CROWDSOURCED PHRASE-BASED TOKENIZATION FOR LOW-RESOURCED NEURAL MACHINE TRANSLATION: THE CASE OF FON LANGUAGE

Anonymous authors
Paper under double-blind review

ABSTRACT

Building effective neural machine translation (NMT) models for very low-resourced and morphologically rich African indigenous languages is an open challenge. Besides the issue of finding available resources for them, a lot of work is put into preprocessing and tokenization. Recent studies have shown that standard tokenization methods do not always adequately deal with the grammatical, diacritical, and tonal properties of some African languages. That, coupled with the extremely low availability of training samples, hinders the production of reliable NMT models. In this paper, using Fon language as a case study, we revisit standard tokenization methods and introduce Word-Expressions-Based (WEB) tokenization, a human-involved super-words tokenization strategy to create a better representative vocabulary for training.

1 INTRODUCTION

Tokenization, the process by which texts are divided into minimal meaningful units, is generally viewed as a solved problem. Yet, in practice, we often encounter difficulties in using standard tokenizers for NMT tasks. This may be because of special tokenization needs for particular domains (like medicine (He & Kayaalp, 2006; Cruz Díaz & Maña López, 2015)), or languages. As an example of a language with special tokenization needs, let's look at Fon language, one of the five classes of the Gbe language clusters (Aja, Ewe, Fon, Gen, and Phla-Phera according to (Capo, 2010)), spoken by approximately 1.7 million people located in southwestern Nigeria, Benin, Togo, and southeastern Ghana. There exists approximately 53 different dialects of Fon spoken throughout Benin. Fon has complex grammar and syntax, is very tonal and diacritics are highly influential (Dossou & Emezue, 2020). Despite being spoken by 1.7 million speakers, Joshi et al. (2020) have categorized Fon as «left behind» or «understudied» in NLP. This poses a challenge when using standard tokenization methods.

In this paper, using Fon in our experiment, we examine standard tokenization methods, and introduce the **Word-Expressions-Based (WEB) tokenization**, and test it on the Fon-French and French-Fon translation tasks. Our main contributions are the dataset, our analysis and the proposal of WEB for extremely low-resourced African languages (ALRLs). The dataset, models and codes will be open-sourced on our Github page.

2 BACKGROUND AND MOTIVATION

Modern NMT models usually require large amount of parallel data in order to effectively learn the representations of morphologically rich source and target languages. While proposed solutions such as transfer-learning

from a high-resource language (HRL) to the low-resource language (LRL) (Gu et al., 2018; Renduchintala et al., 2018; Karakanta et al., 2018), and using monolingual data (Sennrich et al., 2016a; Zhang & Zong, 2016; Burlot & Yvon, 2018; Hoang et al., 2018) have proved effective, they are still not able to produce better translation results for most ALRLs. Standard tokenization methods, like Subword Units (SU) (Sennrich et al., 2015), inspired by the byte-pair-encoding (BPE) (Gage, 1994), have greatly improved current NMT systems. However, studies have shown that BPE does not always boost performance of NMT systems for analytical languages(Abbott & Martinus, 2018). Ngo et al. (2019) show that when morphological differences exist between source and target languages, SU does not significantly improve results. Therefore, there is a great need to revisit NMT with a focus on low-resourced, morphologically complex languages like Fon. This may involve taking a look at how to adapt standard NMT strategies to these languages.

3 Tokenization Strategies and their Challenges for Fon

In this section, we briefly discuss the standard tokenization strategies employed in NMT, as well as challenges faced while applying them to Fon.

Word-Based tokenization (WB) consists of splitting sentences into words, according to a delimiter. We'll show the limits of this method using this Fon expression: *«un ɖo ganji»*. *«un»* on its own is an interjection, to express an emotion of surprise or astonishment. But «un ɖo» already means *"I am"*, *"I am at"*, or *"I have"*, depending on the context in which it is used. The whole expression, *«un ɖo ganji»*, could mean "I am fine" or "I am okay".

Phrase-Based tokenization (PhB) encodes phrases (group of words) as atomic units, instead of words. As a result, models trained on PhB have the ability to learn and interpret language-specific phrases (noun, verbal and prepositional phrases), making it better than WB for Fon language. However, due to the randomness of PhB alignments, some extracted pairs are not always contextually faultless. For example, the computer alignment gave respectively [zɛn, une (a, an, one)] and [azɔn, la (the)], instead of [zɛn, une marmite (a pot)] and [azɔn, la maladie (the disease)].

