach is expensive and not reproducible, while the latter is prone to biases induced by translation systems that could be further amplified by summarization systems.

This paper presents *Global Voices*, a high-quality multilingual dataset of summaries of news articles. The dataset can serve as a standard benchmark in both multilingual and cross-lingual summarization. Global Voices[1] is a multilingual website that reports and translates news about unheard voices across the globe. Translation in this website is performed by the Lingua team,[2] consisting of volunteer translators. As of August 2019, Global Voices provides translations of news articles in 51 languages; many articles are translated into multiple languages. Figure 1 illustrates a sample article from Global Voices. We extract the social-network descriptions of the articles to (cheaply) construct `gv-snippet`, an evaluation set for multilingual and cross-lingual news summarization. Nevertheless, these descriptions usually have poor coverage over the original contents because they were written with the intention of drawing user clicks to read more about the articles. Therefore, besides `gv-snippet`, we construct a smaller but higher-quality dataset of human-written English summaries, called `gv-crowd`, based on our guidelines which explicitly emphasize accuracy, coverage and understandability. The Global Voices dataset is summarized in Table 2. It currently supports 15 languages, which span nine language genera (Romance, Barito, Indic, Slavic, Semitic, Greek, Germanic, Japanese, Bantoid) and five language families (Indo-European, Austronesian, Japanese, Niger-Congo, Afro-Asiatic).

2 Dataset Construction

Data Collection and Pre-Processing. Using Scrapy,[3] we crawl and download HTML source codes of 41,939 English articles and their translations. We use `bs4`[4] to extract each article's main content and remove image captions. Next, we use `html2text`[5] to convert the main content's HTML source code to regular text, removing web-page and image URLs. Since an article may content block-quotes written in original languages, we detect language of each paragraph and remove paragraphs that are not in the article's main language. Language detection is conducted by voting decisions of four packages: `langdetect`,[6] `langid`,[7] `polyglot`,[8] `fastText`[9] (Joulin et al., 2016a,b).

Constructing `gv-snippet`. This split includes articles whose English versions contain from 150 to 500 words. For each article, we extract its Open Graph description by extracting the `meta` tag with property `og:description` in the HTML source code, and use the description as the reference summary of the article. These descriptions are short text snippets that serve as captions of the articles when they appear on social networks (e.g. Facebook, Twitter).

Crowd-sourcing `gv-crowd`. We select English

Language	ISO 639-1	gv-snippet	gv-crowd
Number of articles			
English	en	4,573	529
Spanish	es	3,921	487
Malagasy	mg	2,680	374
Bengali	bn	2,253	352
French	fr	2,130	352
Portuguese	pt	798	162
Russian	ru	795	139
Arabic	ar	745	191
Italian	it	718	135
Macedonian	mk	701	138
Greek	el	694	128
German	de	647	204
Japanese	ja	424	75
Swahili	sw	418	84
Dutch	nl	348	87
Other statistics			
Summarized by		GV authors	MTurkers
Summary languages		All versions	English
Summary lengths (words)		-	40-50
Article lengths (words)		150-500	150-350

Table 1: Summary of the Global Voices dataset. The dataset include articles in 15 languages. English versions of all non-English articles are included. The `gv-snippet` split contains social-network summaries of all articles, while the `gv-crowd` split contains crowd-sourced summaries of English articles.

[1] https://globalvoices.org/
[2] https://globalvoices.org/lingua/
[3] https://scrapy.org/
[4] https://www.crummy.com/software/BeautifulSoup/
[5] https://pypi.org/project/html2text/
[6] https://pypi.org/project/langdetect/
[7] https://github.com/saffsd/langid.py
[8] https://pypi.org/project/polyglot/
[9] https://fasttext.cc/docs/en/language-identification.html

articles that contain 150-350 words, and request workers from Mechanical Turk[10] (MT) to summarize them in 40-50 words. Each HIT[11] asks a worker to summarize five articles in 35 minutes. We recruit Turkers in Canada and the U.S.A. with Masters qualification, a HIT approval rate greater than or equal to 97%, and a number of HITs approved greater than or equal to 1,000. On average, collecting a summary costs 1.50 USD (including taxes and extra fees). We inform workers of our evaluation guidelines, which focus on three criteria:

