Empirical Evaluation of Sequence-to-Sequence Models for Word Discovery in Low-resource Settings
Marcely Zanon Boito, Aline Villavicencio, Laurent Besacier

To cite this version:
Marcely Zanon Boito, Aline Villavicencio, Laurent Besacier. Empirical Evaluation of Sequence-to-Sequence Models for Word Discovery in Low-resource Settings. Interspeech 2019, Sep 2019, Graz, Austria. hal-02193867

HAL Id: hal-02193867
https://hal.archives-ouvertes.fr/hal-02193867
Submitted on 24 Jul 2019

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.
Empirical Evaluation of Sequence-to-Sequence Models for Word Discovery in Low-resource Settings

Marcely Zanon Boito¹, Aline Villavicencio²,³, Laurent Besacier¹

¹Laboratoire d’Informatique de Grenoble, Univ. Grenoble Alpes (UGA), France
²School of Computer Science and Electronic Engineering, University of Essex, UK
³Institute of Informatics, Federal University of Rio Grande do Sul, Brazil

contact: marcely.zanon-boito@univ-grenoble-alpes.fr

Abstract

Since Bahdanau et al. [1] first introduced attention for neural machine translation, most sequence-to-sequence models made use of attention mechanisms [2, 3, 4]. While they produce soft-alignment matrices that could be interpreted as alignment between target and source languages, we lack metrics to quantify their quality, being unclear which approach produces the best alignments. This paper presents an empirical evaluation of 3 of the main sequence-to-sequence models for word discovery from unsegmented phoneme sequences: CNN, RNN and Transformer-based. This task consists in aligning word sequences in a source language with phoneme sequences in a target language, inferring from it word segmentation on the target side [5]. Evaluating word segmentation quality can be seen as an extrinsic evaluation of the soft-alignment matrices produced during training. Our experiments in a low-resource scenario on Mboshi and English languages (both aligned to French) show that RNNs surprisingly outperform CNNs and Transformer for this task. Our results are confirmed by an intrinsic evaluation of alignment quality through the use Average Normalized Entropy (ANE). Lastly, we improve our best word discovery model by using an alignment entropy confidence measure that accumulates ANE over all the occurrences of a given alignment pair in the collection.

Index Terms: sequence-to-sequence models, soft-alignment matrices, word discovery, low-resource languages, computational language documentation

1. Introduction

Sequence-to-Sequence (S2S) models can solve many tasks where source and target sequences have different lengths. For learning to focus on specific parts of the input at decoding time, most of these models are equipped with attention mechanisms [1, 2, 3, 4]. By-products of the attention are soft-alignment probability matrices, that can be interpreted as alignment between target and source. However, we lack metrics to quantify their quality. Moreover, while these models perform very well in a typical use case, it is not clear how they would be quantified their quality. Moreover, while these models perform very well in a typical use case, it is not clear how they would be affected by low-resource scenarios.

This paper proposes an empirical evaluation of well-known S2S models for a particular S2S modeling task. This task consists of aligning word sequences in a source language with phoneme sequences in a target language, inferring from it word segmentation on the target side [5]. We concentrate on three models: Convolutional Neural Networks (CNN) [6], Recurrent Neural Networks (RNN) [1] and Transformer-based models [3]. While this word segmentation task can be used for the extrinsic evaluation of the soft-alignment probability matrices produced during S2S learning, we also introduce Average Normalized Entropy (ANE), a task-agnostic confidence metric to quantify the quality of the source-to-target alignments obtained. Experiments performed on a low-resource scenario for two languages (Mboshi and English) using equivalently sized corpora aligned to French, are, to our knowledge, the first empirical evaluation of these well-known S2S models for a word segmentation task. We also illustrate how our entropy-based metric can be used in a language documentation scenario, helping a linguist to efficiently discover types, in an unknown language, from an unsegmented sequence of phonemes. This work is thus also a contribution to the emerging computational language documentation domain [7, 8, 9, 10, 11], whose main goal is the creation of automatic approaches able to help the documentation of the many languages soon to be extinct [12].