Encoding with SU has made great headway in NMT, especially due to its ability to effectively encode rare out-of-vocabulary words (Sennrich et al., 2016b). Machácek et al. (2018), in analyzing the word segmentation for NMT, reported that the common property of BPE and SU relies on the distribution of character sequences, but disregards any morphological properties of the languages in question. Apart from rule-based tokenization, there are machine learning approaches to tokenization as well, which unfortunately require a substantial amount of training samples (both original and tokenized versions of the same texts) (Riley, 1989; Mikheev, 2000; Jurish & Würzner, 2013). To the best of our knowledge, there is no known language-specific tokenization proposed for ALRLs, although there have been a number of works on adapting NMT specifically to them (like (Orife et al., 2020; van Biljon et al., 2020; Vaswani et al., 2017), to mention but a few).

4 Word-Expressions-Based tokenization (WEB)

WEB involves manually aligning and extracting meaningful expressions based on linguistic components of Fon (phonemes, morphemes, lexemes, syntax, and context), with the crucial help of Fon-French native speakers. Some examples of good alignments are [kuɖo jigbézǎn, joyeux anniversaire (Happy Birthday)], [nɔncé, maman (mum)], and [nɔncé vivɛ, maman chérie (dear mum)]. The following algorithm was designed to encode and decode sentences using the established vocabulary:

1. **Run** through the vocabulary and output a list L of all possible word combinations of words and expressions appearing in the sentence S.
2. Important principle in Fon: higher word orders = more precise and meaningful expressions. Using this principle, for each word, $w \in L$,
   (a) **Check** if there exists a higher word order, $v \in L$, such that $w \subsetneq v$.
   (b) **If** 2a is true, discard w, **else** keep w.
3. The output is a list $\hat{L}$ of optimal expressions that from the initial L, making up the initial sentence S.
4. Encode $\hat{L}$

We argue that WEB scales well because it does not require any linguistic annotations but knowledge and intuitions from bilinguals, meaning, we can crowdsource those phrases. We want to state clearly, in order to avoid any confusion, that WEB is another version of PhB, involving human evaluation. For our study, it took a group of 8 people, all bilinguals speaking Fon and French, and 30 days to align and extract meaningful sentences manually.

5 THE FON-FRENCH DATASET: DATA COLLECTION, CLEANING AND EXPANSION PROCESSES

As our goal is to create a reliable translation system to be used by the modern Fon-speaking community, we set out to gather more data on daily conversations domain for this study. Thanks to many collaborations with Fon-French bilinguals, journalists and linguists, we gathered daily citations, proverbs and sentences with their French translations. After the collection's stage, we obtained a dataset of 8074 pairs of Fon-French sentences.

The cleaning process, which involved the Fon-French bilinguals, mainly consisted of analyzing the contextual meanings of the Fon sentences, and checking the quality of the French translations. In many cases, where the French translations were really bad, we made significant corrections.

Another major observation was the presence of many long and complex sentences. That's where the idea of expanding the dataset came from: we proceeded to split, when possible, Fon sentences into short, independent, and meaningful expressions (expression of 1-6 words), and accordingly add their respective French translations. At the end of these processes, we obtained our final dataset of 25,383 pairs of Fon-French sentences. The experiments, described in this paper, were conducted using the final dataset.

We strongly believe that involving the Fon-French bilinguals into the cleaning process, greatly improved the quality of the dataset. In fact, many initial translation errors were disregarded by standard, rule-based tokenization (like WB, PhB and SU) and cleaning techniques[1]. However, with the help of the **«intuitive or natural»** language knowledge of the Fon-French bilinguals, bunch of those errors has been fixed. This highlights, the importance of having native speakers of ALRLs to clean and review the dataset, during the initial stages of its compilation.

6 METHODOLOGY, RESULTS AND CONCLUSION

In this section, we describe the implementation of WB, PhB, SU, WEB and we compare the results of our NMT model trained on them for our analysis.

[1]Using Python Regex and String packages (https://docs.python.org/3/library/re.html) together with NLTK preprocessing library (https://www.nltk.org/)

6.1 CREATION OF VOCABULARIES FOR WB, PHB, SU AND WEB

For WB, we split the sentences according to the standard 'space' delimiter, using the TensorFlow-Keras text tokenizer[2], getting a vocabulary of 7,845 and 8,756 Fon and French tokens (words) respectively.

For PhB, we use the IBM1 model from nltk.translate.api module[3] to align and extract all possible pairs of sentences. Our main observation was that, some pairs generated were either not meaningful or not maching, but we didn't try to rearrange them in order to see how well the generated pairs, without human intervention, would affect the translation quality. In so doing, we got a vocabulary of 10,576 and 11,724 Fon and French tokens respectively (word and expressions).