- *Accuracy*: information in a summary should be based on the original article only. It can be paraphrased from but should not disagree with information in the article.
- *Coverage*: a summary should reflect the most important messages/stories in the original article. Each message/story should be captured as detailed as possible, without missing other important messages/stories.
- *Understandability*: a summary must be written in standard, fluent English. Readers must be able to understand the summary without reading the original article. Understanding the summary must not require any additional knowledge beyond knowledge required to understand the article.

In comparison, the DUC-2004 dataset (Over et al., 2007) only provides subtle format suggestions and leaves the summary contents almost entirely to the decisions of the writers:

> "...Imagine that to save time, rather than read through a set of complete documents, you first read a list of very short summaries of those documents and based on these summaries you choose which documents to read in their entirety. Create your very short summaries to be useful in such a scenario. A very short summary could look like a newspaper headline, be a list of important terms or phrases separated by commas, a sentence, etc. It should not contain any formatting, i.e., no indented lists, etc. Feel free to use your own words."
>
> Source: https://duc.nist.gov/duc2004

Our guideline criteria are similar to those of the TAC 2010's guided summarization task (Dang and Owczarzak, 2010) but we do not restrict the summary format using domain-specific templates.

[10] https://www.mturk.com/

[11] a Mechanical Turk task.

Some articles may read disrupted due to removals of images and videos, and may contain non-English texts. To ensure the summaries are based on the English texts only, we advise workers to (a) *not* web-search for the original content and (b) ignore the non-English contents. We also emphasize spelling words correctly and recommend copying difficult-to-spell words from the original articles. In the end, we collect 840 summaries for 738 articles.

Human Evaluation of `gv-crowd`. The summary-collecting task receives mostly positive feedback from workers. The task is widely regarded as "fun", "interesting", and "challenging". However, many workers raised concern about the strict time constraint. To evaluate the quality of the dataset, we launch another MT task in which we ask workers to rate and post-edit the summaries collected in the previous task. Each task HIT requires evaluating ten summaries in 60 minutes. We recruit workers in Canada and the U.S.A. with a HIT approval rate greater than or equal to 97%, and a number of HITs approved greater than or equal to 1,000.

Specifically, we ask workers to provide two types of ratings: *criterion-based* ratings and *overall* ratings. Each worker is instructed to first give a 1-to-5 rating of a summary in each of our three criteria (accuracy, coverage, understandability), and then to give an overall rating of the summary. We define three levels of the overall rating:

- *Bad*: the summary misrepresents the original article. It contains factual errors that disagree with the content of the article. OR it does not cover the most important message/story of the article. OR it is missing other important points that could easily be included without violating the 50-word constraint.
- *Acceptable*: the summary covers the most important message/story of the article. It does *not* contain factual errors. It is missing one or two important points that would be difficult to include in a 50-word summary.
- *Good*: the summary covers the most important message/story of the article. It does *not* contain factual errors. All important points are captured.

In addition, the worker is required to write short reasons (each in 5-25 words) to justify their ratings.

Among 840 summaries collected, 383 (45.60%) were rated as *Good*, 264 (31.43%) *Acceptable*, and 193 (22.98%) *Bad*. We observe that among the three criteria, understandability is easiest to meet while coverage is the most challenging: the mean understandability rating is 4.06 while the mean coverage rating is only 3.47; about 90% of the summaries attain understandability ratings of at least 3. By computing Pearson correlation coefficients, we find that the overall rating most strongly linearly correlates with the coverage rating (0.81) and least with the understandability rating (0.57). Common flaws identified by the human evaluators include: missing important points, factual errors, abstruse and/or verbose writing.

