# Unsupervised Word Discovery Using Attentional Encoder-Decoder Models

**Marcely Zanon Boito**
LIG, Grenoble Alpes University
Institute of Informatics, UFRGS
marcely.zanon-boito
@grenoble-inp.org

**Laurent Besacier**
LIG, Grenoble Alpes University
laurent.besacier@imag.fr

**Aline Villavicencio**
Institute of Informatics, UFRGS
avillavicencio@inf.ufrgs.br

## Abstract

Attention-based sequence-to-sequence neural machine translation systems have been shown to jointly align and translate source sentences into target sentences. In this project we use unsegmented symbol sequences (characters and phonemes) as source, aiming to explore the soft-alignment probability matrices generated during training and to evaluate if these soft-alignments allow us to discover latent lexicon representations.

If successful, such approach could be useful for documenting unwritten and/or endangered languages. However, for this to be feasible, attention models should be robust to low-resource scenarios, of several thousand of sentences only. We use a parallel corpus between the endangered language Mboshi and French, as well as a larger and more controlled English-French parallel corpus. Our goal is to explore different representation levels and study their impact, together with the impact of different data set sizes, in the quality of the generated soft-alignment probability matrices.

## 1 Introduction

The general consensus between specialists is that there are around 7000 languages currently spoken in the world, and between 50 and 90% of them will become extinct by the year 2100 (Austin and Sallabank, 2011). Even with a joint effort from the linguistics community, manually documenting all these languages before their extinction is not feasible. Recently, initiatives for helping with this issue include organizing tasks (Versteegh et al., 2016; Jansen et al., 2013) and offering tools and methodologies to help to automate (part of) this documentation process (Besacier et al., 2006; Bartels et al., 2016; Bansal et al., 2016; Lignos and Yang, 2010; Anastasopoulos and Chiang, 2017).

Endangered languages are characterized by the small number of speakers and often by the lack of a well-defined written form, which makes their documentation an even more challenging task. To deal with the absence of standard written form, collected speech can be aligned to its translation in a well-documented language. The resulting parallel corpora, though, usually lack size.

Nonparametric bayesian models (Goldwater et al., 2009; Lee et al., 2015; Elsner et al., 2013; Adams et al., 2015, 2016) and Neural Network systems (Duong et al., 2016; Bérard et al., 2016; Franke et al., 2016) emerged as popular approaches for phonetic unit discovery, unsupervised segmentation and lexicon discovery, common subtasks to the documentation process. Our ongoing project covers unsupervised segmentation and lexicon discovery, and we are interested in examining the performance impact of executing these tasks from different representation levels. We approach grapheme and phonetic representation, and later we wish to extend our methodology to raw speech.

In this work, we present preliminary results using the attention models soft-alignment probability matrices from a global attention-based sequence-to-sequence Neural Machine Translation (NMT) system as the starting point in our unsupervised segmentation process. We investigate if this approach is realistic using a small corpus from an endangered language, and we compare our results against a nonparametric bayesian model (Goldwater et al., 2009).

We define our architecture in a way that allows us to easily extend it for working directly from raw signal (Bérard et al., 2016; Weiss et al., 2017) in the future, which would be ideal for endangered languages that lack written form. We are also interested in discovering how much data is necessary to achieve good segmentation and lexicon discov-

ery results, and consequently, how applicable this approach is to the endangered languages case.

## 2 Related Work

Encoder-decoder NMT architectures using attention were first presented in Bahdanau et al. (2014), and we use the implementation of Bérard et al. (2016), an end-to-end translation architecture that can work directly from raw speech. Attention-based NMT systems are known for producing not only good translations, but also attentional information in the form of soft-alignment probability matrices. They demonstrate how these architectures jointly learn to align and translate. We believe this information can be useful for both segmentation and lexicon discovery.

The work by Duong et al. (2016) is the most similar to ours. They also used attention models for their unsupervised segmentation task, achieving very good results compared to three baselines. The Spanish-English parallel corpus used in their work was approximately 18,300 sentences long.

In comparison to that, in this project we use a small parallel corpus from a real unwritten language, for which we study the applicability of the proposed approach for language documentation considering the limitations in data size.

## 3 Methodology and Preliminary Results

We use a 5,157 sentences parallel corpus in an unwritten[1] African language called Mboshi (Bantu C 25), aligned to French translations on sentence level. Mboshi is a language spoken in the north of the Republic of the Congo, and it counts with 32 different phonemes (25 consonants and 7 vowels) and two tones (high and low). The corpus was recorded using the LIG-AIKUMA tool (Blachon et al., 2016) in the scope of the BULB project (Adda et al., 2016), and preliminary experiments were reported by Godard et al. (2016).