To implement SU , we used TensorFlow's SubwordTextEncoder[4] and built a vocabulary of 7,952 and 8,116 Fon and French tokens (words and subwords) respectively.

To implement WEB, we considered unique expressions as atomic units. Using the steps highlighted for WEB in section 4, we encoded those atomic units and obtained a vocabulary of 18,759 and 19,785 Fon and French tokens (word and expressions) used for the model training.

6.2 DATASET SPLITTING, MODEL'S ARCHITECTURE AND TRAINING.

From the dataset, we carefully selected 155 long and complex sentences i.e. sentences made of 5 or more expressions, as test data; sentences that we believe, would test the model's ability to correctly translate higher word order expressions in Fon. 10% of the training data, was set aside for validation.

For training, we used an encoder-decoder-based architecture (Sutskever et al., 2014), made up of 128-dimensional gated rectified units (GRUs) recurrent layers(Cho et al., 2014), with a word embedding layer of dimension 256 and a 10-dimensional attention model (Bahdanau et al., 2015).

We trained with a batch size of 100, learning rate of 0.001 and 500 epochs, using validation loss to track model performance.

All training processes took 14 days on a 16GB Tesla K80 GPU. We evaluated our NMT models performances using BLEU (Papineni et al., 2002), METEOR (Banerjee & Lavie, 2005), CharacTER (TER) (Wang et al., 2016), and GLEU (Wu et al., 2016) metrics.

6.3 RESULTS AND CONCLUSION:

The results from Table 2 and Table 1 show that our baseline model performs better with PhB, and best with WEB, in terms of metric and translation quality. It is important to note that while BLEU of other methods reduced on the Fr→Fon task, WB improved on it. We speculate that this might be because WB enhanced the model's understanding of French expressions over Fon. This corroborates our argument that in order to help NMT systems to translate ALRLs better, it is paramount to create adequate tokenization processes that can represent their morphology. We also speculate that SU performed quite poorly because the dataset contains mostly short-medium sentences.

This is a pilot project and there is headroom to be explored with improving WEB. We are also working on combining WEB with SU, to get the best of both worlds. To promote research and reproducibility in this direction, the dataset and model will be made publicly available on Github after the review. Simultaneously, we are working on releasing the platform for the service, by the end of 2020. We believe that it would be a good way to gather more data and keep constantly improving the model's performance.

[2]https://www.tensorflow.org

[3]https://www.nltk.org/api/nltk.translate.html

[4]http://www.tensorflow.org

	Sentences(with English for better understanding)
Input	wa miɗu wazɔ ɗokpɔ
Target (Eng)	viens qu'on travaille ensemble (come, let's work together)
WB	imbécile (coward)
PhB	venez travaille ensemble (come work together)
SU	travaille ensemble (work together)
WEB	venez qu'on travaille ensemble (come, let's work together)
Input	kuɗo jigbézǎn nɔncé vivɛ
Target (Eng)	joyeux anniversaire maman chérie (happy birthday dear mum)
WB	manoeuvre (maneuver)
PhB	joyeux anniversaire maman (happy birthday mum)
SU	maman chérie (dear mum)
WEB	joyeux anniversaire maman chérie (happy birthday dear mum)

Table 1: Model translations with WB, PhB, SU and WEB

Translation	**Tokenization**	**BLEU ↑**	**METEOR ↑**	**TER ↓**	**GLEU↑**
Fon → Fr	WB	6.8	12.2	86.2	9
Fon → Fr	SU	7.6	13.6	87.4	10
Fon → Fr	PhB	38.9	53.7	43.9	42
Fon → Fr	**WEB**	**66.6**	**77.77**	**24.2**	**67**
Fr → Fon	WB	15.65	-	-	8
Fr → Fon	SU	25.68	-	-	9
Fr → Fon	PhB	38.74	-	-	25.74
Fr → Fon	**WEB**	**49.37**	-	-	**43**

Table 2: Experimental results of our model trained on WB, SU, PhB and WEB.

REFERENCES

Jade Z. Abbott and Laura Martinus. Towards neural machine translation for african languages. *CoRR*, abs/1811.05467, 2018. URL `http://arxiv.org/abs/1811.05467`.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL `http://arxiv.org/abs/1409.0473`.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Mea-*

sures for Machine Translation and/or Summarization, pp. 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W05-0909.

Franck Burlot and François Yvon. Using monolingual data in neural machine translation: a systematic study. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 144–155, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6315. URL https://www.aclweb.org/anthology/W18-6315.