To construct the `gv-crowd` split, we pair each article with its highest-rated summaries[12] and excluded articles that (a) are paired with *Bad* summaries or (b) have a criterion-based rating below 3. We also ask workers to correct spelling and factual errors in the *Bad* summaries, but these post-edited summaries require further evaluation to be included in the dataset in the future. To facilitate summarization evaluation studies, we will release all the summaries accompanied with their ratings, reasons, and post-edit versions.

For a (randomly selected) subset of 50 articles, we collect *three* summaries per article to study the diversity in quality and language usage among human-written summaries of the same documents. We find that the summary quality does not vary greatly: the overall-rating difference between the highest and lowest rated summaries is at most 1 in 74% of these articles. To quantify the diversity of summaries, we calculate the *pairwise* ROUGE scores, using one summary as the reference and another as the predicted

$$\text{ROUGE}_{\text{pair}} = \frac{1}{3 \cdot 50} \sum_{i=1}^{50} \sum_{1 \le j < k \le 3} \text{ROUGE}(s_{i,j}, s_{i,k}) \quad (1)$$

where $s_{i,j}$ and $s_{i,k}$ are distinct summaries of the i-th article. The $\text{ROUGE}_{\text{pair}}$-1,2,L F-1 scores are relatively low (39.44, 12.39, and 32.85, respectively), indicating that the summaries highly vary in vocabulary and sentence structure.

[12]For a pair of summaries, we first compare their overall ratings, then sums of three criterion-based ratings, then the individual accuracy, coverage, understandability ratings (in this specific order).

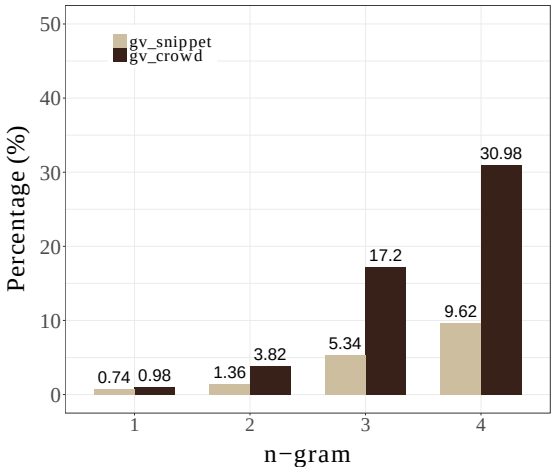

Figure 2: Average fraction of n-grams in the summary that are not seen in the original article.

Human Comparison of `gv-snippet` and `gv-crowd`. To ensure that the `gv-crowd` summaries are of higher quality than the `gv-snippet` summaries, we conduct a survey that asks MT workers to compare the two types of summary. Concretely, each worker reads an article and its `gv-snippet` and `gv-crowd` summaries. We ask the worker to specify which summary (or none) is better in each of the three criteria and is better overall. We remove partial sentences that end with "..." in the `gv-snippet` summaries to ensure that the workers rate the two types of summary mainly based on their contents, not based on any peculiar features. We also randomly shuffle the order of the summaries in a pair so that the workers cannot rely on the order to determine the summary type. Each worker is given 45 minutes to compare five summary pairs. Each summary pair is evaluated by three workers. We recruit workers with similar qualifications to those in the `gv-crowd` evaluation task.

The outcome of this survey is positive. In 22 out of the 30 articles included in the survey (75.9%), at least 2 out of 3 workers prefer the `gv-crowd` summary. Overall, 63 out of 90 workers (70.0%) prefer the `gv-crowd` summaries to its `gv-snippet` counterparts. As expected, coverage is the criterion where the `gv-crowd` summaries show most strength against the `gv-crowd` summaries, with a preference ratio of 83.3% (25/30) compared to 66.7% (20/30) of accuracy or understandability.

