Optimizing Word Alignments with Better Subword Tokenization

Anh Khoa Ngo Ho, François Yvon

Optimizing Word Alignments with Better Subword Tokenization

Anh Khoa Ngo Ho anh-khoa.ngo-ho@limsi.fr
François Yvon francois.yvon@limsi.fr
Université Paris-Saclay, CNRS, LISN
Bât. 508, rue John von Neumann, Campus Universitaire, F-91405 Orsay

Abstract

Word alignments identify translational correspondences between words in a parallel sentence pair and are used, for example, to train statistical machine translation, learn bilingual dictionaries or to perform quality estimation. Subword tokenization has become a standard preprocessing step for a large number of applications, notably for state-of-the-art open vocabulary machine translation systems. In this paper, we thoroughly study how this preprocessing step interacts with the word alignment task and propose several tokenization strategies to obtain well-segmented parallel corpora. Using these new techniques, we were able to improve baseline word-based alignment models for six language pairs.

1 Introduction

Word alignment is a basic task in multilingual Natural Language Processing (NLP) and is used, for instance, to learn bilingual dictionaries, to train statistical machine translation (SMT) systems (Koehn, 2010), to filter out noise from translation memories (Pham et al., 2018) or in quality estimation applications (Specia et al., 2018). Word alignment can also serve to *explain MT decisions* (Stahlberg et al., 2018). Given pairs associating a sentence in a source language and a translation in a target language, word alignment aims to identify translational equivalences at the level of individual word tokens and has been initially approached with generative probabilistic models learning alignment in an unsupervised manner (Och and Ney, 2003; Tiedemann, 2011).

With rapid advances in neural based NLP, word alignment has recently regained some traction (Legrand et al., 2016) and improvements of the state of the art for multiple language pairs have been reported thanks to neuralized generative models (Alkhouli and Ney, 2017; Alkhouli et al., 2018; Ngo-Ho and Yvon, 2019), pre-trained multilingual embeddings (Jalili Sabet et al., 2020; Nagata et al., 2020; Dou and Neubig, 2021) or more powerful architectures based on the Transformer translation model of Vaswani et al. (2017), as reported for instance by Garg et al. (2019); Chen et al. (2020) and Chen et al. (2021).

In addition to using neural architectures, these new models differ from past approaches in that they compute alignments based on a decomposition into subword units (Sennrich et al., 2016; Kudo, 2018), which makes it possible to easily accommodate open-ended vocabularies and mitigate issues related to the alignment of unknown words, which has always been a challenge for discrete models. Another interesting property of subword units in the context of word alignment is that (a) they ease the generation of many-to-one / one-to-many links, which are difficult to handle in standard asymmetric models such as IBM-1 and IBM-4 (Liu et al., 2015; Tomeh et al., 2014; Wang and Lepage, 2016); (b) they also enable to actively manipulate the lengths of the

source and target sentences so as to make them more even, arguably a facilitating factor for alignment and translation models (Deguchi et al., 2020).

In this work, we take a closer look at the interaction between alignment and subword tokenization and try to address the following research questions: how much of the reported improvements in alignment performance can be linked to subword splitting? which issue(s) of basic alignment models do they mitigate? is it possible to design more active segmentation strategies that would target the alignment problem for specific language pairs? Our conclusions rests on the analysis of a systematic study of word alignment for 6 language pairs from multiple language families. We notably show that subword tokenization also help discrete alignment models. We also study techniques aimed at optimizing tokenization, which enable us to further improve the alignment accuracy and mitigate the problems cause by rare / unaligned words.

This paper is organized as follows: in § 2 we review the pitfalls of generative word alignment models, and analyse in § 3 how their performance vary with changing subword tokenizations. These analyses help to understand why such preprocessing actually improves word based models. Our main proposals are sketched in § 4, where we show how to optimize subword tokenization for better alignments. In § 5, we then briefly review related work, before concluding in § 6.

2 Pitfalls and limitations of word alignments models

In this section, we experiment with well-known word alignment packages (Fastalign (Dyer et al., 2013), Giza++ (Och and Ney, 2003), Eflomal (Östling and Tiedemann, 2016) as well as `Simalign` (Jalili Sabet et al., 2020)[1]), outlining difficult issues for word alignment models such as the prediction of null links, of many-to-one links, as well as the alignment of rare words. Detailed analyses are in (Ngo Ho, 2021). Asymmetric alignment models associate each source word with exactly one target word; such alignments are denoted as English $\rightarrow$ Foreign, when English is the source language. As a preamble, we start with our data condition.

