HAL
open science

How Does Language Influence Documentation Workflow? Unsupervised Word Discovery Using Translations in Multiple Languages

Marcely Zanon Boito, Aline Villavicencio, Laurent Besacier

▶ **To cite this version:**

Marcely Zanon Boito, Aline Villavicencio, Laurent Besacier. How Does Language Influence Documentation Workflow? Unsupervised Word Discovery Using Translations in Multiple Languages. Journées Scientifiques du Groupement de Recherche: Linguistique Informatique, Formelle et de Terrain (LIFT)., Nov 2019, Orléans, France. hal-02895895

HAL Id: hal-02895895

https://hal.science/hal-02895895v1

Submitted on 10 Jul 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

arXiv:1910.05154v1 [cs.CL] 11 Oct 2019

How Does Language Influence Documentation Workflow? Unsupervised Word Discovery Using Translations in Multiple Languages

Marcely Zanon Boito[1] Aline Villavicencio[2,3] Laurent Besacier[1]
(1) Laboratoire d'Informatique de Grenoble (LIG), UGA, G-INP, CNRS, INRIA, France
(2) Department of Computer Science, University of Sheffield, England
(3) Institute of Informatics (INF), UFRGS, Brazil
contact: marcely.zanon-boito@univ-grenoble-alpes.fr

RÉSUMÉ

Comment la langue influence le processus de documentation ? Découverte non supervisée de mots basée sur des traductions en langues multiples

Pour la documentation des langues, la transcription est un processus très coûteux : une minute d'enregistrement nécessiterait environ une heure et demie de travail pour un linguiste (Austin and Sallabank, 2013). Récemment, la collecte de traductions (dans des langues bien documentées) alignées aux enregistrements est devenue une solution populaire pour garantir l'interprétabilité des enregistrements (Adda et al., 2016) et aider à leur traitement automatique. Dans cet article, nous étudions l'impact de la langue de traduction sur les approches automatiques en documentation des langues. Nous traduisons un corpus parallèle bilingue Mboshi-Français (Godard et al., 2017) dans quatre autres langues, et évaluons l'impact de la langue de traduction sur une tâche de segmentation en mots non supervisée. Nos résultats suggèrent que la langue de traduction peut influencer légèrement la qualité de segmentation. Cependant, combiner l'information apprise par différents modèles bilingues nous permet d'améliorer ces résultats de manière marginale.

ABSTRACT

For language documentation initiatives, transcription is an expensive resource: one minute of audio is estimated to take one hour and a half on average of a linguist's work (Austin and Sallabank, 2013). Recently, collecting aligned translations in well-resourced languages became a popular solution for ensuring posterior interpretability of the recordings (Adda et al., 2016). In this paper we investigate language-related impact in automatic approaches for computational language documentation. We translate the bilingual Mboshi-French parallel corpus (Godard et al., 2017) into four other languages, and we perform bilingual-rooted unsupervised word discovery. Our results hint towards an impact of the well-resourced language in the quality of the output. However, by combining the information learned by different bilingual models, we are only able to marginally increase the quality of the segmentation.

MOTS-CLÉS : découverte non supervisée du lexique, documentation des langues, approches multilingues.

KEYWORDS: unsupervised word discovery, language documentation, multilingual approaches.

1 Introduction

The *Cambridge Handbook of Endangered Languages* (Austin and Sallabank, 2011) estimates that at least half of the 7,000 languages currently spoken worldwide will no longer exist by the end of this century. For these *endangered* languages, data collection campaigns have to accommodate the challenge that many of them are from oral tradition, and producing transcriptions is costly. This *transcription bottleneck* problem can be handled by translating into a widely spoken language to ensure subsequent interpretability of the collected recordings, and such parallel corpora have been recently created by aligning the collected audio with translations in a well-resourced language (Adda et al., 2016; Godard et al., 2017; Boito et al., 2018). Moreover, some linguists suggested that more than one translation should be collected to capture deeper layers of meaning (Evans and Sasse, 2004).

This work is a contribution to the Computational Language Documentation (CLD) research field, that aims to replace part of the manual steps performed by linguists during language documentation initiatives by automatic approaches. Here we investigate the unsupervised word discovery and segmentation task, using the bilingual-rooted approach from Godard et al. (2018). There, words in the well-resourced language are aligned to unsegmented phonemes in the endangered language in order to identify group of phonemes, and to cluster them into word-like units. We experiment with the Mboshi-French parallel corpus, translating the French text into four other well-resourced languages in order to investigate language impact in this CLD approach. Our results hint that this language impact exists, and that models based on different languages will output different word-like units.