We also evaluate these two types of summary in terms of how novel their summaries are compared to the original articles. Figure 2 shows the

Model	Train	Validation
Translation (sentences)		
Spanish-English	4.1M	3K
French-English	5.6M	3K
German-English	151.6K	2K
Arabic-English	174.3K	2K
Summarization (pairs of documents and summaries)		
English	287.2K	13.4K

Table 2: Data used to train and validate translation and summarization models.

average fractions of novel n-grams of each type of summary. Overall, the summaries reuse most words in the articles. The `gv-crowd` summaries contain substantially more novel 3-grams and 4-grams than the `gv-snippet` summaries, partly because each sentence of a `gv-crowd` summary usually includes information from multiple sentences in the original article. On 73% of the articles in the `gv-crowd` split, the `gv-crowd` summary has higher fractions of novel n-grams than the `gv-snippet` counterpart (with $n = 1, 2, 3, 4$).

3 Experiments

We study the task of generating English summaries of non-English news articles. This task can naturally be decomposed into two subtasks: translation and summarization. We follow a *translate-then-summarize* approach where each article is first translated into English using a pre-trained machine translation model, then the translation is summarized using a pre-trained English summarization model. Data for training models in both subtasks are publicly available, allowing solving the joint task in a *zero-shot* manner, in the sense that no parallel pairs of (original document, English summary) are provided during training. On the other hand, a *summarize-then-translate* approach is practically difficult to implement because of the lack of large-scale datasets for training reliable summarization models in non-English languages.

Translation models. Our goal is to study the effect of translation quality in this task. Hence, we employ translation models trained under various amounts of resources. We conduct experiments in four source languages: Spanish (es), French (fr), German (de), and Arabic (ar). Concretely, we train the {es,fr}-en models using the large-scale CommonCrawl and News Commentary datasets, and train the {de,ar}-en models using the low-resource multilingual TED (Duh, 2018) dataset. We apply standard machine translation pre-processing steps, normalizing and tokenizing the data with Moses scripts. We tokenize Arabic texts with the PyArabic tool (Zerrouki, 2010). Our translation models implement the Transformer architecture (Vaswani et al., 2017). The {es,fr}-en models have the same hyperparameters as those of the base Transformer architecture described in Table 3 of Vaswani et al. (2017). The {de,ar}-en models have less parameters, using 4 attention heads and a feed-forward hidden size of 1024. We train the models using the `fairseq-py` toolkit (Ott et al., 2019). Since the models are trained to perform sentence-level translation, we split the source articles into sentences, perform translation, and join the output sentences into articles. The training settings are the same as those of Vaswani et al. (2017) except that: (a) the maximum tokens in a batch is 4,000, (b) the {es,fr}-en models and the {de,ar}-en models are trained for $5 \cdot 10^4$ and $8 \cdot 10^5$ iterations, respectively, and (c) the {de,ar}-en models use a dropout ratio of 0.3. Training with an Nvidia Titan Xp GPU took place in approximately 5 hours for the smaller models and 3.5 days for the larger models.

Summarization models. We employ the state-of-the-art Bi-LSTM bottom-up abstractive summarization model (Gehrmann et al., 2018). We make use of a pre-trained instance of this model provided by OpenNMT-py (Klein et al., 2017) and trained on the CNN/DailyMail dataset (Hermann et al., 2015).

Baselines. We compare the following approaches:

- FIRST50: copies the first 50 words of the English version of the source article.
- PERFECTTRANS: directly summarizes the English version of the source article.
- TRANSTHENSUM: our approach which first translates the source article into English then summarizes the translation.

Evaluation. Translation quality is measured by corpus-level BLEU, treating each article as a data point. Summarization quality is determined by computing ROUGE-1, ROUGE-2, ROUGE-L F-1 scores.

