On the use of linguistic similarities to improve Neural Machine Translation for African Languages

Pascal Tikeng[1]**, Brice Nanda** [1]**, James Assiene**[2]
[1]National Advanced School of Engineering Yaounde
[2] MILA-Quebec AI Institute

Abstract

In recent years, there has been a resurgence in research on empirical methods for machine translation. Most of this research has been focused on high-resource, European languages. Despite the fact that around 30% of all languages spoken worldwide are African, the latter have been heavily under investigated and this, partly due to the lack of public parallel corpora online. Furthermore, despite their large number (more than 2,000) and the similarities between them, there is currently no publicly available study on how to use this multilingualism (and associated similarities) to improve machine translation systems performance on African languages. So as to address these issues, we propose a new dataset (from a source that allows us to use and release) for African languages that provides parallel data for vernaculars not present in commonly used dataset like JW300. To exploit multilingualism, we first use a historical approach based on migrations of population to identify similar vernaculars. We also propose a new metric to automatically evaluate similarities between languages. This new metric does not require word level parallelism like traditional methods but only paragraph level parallelism. We then show that performing Masked Language Modelling and Translation Language Modeling in addition to multi-task learning on a cluster of similar languages leads to a strong boost of performance in translating individual pairs inside this cluster. In particular, we record an improvement of 29 BLEU on the pair Bafia-Ewondo using our approaches compared to previous work methods that did not exploit multilingualism in any way. Finally, we release the dataset and code of this work to ensure reproducibility and accelerate research in this domain.

Introduction

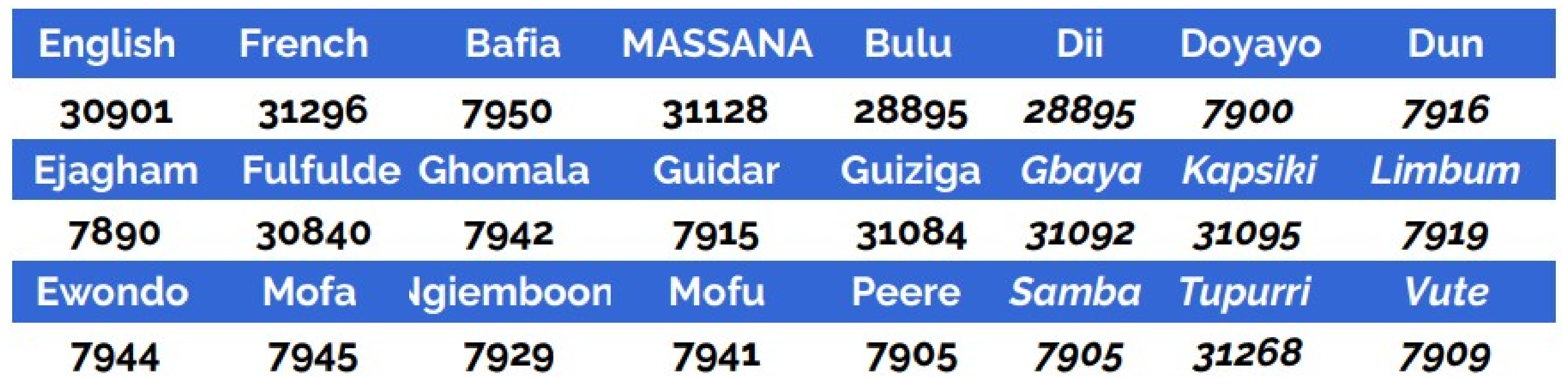
Machine Translation (MT) of African languages (AL) is a challenging problem because of multiple reasons. As pointed out by Martinus et al., 2019, the main ones are: morphological complexity and diversity of AL, lack/absence of large parallel datasets for most language pairs, discoverability (the existing resources for AL are often hard to find), reproducibility (data and code of existing research are rarely shared), lack of benchmarks. However, despite the strong multilingualism of the continent, previous research focused solely on translating individual pairs of language without taking advantage of the potential similarities between all of them. The main purpose of this work is to exploit this multilingualism, starting from a few languages spoken in Central and West Africa, to produce better machine translation systems. Our contributions can be summarised as follows :

1. We provide new parallel corpora extracted from the Bible for several pairs of African vernacular that were not available until now as well as the code used to perform this extraction.
2. We present a method for aggregating languages together based on their historical origins, their morphologies, their geographical and cultural distributions etc... We also propose of a new metric to evaluate similarity between languages : this metric, based on language models, doesn't require word level parallelism (contrary to traditional like (Levenshtein, 1965) and (Swadesh, 1952)) but only paragraph level parallelism. It also takes into account the lack of translation of words present in (Swadesh, 1952) but not in African vernaculars (like "snow").
3. Using the language clusters created using the previous similarities, we show that Translation Language Modelling (TLM) (Lample et al., 2019) and multi-task learning generally improve the performance on individual pairs inside these clusters.
4. We make our data, code and benchmark publicly available.