Lastly, studies focused on comprehensive attention mechanisms for NMT [13, 14, 15] lack evaluation of the resulting alignments, and the exceptions [16] do so for the task of word-to-word alignment in well-resourced languages. Differently, our work is not only an empirical evaluation of NMT models focused on alignment quality, but it also tackles data scarcity of low-resource scenarios.

2. Experimental Settings

2.1. Unsupervised Word Segmentation from Speech

As in language documentation scenarios available corpora usually contain speech in the language to document aligned with translations in a well-resourced language, Godard et al. [5] introduced a pipeline for performing Unsupervised Word Segmentation (UWS) from speech. The system outputs timestamps delimiting stretches of speech, associated with class labels, corresponding to real words in the language. The pipeline consists of first transforming speech into a sequence of phonemes, either through Automatic Unit Discovery (e.g. [17]) or manual transcription. The phoneme sequences, together with their translations, are then fed to an attention-based S2S system that produces soft-alignment probability matrices between target and source languages. The alignment probability distributions between the phonemes and the translation words (as in Figure 1) are used to cluster (segment) together neighbor phonemes whose alignment distribution peaks at the same word. The final speech segmentation is evaluated using the Zero Resource Challenge (ZRC) 2017 evaluation suite (track 2) [1].

¹Available at http://zerospeech.com/2017
²We increment over [5] by removing silence labels before training, and using them for segmentation. This results in slightly better scores.
Figure 1: Soft-alignment probability matrices from the UWS task. ANE values (from left to right) are 0.11, 0.64 and 0.83. The gold segmentation is “BAH1T MAA1MAH0 PAA1PAH0 IH0Z AW1IT”, which corresponds to the English sentence “But mama, papa is out”.

2.2. Parallel Speech Corpora

The parallel speech corpora used in this work are the English-French (EN-FR) [13] and the Mboshi-French (MB-FR) [19] parallel corpora. EN-FR corpus is a 33,192 sentences multilingual extension from librispeech [20], with English audio books automatically aligned to French translations. MB-FR is a 5,130 sentences corpus from the language documentation process of Mboshi (Bantu C25), an endangered language spoken in Congo-Brazzaville. Thus, while the former corpus presents larger vocabulary and longer sentences, the latter presents a more tailored environment, with short sentences and simpler vocabulary. In order to provide a fair comparison, as well as to study the impact of corpus size, the EN-FR corpus was also down-sampled to 5K utterances (to the exact same size than the MB-FR corpus). Sub-sampling was conducted preserving the average number of tokens per sentence, shown in Table 1.

2.3. Introducing Average Normalized Entropy (ANE)

In this paper, we focus on studying the soft-alignment probability matrices resulting from the learning of S2S models for the UWS task. To assess the overall quality of these matrices without having gold alignment information, we introduce Average Normalized Entropy (ANE).

Definition: Given the source and target pair \((s, t)\) of lengths \(|s|\) and \(|t|\) respectively, for every phone \(t_i\), the normalized entropy (NE) is computed considering all possible words in \(s\) (Equation 1), where \(P(t_i, s_j)\) is the alignment probability between the phone \(t_i\) and the word \(s_j\) (a cell in the matrix). The ANE for a sentence is then defined by the arithmetic mean over the resulting NE for every phone from the sequence \(t\) (Equation 2).

\[
NE(t_i, s) = -\sum_{j=1}^{|s|} P(t_i, s_j) \cdot \log_2(s_j) (P(t_i, s_j))
\]

\[
ANE(t, s) = \frac{\sum_{i=1}^{|t|} NE(t_i, s)}{|t|}
\]

From this definition, we can derive ANE for different granularities (sub or supra-sentential) by accumulating its value for the full corpus, for a single type or for a single token. Corpus ANE will be used to summarize the overall performance of a S2S model on a specific corpus. Token ANE extends ANE to tokens by averaging NE for all phonemes from a single (discovered) token. Type ANE results from averaging the ANE for every token instance of a discovered type. Finally, Alignment ANE is the result of averaging the ANE for every discovered (type, translation word) alignment pair. Intuition that lower ANEs correspond to better alignments is exemplified in Figure 1.