2.1 Datasets

Our experiments consider multiple language pairs all having English on one side. Our training sets for French and German are made of sentences from Europarl (Koehn, 2005). For Romanian, we use both the NAACL 2003 corpus (Mihalcea and Pedersen, 2003) and the SETIMES corpus used in WMT'16 MT evaluation. For Czech, the parallel data from News Commentary V11 (Tiedemann, 2012) is considered, while we use the preprocessed parallel data for Vietnamese in IWSLT'15 (Luong and Manning, 2015) and the Japanese data from the KFTT (Neubig, 2011).

Our evaluations use standard test sets whenever applicable: for French and Romanian, we use data from the 2003 word alignment challenge (Mihalcea and Pedersen, 2003); the German test data is Europarl;[2] for Czech we use the corpus designed by Mareček (2016); the Japanese test data is from the KFTT and the test corpus for Vietnamese is generated from the EVBCorpus.[3] As is custom when evaluating unsupervised alignments, we append the test set to the training corpus at training time, meaning that there is no unknown word in the reference alignments.

Basic statistics for these corpora are in Table 1.[4] English-French and English-German training data ($\geq$ 1.5M) are much larger than the rest (from 122K to under 400K) and we take them as representative of a "large data" condition. Unsurprisingly, the vocabulary sizes of the German, Romanian and Czech corpora are substantially greater than the corresponding English,

[1] A method of generating alignment links based on the matrix of embedding similarities without parallel data. The options are to use mBert (Devlin et al., 2019) or the multilingual version of Fasttext are used to generate multilingual embeddings from monolingual data. In our experiments, we use the setting: mBert + Argmax.

[2] http://www-i6.informatik.rwth-aachen.de/goldAlignment/

[3] https://code.google.com/archive/p/evbcorpus/

[4] We only use training sentences of length lower than 50.

which contains a smaller number of inflected variants. The opposite pattern is observed for Japanese and Vietnamese, two synthetic languages with less inflectional variability than English.

Corpus	Training data					Test data			
	# sent.	word vocab.		char. vocab.		# sent.	# words		# non-null
	pairs	Eng.	For.	Eng.	For.	pairs	Eng.	For.	links
En-Fr	∼1.7M	∼106K	∼112K	111	115	447	7 020	7 761	17 438
En-Ge	∼1.5M	∼96K	∼311K	218	235	509	10 413	9 945	10 533
En-Ro	∼250K	∼74K	∼115K	124	131	246	5 455	5 315	5 991
En-Cz	∼182K	∼62K	∼147K	246	157	2 501	59 724	52 881	67 423
En-Ja	∼377K	∼156K	∼126K	∼2K	∼5K	1 235	30 822	34 403	33 377
En-Vi	∼122K	∼42K	∼19K	133	171	3 447	70 049	94 753	81 748

Table 1: Basic statistics for the training data and test data

2.2 Evaluation protocol

We use the alignment error rate (AER) (Och, 2003), F-score (F1), precision and recall as measures of performance. AER is based on a comparison of predicted alignment links (A) with a human reference including sure (S) and possible (P) links, and is defined as an average of the recall and precision taking into account the sets P and S. AER is defined as:

$$AER = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|} \tag{1}$$

where A is the set of predicted alignments. Note that the English-Romanian, English-Japanese and English-Vietnamese reference data only contain "sure" links, meaning that for these languages pairs, AER and F-measure are deterministically related.

2.3 Main observations

Detailed analyses of automatic word alignments, fully documented in (Ngo Ho, 2021), show that:

- Unaligned words are poorly predicted: we collect correctly/incorrectly unaligned words on the source side for the asymmetrical models. For English→ Czech, there are too few English words aligning with Czech words for `IBM-1` whereas `IBM-4` produces too many unaligned English words (Figure 1).
- Many-to-one/one-to-many links are also poorly predicted, even with symmetrization.[5] This can be seen in Figure 2.
- Larger length differences between parallel sentences yield more errors, as shown in Figure 3. This again hints at the tendency of discrete word models to generate one-to-one alignments.