Methodology and Results

Data

YouVersion provides biblical content in 1,756 languages [1] among which many (local) African ones. We have extracted this content in the desired languages to build our dataset. In particular, for some countries like Cameroon, YouVersion provide more languages (22 for the latter) than JW300 (Agić et al., 2019) (18 for Cameroon) that is the most used data source for work on African languages (e.g Masakhane (2020)).



English	French	Bafia	MASSANA	Bulu	Dii	Doyayo	Dun
30901	31296	7950	31128	28895	*28895*	*7900*	*7916*
Ejagham	Fulfulde	Ghomala	Guidar	Guiziga	*Gbaya*	*Kapsiki*	*Limbum*
7890	30840	7942	7915	31084	*31092*	*31095*	*7919*
Ewondo	Mofa	ŋgiemboon	Mofu	Peere	*Samba*	*Tupurri*	*Vute*
7944	7945	7929	7941	7905	*7905*	*31268*	*7909*

Figure 1: Number of sentences per language. The number of sentences for a language pair (L_1, L_2) is equal to $min(|L_1|, |L_2|)$ where $|L_i|$ represents the number of sentences of the language L_i

Historical approach for languages aggregation.

The intersection between our dataset and Eberhard et al. (2020) analysis led to the following languages clusters.

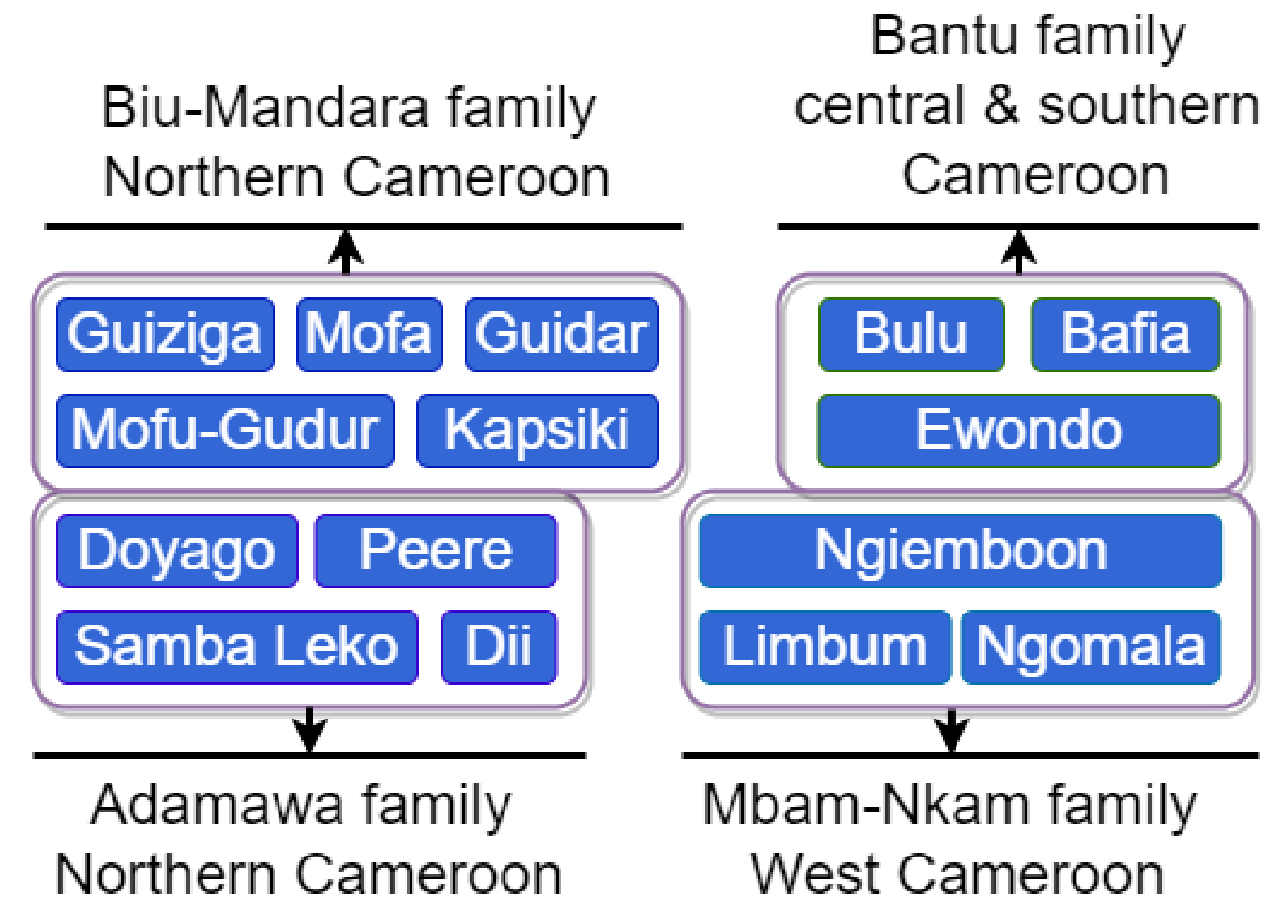


Figure 2: Historical approach for languages aggregation

Language model based approach for lingustic similarities.