3. Empirical Comparison of S2S Models

We compare three NMT models [31, 32, 33] for UWS, focusing on their ability of aligning words (French) with phonemes (English or Mboshi) in medium-low resource settings. The results, an analysis of the impact of data size and quality, and the correlation between intrinsic (ANE) and extrinsic (boundary F-score) metrics are presented in [34]. The application of ANE for type discovery in low-resource settings is presented in [35].

3.1. RNN: Attention-based Encoder-Decoder

The classic RNN encoder-decoder model [1] connects a bidirectional encoder with an unidirectional decoder by the use of an alignment module. The RNN encoder learns annotations for every source token, and these are weighted by the alignment module for the generation of every target token. Weights are defined as context vectors, since they capture the importance of every source token for the generation of each target token.

Attention mechanism: a context vector for a decoder step \(t\) is computed using the set of source annotations \(H\) and the last state of the decoder network (translation context). The attention is the result of the weighted sum of the source annotations \(H\) (with \(H = h_1, \ldots, h_A\)) and their \(\alpha\) probabilities (3) obtained through a feed-forward network align (4).

\[
c_t = \text{Att}(H, s_{t-1}) = \sum_{i=1}^A \alpha_i^t h_i
\]

\[
\alpha_i^t = \text{softmax}(\text{align}(h_i, s_t))
\]

3.2. Transformer

Transformer [3] is a fully attentional S2S architecture, which has obtained state-of-the-art results for several NMT shared tasks. It replaces the use of sequential cell units (such as LSTM) by Multi-Head Attention (MHA) operations, which make the architecture considerably faster. Both encoder and decoder networks are stacked layers sets that receive source and target sequences, embedded and concatenated with positional encoding. An encoder layer is made of two sub-layers: a Self-Attention MHA and a feed-forward. A decoder layer is made of three sub-layers: a masked Self-Attention MHA (no access to subsequent positions); an Encoder-Decoder MHA (operation over the encoder stack’s final output and the decoder self-attention output); and a feed-forward sub-layer. Dropout and residual connections are applied between all sub-layers. Final output probabilities are generated by applying a linear projection over the decoder stack’s output, followed by a softmax operation.

Multi-head attention mechanism: attention is seen as a mapping problem: given a pair of key-value vectors and a query vector, the task is the computation of the weighted sum of the given values (output). In this setup, weights are learned by compatibility functions between key-query pairs (of dimension \(d_k\)). For a given query (Q), keys (K) and values (V) set, the Scaled
Dot-Product (SDP) Attention function is computed as:

$$\text{Att}(V, K, Q) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$ (5)

In practice, several attentions are computed for a given QKV set. The QKV set is first projected into \( h \) different spaces (multiple heads), where the SDP attention is computed in parallel. Resulting values for all heads are then concatenated and once again projected, yielding the layer’s output. (6) and (7) illustrate the process, in which \( H \) is the set of \( h \) heads \((H = h_1, ..., h_h)\) and \( f \) is a linear projection. Self-Attention defines the case where query and values come from same source (learning compatibility functions within the same sequence of elements).

$$\text{MultiHead}(V, K, Q) = f(\text{Concat}(H))$$ (6)

$$h_i = \text{Att}(f_i(V), f_i(K), f_i(Q))$$ (7)

3.3. CNN: Pervasive Attention

Different from the previous models, which are based on encoder-decoder structures interfaced by attention mechanisms, this approach relies on a single 2D CNN across both sequences (no separate coding stages)\(^3\). Using masked convolutions, an auto-regressive model predicts the next output symbol based on a joint representation of both input and partial output sequences. Given a source-target pair \((s, t)\) of lengths \(|s|\) and \(|t|\) respectively, tokens are first embed in \( d_s \) and \( d_t \) dimensional spaces via look-up tables. Token embeddings \( \{x_1, ..., x_{|s|}\} \) and \( \{y_1, ..., y_{|t|}\} \) are then concatenated to form a 3D tensor \( X \in \mathbb{R}^{|t| \times |s| \times d} \) with \( f_0 = d_1 + d_2 \), where:

$$X_{ij} = [y_j, x_j]$$ (8)