Results. Table 3 presents our results. A qualitative example is illustrated in Figure 3. As expected, translation quality varies among different pairs of languages. The Spanish-English model

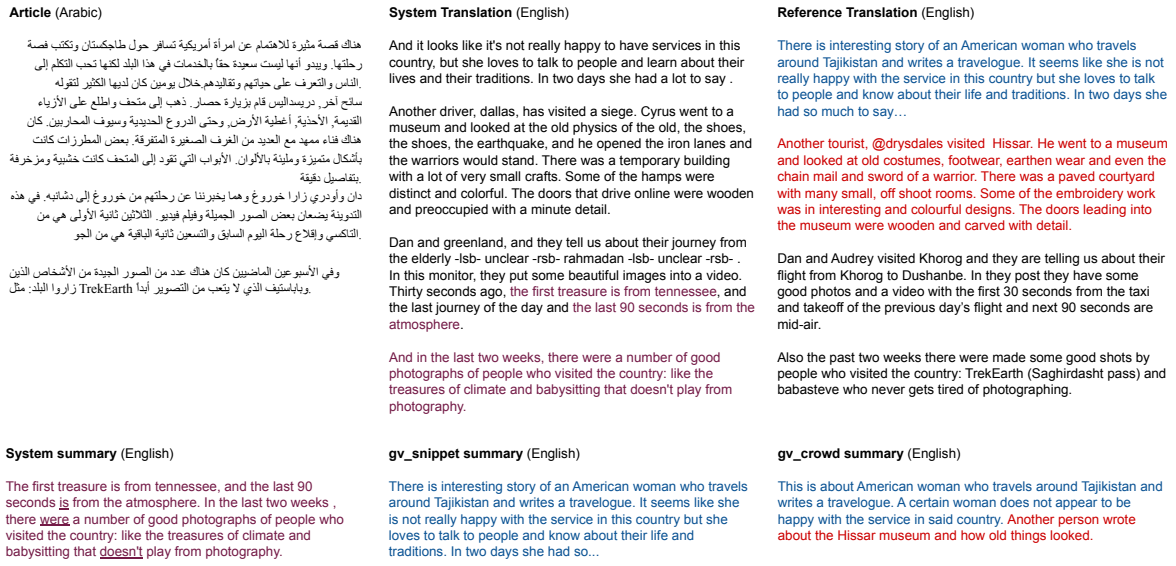

Figure 3: An example in our dataset. The source document is originally written in English and is translated into Arabic by a Global Voices translator. Our translation system translates the Arabic article into English poorly. The summarization system mostly copies segments from the translation and carries grammatical errors (underlined) from the translation to its summary. The `gv-snippet` summary is a mere copy of the first few sentences of the English version of the article (though this may not always be the case in other articles). On the other hand, the `gv-crowd` summary offers better coverage, including information in the second paragraph. Note that this article is challenging to summarize perfectly in 50 words because it features four different parallel stories at the same time. Here, the `gv-crowd` summarizer trades off coverage for specificity of the stories.

Method	Spanish-English	French-English	German-English	Arabic-English
Translation quality (BLEU ↑)				
Transformer	37.45	29.80	19.34	10.77
Summarization quality evaluated on `gv-snippet` (ROUGE-1|2|L F-1 scores ↑)				
FIRST50	63.7 | 55.1 | 61.3	64.7 | 56.2 | 62.3	65.2 | 57.1 | 63.0	62.9 | 53.5 | 60.5
PERFECTTRANS	38.0 | 22.1 | 34.0	38.1 | 21.8 | 34.0	37.7 | 21.9 | 33.6	36.8 | 20.0 | 32.7
TRANSTHENSUM	33.0 | 12.4 | 28.4	32.0 | 10.6 | 27.2	28.3 | 7.4 | 23.7	24.5 | 4.3 | 20.4
Summarization quality evaluated on `gv-crowd` (ROUGE-1|2|L F-1 scores ↑)				
FIRST50	46.4 | 23.4 | 40.4	46.0 | 22.8 | 40.1	47.4 | 25.7 | 40.9	45.9 | 22.9 | 40.4
PERFECTTRANS	36.1 | 13.5 | 31.3	36.7 | 13.7 | 31.7	36.6 | 14.1 | 31.6	36.9 | 14.0 | 31.9
TRANSTHENSUM	35.1 | 10.6 | 30.0	33.3 | 8.9 | 28.5	29.4 | 6.0 | 25.0	26.0 | 3.8 | 22.1