Our approach consists of using the entire corpus for training[2] a global attention sequence-to-sequence NMT system, leaving nothing for testing, since we are not interested in the translations. Then we extract the soft-alignments probability matrices for all the sentences used for training, and we use these matrices to transform the soft-alignment information in hard alignment. We do so by selecting the target word that maximizes the probability of the input symbol given all the target possibilities.

| | **Recall** | **Precision** | **F-Score** |
|---|---|---|---|
| base_model | 6.53 | 3.17 | 4.27 |
| base_s | 8.39 | 5.38 | 6.56 |
| reverse | 20.04 | 10.02 | 13.36 |
| reverse_s | 22.29 | 17.15 | 19.39 |
| dpseg* | 19.73 | 36.20 | 25.54 |

Table 1: Results for the unsupervised segmentation task of tokens using 4,643 parallel sentences. The "s" identifies the models' smoothed versions.

In order to validate our architecture we executed a version using the gold standard segmentation for Mboshi as source. That allowed us to discover if our data set was enough to generate good soft-alignments in the ideal scenario where we already have the segmentation. For this analysis, the evaluation was qualitative, and we observed very good alignments between known Mboshi words and their translations[3].

For a more realistic setup we replaced the source by its unsegmented version. The results had noisy and unhelpful soft-alignment probability matrices, what can be verified by precision and recall being both low. We also trained a model using the alignment smoothing described in Duong et al. (2016), what helped the model's performance. The results are respectively base_model and base_s at Table 1.

In more details, this alignment smoothing is applied by training the model with a temperature factor in the softmax function. The resulting probability matrices are further smoothed by replacing each probability $\alpha_{ij}$ by $\frac{1}{3}(\alpha_{i,j-1} + \alpha_{i,j} + \alpha_{i,j+1})$, $i$ and $j$ being respectively the target words and source symbols indexes.

Evaluating the matrices generated by this first model, we observed that the system was consistently ignoring part of the source symbols when generating the translation. In NMT systems, the soft-alignments are created forcing the probabilities for each target word $j$ to sum to one, what ensures all the target words are used. However, there is no similar constraint for the source symbols, as discussed in Duong et al. (2016).

[1] Even if it is unwritten, we have a non-standard grapheme form, considered to be close to the language phonology.

[2] 10% for development set, which corresponds to 514 sentences, and the remaining 4,643 sentences for training.

[3] We had access to a small Mboshi-French dictionary (Beapami et al., 2000).

Considering that we are interested in segmentation, our system must use all these source units from the unsegmented input when processing a sentence. To solve this, we reversed the system input, creating a French-Mboshi words-to-characters system. As we can see in Table 1, this constraint impacted greatly in the segmentation performance. The addition of alignment smoothing further improved the system performance.

Finally, for comparison, we executed the nonparametric bayesian model implemented in dpseg[4] (Goldwater et al., 2009), using it as an out-of-the-box tool. We used default configurations for the bigram model and 20,000 iterations. We considered the achieved result to be a lower bound result for this technique in this scenario.

The out-of-the-box trained nonparametric bayesian model presented better overall results than our reverse neural model. This is consistent and expected, since bayesian models are known for being able to achieve good segmentation with small amounts of data. In the other hand, neural approaches are known for needing large data sets to train their parameters.

Moreover, even if we still can apply some optimizations to our model, we do believe there is a limit of how much is achievable with this amount of data. Unpublished results to which we had access in our laboratory investigated the dependency between data set size and the soft-alignment probability matrices quality.

It seems that, even when the model performs well in translation, sometimes that amount of data is not enough to create consistent soft-alignment matrices, and the network learns a global sentence representation which is not meaningful for us. In that case, adding more data to the model seems to make the soft-alignment matrices "converge" to the desirable representation.

## 4 Conclusion and Future Experiments

In this work we presented our preliminary results approaching the task of unsupervised segmentation. We used a neural machine translation system to retrieve soft-alignment information using a data set from a real endangered language.

By reversing source and target languages, we were able to achieve interesting results considering the amount of data available. However, these results are still inferior to what we can achieve with bayesian systems such as dpseg.

We are following our experiments by using a large English-French corpus to study the impact that more data can have in the soft-alignment matrices quality. Doing so, we want to narrow down the amount of data needed in order to retrieve enough information from these alignments. This will answer how applicable this approach is for language documentation scenarios.

We are also investigating a semi-supervised approach. We believe that by offering some already segmented units (such as function words), we could improve the system's performance segmenting the rest of the vocabulary.

Finally, we want to investigate how different representation levels for the source impact the amount of data needed, and if it is possible to achieve good results working directly from speech. In this scenario, we would like to explore how reducing information in the target side, by replacing the translations by their lemmas or part of speech, could help to decrease the amount of data needed for this task.

[4] Available at http://homepages.inf.ed.ac.uk/sgwater/.