Each convolutional layer \( l \in \{1, ..., L\} \) of the model produces a tensor \( H_l \) of size \(|t| \times |s| \times f_l \), where \( f_l \) is the number of output channels for that layer. To compute a distribution over the tokens in the output vocabulary, the second dimension of the tensor is used. This dimension is of variable length (given by the input sequence) and it is collapsed by max or average pooling to obtain the tensor \( H^p_{pool} \) of size \(|t| \times f_l \). Finally, 1×1 convolution followed by a softmax operation are applied, resulting in the distribution over the target vocabulary for the next output token.

Attention mechanism: joint encoding acts as an attention-like mechanism, since individual source elements are re-encoded as the output is generated. The self-attention approach of \(21\) is applied. It computes the attention weight tensor \( \alpha \), of size \(|t| \times |s|\), from the last activation tensor \( H_L \), to pool the elements of the same tensor along the same dimension, as follows:

$$\alpha = \text{softmax}(W_1 \tanh(H_L W_2))$$ (9)

$$H^\alpha_{full} = \alpha H_L$$ (10)

where \( W_1 \in \mathbb{R}^{d \times f} \) and \( W_2 \in \mathbb{R}^{f \times f} \) are weight tensors that map the \( f \) dimensional features in \( H_L \) to the attention weights via an \( f \) dimensional intermediate representation.

3.4. Comparing S2S Architectures

For each S2S architecture, and each of the three corpora, we train five models (runs) with different initialization seeds\(^4\). Before segmenting, we average the produced matrices from the five different runs as in \(3\). Evaluation is done in a bilingual segmentation condition that corresponds to the real UWS task. In addition, we also perform segmentation in a monolingual condition, where a phoneme sequence is segmented with regards to the corresponding word sequence (transcription) in the same language (hence monolingual)\(^5\). Our networks are optimized for the monolingual task. Across all architectures, we use embeddings of size 64 and batch size of 32 (5K data set), or embeddings of size 128 and batch size of 64 (33K data set). Dropout of 0.5 and early-stopping procedure are applied in all cases. RNN models have only one layer, a bi-directional encoder, and cell size equal to the embedding size, as in \(3\). CNN models use the hyper-parameters from \(5\) with only 3 layers (5K data set), or 6 (33K data set), and kernel size of 3. Transformer models were optimized starting from the original hyper-parameters of \(3\). Best results (among 50 setups) were achieved using 2 heads, 3 layers (encoder and decoder), warm-up of 5K steps, and using cross-entropy loss without label-smoothing. Finally, for selecting which head to use for UWS, we experimented using the last layer’s averaged heads, or by selecting the head with minimum corpus ANE. While the results were not significantly different, we kept the ANE selection.

3.4.1. Unsupervised Word Segmentation Results

The word boundary F-scores\(^6\) for the task of UWS from phoneme sequence (in Mboshi or English) are presented in Table 2 with monolingual results shown for information only (topline). Surprisingly, RNN models outperform the more recent (CNN and Transformer) approaches. One possible explanation is the lower number of parameters (for a 5K setup, in average 700K parameters are trained, while CNN needs an additional 30.79% and Transformer 5.31%). However, for 33K setups, CNNs actually need 30% less parameters than RNNs, but still perform worse. Transformer’s low performance could be due to the use of several heads “distributing” alignment information across different matrices. Nonetheless, we evaluated averaged heads and single-head models, and these resulted in significant decreases in performance. This suggests that this architecture may not need to learn explicit alignment to translate, but instead it could be capturing different kinds of linguistic information, as discussed in the original paper and in its examples \(3\). Also, on the decoder side, the behavior of the self-attention mechanism on phoneme units is unclear and under-

\(^3\)RNN, CNN and Transformer implementations from \(22, 2, 23\) respectively.