2 Methodology

The Multilingual Mboshi Parallel Corpus: In this work we extend the bilingual Mboshi-French parallel corpus (Godard et al., 2017), fruit of the documentation process of Mboshi (Bantu C25), an endangered language spoken in Congo-Brazzaville. The corpus contains 5,130 utterances, for which it provides audio, transcriptions and translations in French. We translate the French into four other well-resourced languages through the use of the $DeepL$ translator.[1] The languages added to the dataset are: English, German, Portuguese and Spanish. Table 1 shows some statistics for the produced *Multilingual Mboshi* parallel corpus.[2]

Bilingual Unsupervised Word Segmentation/Discovery Approach: We use the bilingual neural-based Unsupervised Word Segmentation (UWS) approach from Godard et al. (2018) to discover words in Mboshi. In this approach, Neural Machine Translation (NMT) models are trained between language pairs, using as source language the translation (word-level) and as target, the language to document (unsegmented phonemic sequence). Due to the attention mechanism present in these networks (Bahdanau et al., 2014), posterior to training, it is possible to retrieve *soft-alignment probability matrices* between source and target sequences. These matrices give us sentence-level source-to-target alignment information, and by using it for clustering neighbor phonemes aligned to the same translation word, we are able to create segmentation in the target side. The product of this approach is a set of (discovered-units, translation words) pairs.

[1] Available at `https://www.deepl.com/translator`

[2] Available at `https://github.com/mzboito/mmboshi`

	MB	FR	EN	ES	DE	PT
# Types	6,633	5,178	4,392	5,473	5,641	5,465
# Tokens	30,556	42,715	37,379	37,428	37,515	37,095
Avg. Token Length	4.18	4.41	4.19	4.36	4.91	4.40
Avg. Tokens/Sentence	5.96	8.33	7.29	7.30	7.31	7.23

Table 1: Statistics for the Multilingual Mboshi parallel corpus. The French text is used for generating translation in the four other languages present in the right side of the table.

Bilingual			Multilingual Voting					ANE Selection
1	**FR**	73.40		25%	50%	75%	100%	
2	**EN**	73.10	**(1-2)**	73.10	73.10	73.30	73.30	73.80
3	**PT**	72.80	**(1-3)**	72.40	74.60	72.10	72.10	73.90
4	**ES**	72.60	**(1-4)**	71.60	74.80	74.20	70.90	73.90
5	**DE**	71.00	**(1-5)**	74.30	74.90	73.10	70.00	73.90

Table 2: From left to right, results for: bilingual UWS, multilingual leveraging by voting, ANE selection.

Multilingual Leveraging: In this work we apply two simple methods for including multilingual information into the bilingual models from Godard et al. (2018). The first one, **Multilingual Voting**, consists of merging the information learned by models trained with different language pairs by performing a voting over the final discovered boundaries. The voting is performed by applying an agreement threshold T over the output boundaries. This threshold balances between accepting all boundaries from all the bilingual models (zero agreement) and accepting only input boundaries discovered by all these models (total agreement). The second method is **ANE Selection**. For every language pair and aligned sentence in the dataset, a soft-alignment probability matrix is generated. We use *Average Normalized Entropy* (ANE) (Boito et al., 2019a) computed over these matrices for selecting *the most confident one* for segmenting each phoneme sequence. This exploits the idea that models trained on different language pairs will have language-related behavior, thus differing on the resulting alignment and segmentation over the same phoneme sequence.

3 Experiments

The experiment settings from this paper and evaluation protocol for the Mboshi corpus (Boundary F-scores using the ZRC speech reference) are the same from Boito et al. (2019a). Table 2 presents the results for bilingual UWS and multilingual leveraging. For the former, we reach our best result by using as aligned information the French, the original aligned language for this dataset. Languages closely related to French (Spanish and Portuguese) ranked better, while our worst result used German. English also performs notably well in our experiments. We believe this is due to the statistics features of the resulting text. We observe in Table 1 that the English portion of the dataset contains the smallest vocabulary among all languages. Since we train our systems in very low-resource settings, vocabulary-related features can impact greatly the system's capacity to language-model, and consequently the final quality of the produced alignments. Even in high-resource settings, it was already attested that some languages are more difficult to model than others (Cotterell et al., 2018).

For the multilingual selection experiments, we experimented combining the languages from top to bottom as they appear Table 2 (ranked by performance; e.g. 1-3 means the combination of FR(1),