3 Studying the interaction between alignment and segmentation

3.1 Implementation

In this section, we restrict our analysis to `Fastalign` and `Eflomal` and study how their performance vary when the subword vocabulary changes. We perform the alignment between

[5]We heuristically merge two alignments with opposite directions to produce a symmetric alignment, by using the grow-diag-final (GDF) heuristic proposed in Koehn (2005).

Figure 1: Number of correctly/incorrectly unaligned English and Czech words for English→Czech (left) and Czech→English (right).

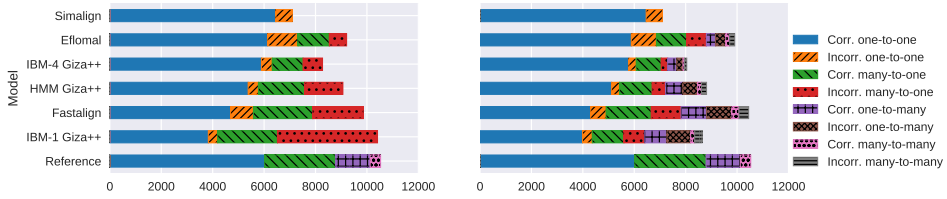

Figure 2: Alignment types for asymmetrical alignments for English→German (left) and symmetrical alignments using Grow-diag-final (right).

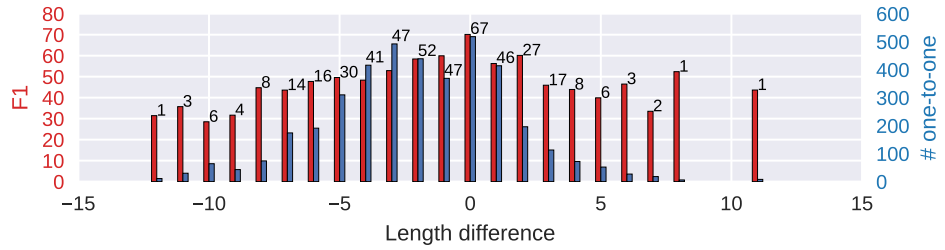

Figure 3: F-score (red) and number of correct one-to-one alignments (blue) as a function of a length difference for the direction English-French, computed by `Fastalign`. The numbers in black are the corresponding number of sentences.

subword units generated by Byte-Pair-Encoding (Sennrich et al., 2016) and the unigram method of (Kudo, 2018), both implemented with the SentencePiece package (Kudo and Richardson, 2018). All parameters of these models are set to their default values. We independently segment sentences in each language with varying vocabulary sizes $V \in \{2K, 4K, 8K, 16K, 32K, 48K\}$. For Japanese, we do not use the vocabulary size of 2K because it is smaller than the character-based vocabulary size. For English-Vietnamese, experiments for English vocabulary size of 48K and Vietnamese vocabulary size larger than 32K were not performed. This is because they would imply larger vocabularies than their word-based counterparts. When using the sampling strategy of SentencePiece, we use $\alpha = 0.1$.

Our results and analyses are however based on *word-level alignments*. Subword-level alignments are thus converted into word-level alignments as follows: a link between a source and a target word exists if there is at least one alignment link between their any of their subwords.

English+ Model	French		German		Romanian		Czech		Japanese		Vietnamese	
	En-Fr	Fr-En	En-De	De-En	En-Ro	Ro-En	En-Cz	Cz-En	En-Ja	Ja-En	En-Vi	Vi-En
`Fastalign`												
Word	15.1	**16.2**	28.9	31.2	33.3	**32.9**	25.7	25.3	50.6	49.3	48.8	32.8
BPE	**14.7** (32K-32K)	16.3 (8K-8K)	**26.7** (4K-32K)	**29.3** (16K, 16K)	**31.4** (16K-8K)	35.0 (16K-2K)	**24.6** (16K-32K)	**24.3** (32K-16K)	**47.5** (8K-8K)	**46.9** (8K-16K)	**45.7** (4K-4K)	**29.5** (4K-8K)
Unigram	18.6 (45K-16K)	20.1 (48K-32K)	31.3 (4K-48K)	33.2 (16K-16K)	36.6 (39K-16K)	40.0 (32K-4K)	30.5 (16K-32K)	31.4 (48K-16K)	49.7 (8K-8K)	48.0 (8K-32K)	49.3 (16K-2K)	35.3 (4K-8K)
`Eflomal`												
Word	8.0	8.7	22.8	24.8	26.3	25.4	14.1	13.4	46.5	46.7	44.1	27.6
BPE	**6.1** (16K-32K)	**7.7** (32K-16K)	**20.7** (4K-32K)	**21.7** (32K-16K)	**24.4** (16K-48K)	**24.5** (8K-48K)	**12.5** (8K-32K)	**11.9** (48K-16K)	**42.5** (8K-32K)	**41.7** (8K-32K)	**36.1** (2K-8K)	**24.9** (2K-32K)
Unigram	11.3 (45K-48K)	14.4 (32K-32K)	23.9 (32K-32K)	26.7 (48K-32K)	26.9 (32K-48K)	28.7 (48K-16K)	17.5 (32K-32K)	17.5 (48K-16K)	45.3 (16K-8K)	42.7 (30K-16K)	43.5 (16K-8K)	29.7 (2K-16K)