\(^4\)This task can be seen as an automatic extraction of a pronunciation lexicon from parallel words/phonemes sequences.

\(^5\)For CNN and RNN, average standard deviation for the bilingual task is of less than 0.8%. For Transformer, it is almost 4%.

Table 1: Statistics of the three source-target data sets.

<table>
<thead>
<tr>
<th>corpus</th>
<th>#types</th>
<th>#tokens</th>
<th>average( token length)</th>
<th>average( #tokens / sentence)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>source</td>
<td>target</td>
<td>source</td>
<td>target</td>
</tr>
<tr>
<td>EN-FR (33k)</td>
<td>21,083</td>
<td>33,135</td>
<td>381,044</td>
<td>467,475</td>
</tr>
<tr>
<td>EN-FR (5k)</td>
<td>8,740</td>
<td>12,226</td>
<td>59,090</td>
<td>72,670</td>
</tr>
<tr>
<td>MB-FR (5k)</td>
<td>6,633</td>
<td>5,162</td>
<td>30,556</td>
<td>42,715</td>
</tr>
</tbody>
</table>
studied so far. For the encoder, Voita et al. performed after-training encoder head removal based on head confidence, showing that after initial training, most heads were not necessary for maintaining translation performance. Hence, we find the Multi-head mechanism interpretation challenging, and maybe not suitable for a direct word segmentation application, such as our method.

As in, our best UWS method (RNN) for the bilingual task does not reach the performance level of a strong Bayesian baseline with F-scores of 89.80 (EN33K), 87.93 (EN5K), and 77.00 (MB5K). However, even if we only evaluate word segmentation performance, our neural approaches learn to segment and align, whereas this baseline only learns to segment. Section 3.5 will leverage those alignments for a type discovery task useful in language documentation.

The Pearson’s ρ correlation coefficients between ANE and boundary F-scores for all mono and bilingual runs of all corpora (N = 30) are −0.98 (RNN), −0.97 (CNN), and −0.66 (Transformer), with p-values smaller than 10⁻²⁶. These strong negative correlations confirm our hypothesis that lower ANEs correspond to sharper and better alignments.

### 3.4.2. Impact of Data Size and Quality

EN33K and EN5K results of Table 2 allow us to analyze the impact of data size on the S2S models. For the bilingual task, RNN performance drops by 7% on average, whereas performance drop is bigger for CNN (14-15%). Transformer performs poorly in both cases, and increasing data size from 5K to 33K seems to help only for a trivial task (see monolingual results).

The EN5K and MB5K results of Table 2 reflect the impact of language pairs on the S2S models. We know from that English should be easier to segment than Mboost, and this was confirmed by both dpseg and monolingual results. However, this trend is not confirmed in the bilingual task, where the quality of the (sentence aligned) parallel corpus seems to have more impact (higher boundary F-scores for MB5K than for EN5K for all S2S models). As shown in Table 2 MB-FR corpus has shorter sentences and smaller lexicon diversity, while EN-FR is made of automatically aligned books (noisy alignments), what may explain our experimental results.

### 3.5. Type Discovery in Low-Resource Settings

We investigate the use of Alignment ANE as a confidence measure. From the RNN models, we extract and rank the discovered types by their ANE, and examine if it can be used to separate true words in the discovered vocabulary from the rest. The results for low-resource scenarios (only 5K) in Table 3 suggest that low ANE corresponds to the portion of the discovered vocabulary the network is confident about, and these are, in most of the cases, true discovered lexical items (first row, P ≥ 70%). As we keep higher Alignment ANE values, we increase recall but loose precision. This suggests that, in a documentation scenario, ANE could be used as a confidence measure by a linguist to extract a list of types with higher precision, without having to pass through all the discovered vocabulary. Moreover, as exemplified for EN5K in Table 3 we also retrieve aligned information (translation candidates) for the generated lexicon.