Given a trained language model LM_0 in language L_0 and two parallel corpora C_0 in L_0 and C_1 in language L_1, we define the LM-similarity between L_0 and L_1 according to LM_0 and C_0 by

$$LMS_{LM_0,C_0}(L_0, L_1) = \frac{PPL_{LM_0}(C_1) - PPL_{LM_0}(C_0)}{PPL_{LM_0}(C_0)} \quad (1)$$

where $PPL_{LM_0}(C_i)$ is the perplexity of the language model LM_0 on the corpus C_i. Despite the fact that this metric is not symmetric, it can be used to build clusters using a hierarchical clustering approach. We start by a set of 2 languages $G_1 = \{L_1, L_2\}$, the language L^* that will be added to G_1 can be determined using the following formula (inspired by the unweighted average linkage clustering criteria) :

$$L^* = \underset{L_i \in \mathcal{L} \setminus G_1}{\arg\min} \frac{1}{2}\left(LMS_{LM_1,C_1}(L_1, L_i) + LMS_{LM_2,C_2}(L_2, L_i)\right) \quad (2)$$

In other words, L^* is the language in the set of possible languages $\mathcal{L}$ (but we exclude $G_1 = \{L_1, L_2\}$) that minimises the average "distance" (LM-similarity) from L_1 and L_2. By recursively applying this formula, one can build a cluster of any size.

Experiments and Results

For each ordered language pair in the cluster of interest *(Bafia, Ewondo, Bulu)*, we examine the MT performance (using BLEU score (Papineni et al., 2002)) on that pair if we add a third language to the pair fo form a cluster. The model will thus become a multi-task machine translation model capable of translating 3 languages rather than 2. The language (among the 22 in our dataset) to add is chosen in 3 ways in our work: using a historical (**Hist**) approach (*Bulu*), using the LM-similarity (**LM**), purely at random (**Rand**). We also compare the performance with model trained on the pair of interest without any form of MLM/TLM pre-training (**None**), and on the pair of interest with MLM+TLM pre-training (**Pair**). Having identified our clusters and to further exploit multilingualism, we follow the approach of Lample et al., 2019 by pre-training the model's encoder with MLM and TLM objectives on the cluster languages, then fine-tune the model on these same languages.

Table 1: Languages used respectively for the Random, Historical and LM columns

Pair	Random	Historical	LM
Bafia-Bulu	Du_ n	Ewondo	Mofa
Bafia-Ewondo	Guidar	Bulu	Bulu
Bulu-Ewondo	Dii	Bafia	Mofa

Table 2: Machine Translation BLEU scores (Ba=Bafia, Bu=Bulu, Ew=Ewondo, Fr=French)

Pair	None	Pair	Rand	Hist	LM
Ba-Bu	9.19	12.58	23.52	**28.81**	13.03
Bu-Ba	13.5	15.15	24.76	**32.83**	13.91
Ba-Ew	9.30	11.28	8.28	**38.90**	**38.90**
Ew-Ba	13.99	16.07	10.26	**35.84**	**35.84**
Bu-Ew	10.27	12.11	11.82	**39.12**	34.86
Ew-Bu	11.62	14.42	12.27	**34.91**	30.98

Pair	None	Pair	Hist
Fr-Bu	19.91	23.47	**25.06**
Bu-Fr	17.49	22.44	**23.68**
Fr-Ba	14.48	15.35	**30.65**
Ba-Fr	8.59	11.17	**24.49**
Fr-Ew	11.51	13.93	**35.50**
Ew-Fr	10.60	13.77	**27.34**

References

[1] Željko Agić et al. "JW300: A Wide-Coverage Parallel Corpus for Low-Resource Languages". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 3204–3210. DOI: 10.18653/v1/P19-1310. URL: https://www.aclweb.org/anthology/P19-1310.

[2] Eberhard et al. *Ethnologue: Languages of the world*. Dallas, Texas: SIL Internationnal, 2020.

[3] Guillaume Lample et al. "Cross-lingual language model pretraining". In: arXiv:1901.07291 (2019). URL: https://arxiv.org/abs/1901.07291.

[4] Vladimir Levenshtein. "Binary codes capable of correcting deletions, insertions, and reversals". In: *Dokl. Akad. Nauk SSSR* (1965).

[5] Laura Martinus et al. "A focus on neural machine translation for african languages". In: *arXiv preprint arXiv:1906.05685* (2019).

[6] Masakhane. "MACHINE TRANSLATION FOR AFRICA". In: arXiv:2020.11529 (2020). URL: https://arxiv.org/pdf/2003.11529.pdf.

[7] Kishore Papineni et al. "BLEU: a Method for Automatic Evaluation of Machine Translation". In: 2002, pp. 311–318.

[8] Morris Swadesh. "Lexico-statistical dating of prehistoric ethniccontacts: With special reference to north american indians and eskimos". In: *In Proceedings of the American Philosophical Society, volume 96, pages 452–463.* (1952).

[1] https://www.bible.com/en/languages (October 22, 2021)