Table 3: Cross-lingual summarization results with different approaches. Translation quality is measured on the `gv-snippet` articles, of which the `gv-crowd` articles are a subset.

achieves the highest BLEU score (34.45) due to the amount of training data and the closeness between the language pair; on the other spectrum, the Arabic-English model offers poorest translations (10.77). Nevertheless, despite the large gaps in BLEU scores, we observe much smaller divergences in ROUGE-1 and ROUGE-L scores. For example, in the extreme case of Arabic-English, even when the BLEU drops by almost 90% when switch from the reference to the predicted translations, the ROUGE-L F1-score only decreases by only about 30%. This observation highlights a major limitation of ROUGE-1 and ROUGE-L: their insensitivity to the summary readability. Even though a source document may contain meaningless, ungrammatical contents (reflected by a low BLEU score), a model that summarizes by simply copying phrases can easily achieve high ROUGE-1 and ROUGE-L scores. This limitation is difficult to observe in the context of monolingual summarization because the source documents come from natural sources and thus are mostly grammatical and meaningful. Another interesting finding is that the FIRST50 baseline achieves higher ROUGE

scores when evaluated on `gv-snippet` than on `gv-crowd`. This observation indicates that the `gv-snippet` summaries overlap highly with the beginning part of the articles, confirming the results from our human preference survey that these summaries generally have poorer coverage over the entire articles than the `gv-crowd` summaries.

4 Conclusion

This work introduces a dataset for evaluating multilingual and cross-lingual summarization methods in multiple languages. Future work aims to extend the dataset to more languages and construct a large-scale training dataset. Another interesting direction is to study whether multi-task learning can benefit cross-lingual summarization. To take advantage of the fact that translating the entire source article may not be necessary, it would be useful to teach models to devise more efficient translation strategies by informing them of the downstream summarization objective.

Acknowledgement

This material is based upon work supported by the National Science Foundation under Grant No. IIS-1618193. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D Trippe, Juan B Gutierrez, and Krys Kochut. 2017. Text summarization techniques: a brief survey. *International Journal of Advanced Computer Science and Applications(IJACSA)*.

Hoa Trang Dang and Karolina Owczarzak. 2008. Overview of the tac 2008 update summarization task. In *TAC*.

Hoa Trang Dang and Karolina Owczarzak. 2010. Overview of tac 2010 summarization track. In *TAC*.

Yue Dong. 2018. A survey on neural network-based summarization methods. *arXiv preprint arXiv:1804.04589*.

Kevin Duh. 2018. The multitarget ted talks task. http://www.cs.jhu.edu/~kevinduh/a/multitarget-tedtalks/.

Mahak Gambhir and Vishal Gupta. 2017. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, 47(1):1–66.

Sebastian Gehrmann, Yuntian Deng, and Alexander M Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of Emperical Methods in Natural Language Processing*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in neural information processing systems*, pages 1693–1701.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016a. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016b. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. In *Proceedings of the Association for Computational Linguistics*.

C Orsan and Oana Andreea Chiorean. 2008. Evaluation of a cross-lingual romanian-english multi-document summariser. In *International Language Resources and Evaluation*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Jessica Ouyang, Boya Song, and Kathleen McKeown. 2019. 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)*, pages 2025–2031.

Paul Over, Hoa Dang, and Donna Harman. 2007. Duc in context. *Information Processing & Management*, 43(6):1506–1520.

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

Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium*.

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 *Advances in neural information processing systems*, pages 5998–6008.

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. Association for Computational Linguistics.

Taha Zerrouki. 2010. pyarabic, an arabic language library for python.