### 4. Conclusions

We presented an empirical evaluation of different architectures (RNN, CNN and Transformer) with respect to their capacity to align word sequences in a source language with phoneme sequences in a target language, inferring from it word segmentation on the target side (UWS task). Although RNNs have been outperformed by CNN and Transformer-based models for machine translation, for UWS these architectures are still more robust in low-resource scenarios, and present the best segmentation results. We also introduced ANE, an intrinsic measure of alignment quality of S2S models. Accumulating it over the discovered alignments, we showed it can be used as a confidence measure to select true words, increasing Type F-scores.

---

**Table 2:** Boundary F-scores for the UWS task.

<table>
<thead>
<tr>
<th></th>
<th>Bilingual</th>
<th>Monolingual</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN 33K</td>
<td>RNN</td>
<td>77.10</td>
</tr>
<tr>
<td></td>
<td>CNN</td>
<td>71.30</td>
</tr>
<tr>
<td></td>
<td>Transformer</td>
<td>52.70</td>
</tr>
<tr>
<td>EN 5K</td>
<td>RNN</td>
<td>70.40</td>
</tr>
<tr>
<td></td>
<td>CNN</td>
<td>55.90</td>
</tr>
<tr>
<td></td>
<td>Transformer</td>
<td>52.50</td>
</tr>
<tr>
<td>MB 5K</td>
<td>RNN</td>
<td>74.00</td>
</tr>
<tr>
<td></td>
<td>CNN</td>
<td>68.20</td>
</tr>
<tr>
<td></td>
<td>Transformer</td>
<td>66.40</td>
</tr>
</tbody>
</table>

**Table 3:** Type retrieval results (RNN) using ANE for keeping most confident (type, translation) pairs (for ex: ANE=0.4 means we keep only pairs whose ANE is below 0.4).

<table>
<thead>
<tr>
<th>ANE</th>
<th>EN 5K</th>
<th>MB 5K</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>0.1</td>
<td>70.97</td>
<td>0.50</td>
</tr>
<tr>
<td>0.2</td>
<td>55.43</td>
<td>3.85</td>
</tr>
<tr>
<td>0.3</td>
<td>44.99</td>
<td>12.51</td>
</tr>
<tr>
<td>0.4</td>
<td>32.81</td>
<td>21.76</td>
</tr>
<tr>
<td>0.5</td>
<td>23.37</td>
<td>28.17</td>
</tr>
<tr>
<td>0.6</td>
<td>18.54</td>
<td>32.41</td>
</tr>
<tr>
<td>0.7</td>
<td>16.23</td>
<td>34.34</td>
</tr>
<tr>
<td>0.8</td>
<td>15.21</td>
<td>35.16</td>
</tr>
<tr>
<td>0.9</td>
<td>15.01</td>
<td>35.31</td>
</tr>
</tbody>
</table>

**Table 4:** Top 5 low and high ANE ranking for the discovered types (EN5K), with gold transcription and aligned information between parentheses (respectively). “INV” means incorrect type.

<table>
<thead>
<tr>
<th>ANE</th>
<th>EN 5K</th>
<th>MB 5K</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top Low ANE</td>
<td>High ANE</td>
</tr>
<tr>
<td></td>
<td>Type</td>
<td>Translation</td>
</tr>
<tr>
<td>0.1</td>
<td>SER1 (sir, EOS_token)</td>
<td>AH0 (a, convenable)</td>
</tr>
<tr>
<td>0.2</td>
<td>HHAH1SH (hush, chut)</td>
<td>IH1 (INV, ah)</td>
</tr>
<tr>
<td>0.3</td>
<td>FIHISHER0 (fisher, fisher)</td>
<td>D (INV, riant)</td>
</tr>
<tr>
<td>0.4</td>
<td>KLER1K (clerk, clerk)</td>
<td>N (INV, obit)</td>
</tr>
<tr>
<td>0.5</td>
<td>KIHY1 (kiss, embrasse)</td>
<td>YUW1 (you, diable)</td>
</tr>
</tbody>
</table>
5. References