Hounkpati B.C. Capo. *A Comparative Phonology of Gbe*. De Gruyter Mouton, Berlin, Boston, 2010. ISBN 978-3-11-087053-4.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078, 2014. URL http://arxiv.org/abs/1406.1078.

Noa P. Cruz Díaz and Manuel Maña López. An analysis of biomedical tokenization: Problems and strategies. In *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis*, pp. 40–49, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-2605. URL https://www.aclweb.org/anthology/W15-2605.

Bonaventure F. P. Dossou and Chris C. Emezue. Ffr v1.0: Fon-french neural machine translation. *In Proceedings of the AfricanNLP Workshop, International Conference on Learning Representations*, arXiv:arXiv:2003.12111, 2020.

Philip Gage. A new algorithm for data compression. *C Users J.*, 12(2):23–38, February 1994. ISSN 0898-9788.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. Universal neural machine translation for extremely low resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 344–354, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1032. URL https://www.aclweb.org/anthology/N18-1032.

Ying He and Mehmet Kayaalp. *A Comparison of 13 Tokenizers on MEDLINE*. 12 2006. doi: 10.13140/2.1.1133.1206.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pp. 18–24, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-2703. URL https://www.aclweb.org/anthology/W18-2703.

Pratik M. Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the nlp world. *ArXiv*, abs/2004.09095, 2020.

Bryan Jurish and Kay-Michael Würzner. Word and sentence tokenization with hidden markov models. *J. Lang. Technol. Comput. Linguistics*, 28:61–83, 2013.

Alina Karakanta, Jon Dehdari, and Josef Genabith. Neural machine translation for low-resource languages without parallel corpora. *Machine Translation*, 32(1–2):167–189, June 2018. ISSN 0922-6567. doi: 10.1007/s10590-017-9203-5. URL https://doi.org/10.1007/s10590-017-9203-5.

Dominik Machácek, Jonás Vidra, and Ondrej Bojar. Morphological and language-agnostic word segmentation for NMT. *CoRR*, abs/1806.05482, 2018. URL http://arxiv.org/abs/1806.05482.

Andrei Mikheev. Tagging sentence boundaries. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, NAACL 2000, pp. 264–271, USA, 2000. Association for Computational Linguistics.

Thi-Vinh Ngo, Thanh-Le Ha, Phuong-Thai Nguyen, and Le-Minh Nguyen. Overcoming the rare word problem for low-resource language pairs in neural machine translation. In *Proceedings of the 6th Workshop on Asian Translation*, pp. 207–214, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5228. URL https://www.aclweb.org/anthology/D19-5228.

Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo, Musie Meressa, Espoir Murhabazi, Orevaoghene Ahia, Elan van Biljon, Arshath Ramkilowan, Adewale Akinfaderin, Alp Öktem, Wole Akin, Ghollah Kioko, Kevin Degila, Herman Kamper, Bonaventure Dossou, Chris Emezue, Kelechi Ogueji, and Abdallah Bashir. Masakhane – machine translation for africa, 2020.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://www.aclweb.org/anthology/P02-1040.

Adithya Renduchintala, Pamela Shapiro, Kevin Duh, and Philipp Koehn. Character-aware decoder for neural machine translation. *CoRR*, abs/1809.02223, 2018. URL http://arxiv.org/abs/1809.02223.

Michael D. Riley. Some applications of tree-based modelling to speech and language. In *Proceedings of the Workshop on Speech and Natural Language*, HLT '89, pp. 339–352, USA, 1989. Association for Computational Linguistics. ISBN 1558601120. doi: 10.3115/1075434.1075492. URL https://doi.org/10.3115/1075434.1075492.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *CoRR*, abs/1508.07909, 2015. URL http://arxiv.org/abs/1508.07909.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 86–96, Berlin, Germany, August 2016a. Association for Computational Linguistics. doi: 10.18653/v1/P16-1009. URL https://www.aclweb.org/anthology/P16-1009.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725, Berlin, Germany, August 2016b. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL https://www.aclweb.org/anthology/P16-1162.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215, 2014. URL http://arxiv.org/abs/1409.3215.

Elan van Biljon, Arnu Pretorius, and Julia Kreutzer. On optimal transformer depth for low-resource language translation. *In Proceedings of the International Conference on Learning Representations*, https://arxiv.org/abs/2004.04418, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL http://arxiv.org/abs/1706.03762.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. CharacTer: Translation edit rate on character level. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pp. 505–510, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-2342. URL https://www.aclweb.org/anthology/W16-2342.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016. URL http://arxiv.org/abs/1609.08144.

Jiajun Zhang and Chengqing Zong. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1535, Austin, Texas, nov 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1160. URL https://www.aclweb.org/anthology/D16-1160.