MB-DE		MB-EN		MB-ES		MB-FR		MB-PT	
itua	itoua	ibara	ibara	ingobha	ingobha	itua	itoua	oboá+ngá	oboa
mɯndzɯ	monzo	otsɛ́ngɛ	otsenge	ondóngo	ondongo	itúa+ngá	itoua	**ERROR**	nyaamvua
tsimba	tsimba	asúa	asoua	mbía+mbvúlá	amvoulou	itúa+mbía	itoua	itua	itoua
abia	Freunde	okúmú	okoumou	itua	itoua	kánga	pintade	mbembe	mbembe
tsósá	Henne	olangi	bottle	y'+konga	cuerno	**ERROR**	nyobhosi	tsimba	tsimba
ibara	Ibara	tsɛ́si	hare	itúa+ngá	itoua	oboá+ngá	oboa	okwɯɯ	cordão
andzɯi	Elefanten	itua	itoua	oboá+ngá	oboa	kyéma	singe	mómeá	tentar
ondúma	Onduma	kóli	badger	ekoko	ekoko	**ERROR**	amassez	abvúɛ	cunhado
ikinyi	Fliege	andzúe	bees	okubha	herrero	tsimba	tsimba	ekoko	ekoko
itúa+ngá	itoua	itúa+mbía	itoua	ibara	ibara	lekú+yá	guêpes	mbósi	cabras

Table 3: Top 10 confident (discovered type, translation) pairs for the five bilingual models. The "+" mark means the discovered type is a concatenation of two existing true types.

EN(2) and PT(3)). We observe that the performance improvement is smaller than the one observed in previous work (Boito et al., 2019b), which we attribute to the fact that our dataset was artificially augmented. This could result in the available multilingual form of supervision not being as rich as in a manually generated dataset. Finally, the best boundary segmentation result is obtained by performing multilingual voting with all the languages and an agreement of 50%, which indicates that the information learned by different languages will provide additional complementary evidence.

Lastly, following the methodology from Boito et al. (2019a), we extract the most confident alignments (in terms of ANE) discovered by the bilingual models. Table 3 presents the top 10 most confident (discovered type, translation) pairs.[3] Looking at the pairs the bilingual models are most confident about, we observe there are some types discovered by all the bilingual models (e.g. Mboshi word *itua*, and the concatenation *oboá+ngá*). However, the models still differ for most of their alignments in the table. This hints that while a portion of the lexicon might be captured independently of the language used, other structures might be more dependent of the chosen language. On this note, Haspelmath (2011) suggests the notion of *word* cannot always be meaningfully defined cross-linguistically.

4 Conclusion

In this work we train bilingual UWS models using the endangered language Mboshi as target and different well-resourced languages as aligned information. Results show that similar languages rank better in terms of segmentation performance, and that by combining the information learned by different models, segmentation is further improved. This might be due to the different language-dependent structures that are captured by using more than one language. Lastly, we extend the bilingual Mboshi-French parallel corpus, creating a multilingual corpus for the endangered language Mboshi that we make available to the community.

[3]The Mboshi phoneme sequences were replaced by their grapheme equivalents to increase readability, but all results were computed using phonemes.

References

Adda, G., Stüker, S., Adda-Decker, M., Ambouroue, O., Besacier, L., Blachon, D., Bonneau-Maynard, H., Godard, P., Hamlaoui, F., Idiatov, D., Kouarata, G.-N., Lamel, L., Makasso, E.-M., Rialland, A., de Velde, M. V., Yvon, F., and Zerbian, S. (2016). Breaking the unwritten language barrier: The BULB project. *Procedia Computer Science*, 81:8–14.

Austin, P. K. and Sallabank, J. (2011). *The Cambridge handbook of endangered languages*. Cambridge University Press.

Austin, P. K. and Sallabank, J. (2013). *Endangered languages*. Taylor & Francis.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Boito, M. Z., Anastasopoulos, A., Lekakou, M., Villavicencio, A., and Besacier, L. (2018). A small griko-italian speech translation corpus. *arXiv preprint arXiv:1807.10740*.

Boito, M. Z., Villavicencio, A., and Besacier, L. (2019a). Empirical evaluation of sequence-to-sequence models for word discovery in low-resource settings. *arXiv preprint arXiv:1907.00184*.

Boito, M. Z., Villavicencio, A., and Besacier, L. (2019b). Leveraging translations in multiple languages for low-resource unsupervised word segmentation. Unpublished work. Paper under review.

Cotterell, R., Mielke, S. J., Eisner, J., and Roark, B. (2018). Are all languages equally hard to language-model? *arXiv preprint arXiv:1806.03743*.

Evans, N. and Sasse, H.-J. (2004). Searching for meaning in the library of babel: field semantics and problems of digital archiving. Open Conference Systems, University of Sydney, Faculty of Arts.

Godard, P., Adda, G., Adda-Decker, M., Benjumea, J., Besacier, L., Cooper-Leavitt, J., Kouarata, G.-N., Lamel, L., Maynard, H., Müller, M., et al. (2017). A very low resource language speech corpus for computational language documentation experiments. *arXiv preprint arXiv:1710.03501*.

Godard, P., Zanon Boito, M., Ondel, L., Berard, A., Yvon, F., Villavicencio, A., and Besacier, L. (2018). Unsupervised word segmentation from speech with attention. In *Interspeech*.

Haspelmath, M. (2011). The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia linguistica*, 45(1):31–80.