## References


Oliver Adams, Graham Neubig, Trevor Cohn, and Steven Bird. 2015. Inducing bilingual lexicons from small quantities of sentence-aligned phonemic transcriptions. In *12th International Workshop on Spoken Language Translation (IWSLT)*.

Oliver Adams, Graham Neubig, Trevor Cohn, and Steven Bird. 2016. Learning a translation model from word lattices. In *17th Annual Conference of the International Speech Communication Association (INTERSPEECH 2016), San Francisco, California, USA*.

Gilles Adda, Sebastian Stüker, Martine Adda-Decker, Odette Ambouroue, Laurent Besacier, David Blachon, Hélène Bonneau-Maynard, Pierre Godard, Fatima Hamlaoui, Dmitry Idiatov, et al. 2016. Breaking the unwritten language barrier: The bulb project. *Procedia Computer Science* 81:8–14.

Antonios Anastasopoulos and David Chiang. 2017. A case study on using speech-to-translation alignments for language documentation. *arXiv preprint arXiv:1702.04372* .

Peter K Austin and Julia Sallabank. 2011. *The Cambridge handbook of endangered languages*. Cambridge University Press.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly

learning to align and translate. *arXiv preprint arXiv:1409.0473* .

Sameer Bansal, Herman Kamper, Sharon Goldwater, and Adam Lopez. 2016. Weakly supervised spoken term discovery using cross-lingual side information. *arXiv preprint arXiv:1609.06530* .

Chris Bartels, Wen Wang, Vikramjit Mitra, Colleen Richey, Andreas Kathol, Dimitra Vergyri, Harry Bratt, and Chiachi Hung. 2016. Toward human-assisted lexical unit discovery without text resources. In *Spoken Language Technology Workshop (SLT), 2016 IEEE*. IEEE, pages 64–70.

Roch Paulin Beapami, Ruth Chatfield, Guy-Noël Kouarata, and Andrea Embengue-Waldschmidt. 2000. *Dictionnaire Mbochi-Franais*. SIL-Congo Publishers.

Alexandre Bérard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. *arXiv preprint arXiv:1612.01744* .

Laurent Besacier, Bowen Zhou, and Yuqing Gao. 2006. Towards speech translation of non written languages. In *Spoken Language Technology Workshop, 2006. IEEE*. IEEE, pages 222–225.

David Blachon, Elodie Gauthier, Laurent Besacier, Guy-Noël Kouarata, Martine Adda-Decker, and Annie Rialland. 2016. Parallel speech collection for under-resourced language studies using the lig-aikuma mobile device app. *Procedia Computer Science* 81:61–66.

Long Duong, Antonios Anastasopoulos, David Chiang, Steven Bird, and Trevor Cohn. 2016. An attentional model for speech translation without transcription. In *Proceedings of NAACL-HLT*. pages 949–959.

Micha Elsner, Sharon Goldwater, Naomi Feldman, and Frank Wood. 2013. A joint learning model of word segmentation, lexical acquisition, and phonetic variability. In *Proc. EMNLP*.

Joerg Franke, Markus Mueller, Fatima Hamlaoui, Sebastian Stueker, and Alex Waibel. 2016. Phoneme boundary detection using deep bidirectional lstms. In *Speech Communication; 12. ITG Symposium; Proceedings of*. VDE, pages 1–5.

Pierre Godard, Gilles Adda, Martine Adda-Decker, Alexandre Allauzen, Laurent Besacier, Helene Bonneau-Maynard, Guy-Noël Kouarata, Kevin Löser, Annie Rialland, and François Yvon. 2016. Preliminary experiments on unsupervised word discovery in mboshi. In *Interspeech 2016*.

Sharon Goldwater, Thomas L Griffiths, and Mark Johnson. 2009. A bayesian framework for word segmentation: Exploring the effects of context. *Cognition* 112(1):21–54.

Aren Jansen, Emmanuel Dupoux, Sharon Goldwater, Mark Johnson, Sanjeev Khudanpur, Kenneth Church, Naomi Feldman, Hynek Hermansky, Florian Metze, Richard Rose, et al. 2013. A summary of the 2012 jhu clsp workshop on zero resource speech technologies and models of early language acquisition .

Chia-ying Lee, Timothy J O'Donnell, and James Glass. 2015. Unsupervised lexicon discovery from acoustic input. *Transactions of the Association for Computational Linguistics* 3:389–403.

Constantine Lignos and Charles Yang. 2010. Recession segmentation: simpler online word segmentation using limited resources. In *Proceedings of the fourteenth conference on computational natural language learning*. Association for Computational Linguistics, pages 88–97.

Maarten Versteegh, Xavier Anguera, Aren Jansen, and Emmanuel Dupoux. 2016. The zero resource speech challenge 2015: Proposed approaches and results. *Procedia Computer Science* 81:67–72.

Ron J Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017. Sequence-to-sequence models can directly transcribe foreign speech. *arXiv preprint arXiv:1703.08581* .