Table 2: AER scores of subword-based models and word-based models. We only report the best result obtained by subword-based models, and the corresponding vocabulary sizes.

3.2 Main results

In order to observe how the alignment accuracy varies with the size of the subword vocabulary, we plot precision and recall as a function of the target vocabulary size for each source vocabulary size. As can be seen in Figure 4, having short units (top-left zones) on both sides yields a better recall but a much worse precision. The opposite trend is found in bottom-right zones where we approach word-based models. Note that however with a proper choice of unit size, BPE-based models are able to outperform their word-based counterparts, with a gain of about 2 AER points. This improvement is not clear for unigram-based models (see Table 2).

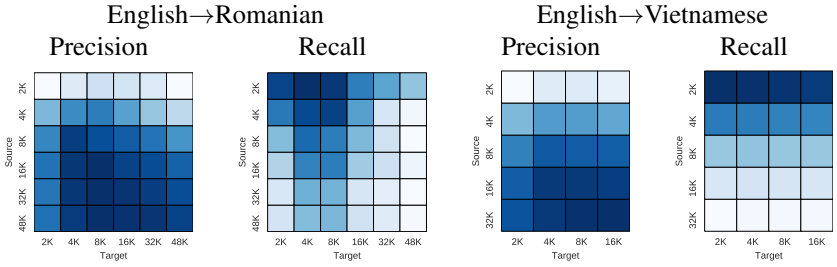

Figure 4: Precision and recall of BPE-based alignments for English→Romanian and English→Vietnamese, computed by `Fastalign`. The darker the cell, the greater the score.

3.3 Complementary analyses

3.3.1 Unaligned words and alignment types

Figure 5 displays unaligned word patterns generated by several BPE-based models for English-German. Choosing small inventories on the target side yields more fragmented sequences and a reduced number of non-aligned words in the source, as is expected for asymmetrical models. Significantly increasing both recall and precision proves difficult, and we only observe small improvements with respect to the word-based baselines: for instance, with `Fastalign`, the best BPE-model (4K-32K) removes 40 incorrectly unaligned words and finds 10 correctly unaligned words. Compared with `HMM` or `IBM-4`, we also notice that BPE-based models are less prone to over-generate null links. Similar trends were observed for the other language pairs/directions.

We now study how the number of links for each alignment type changes with the vocabulary size (Figure 6). The most noticeable observation is that shorter BPE units (e.g., 2K-2K) generate less one-to-one links and accordingly more of the other alignment types, especially one-to-many

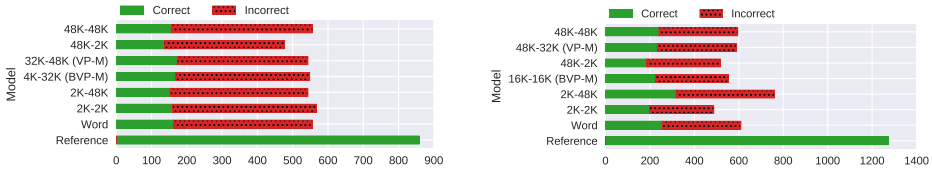

Figure 5: Number of correctly/incorrectly unaligned English (left) and German (right) words generated by `Fastalign` for respectively English→German and German→English. VP-M denotes the vocabulary pair for which the average length difference between source and target sentences is smallest; BVP-M denotes the vocabulary pair yielding the best AER.

and many-to-many links. In other words, tokens that decompose into a sequence of shorter units in the source side have more chance to align with several target tokens. However, this does not result in an increased number of correct one-to-many/many-to-many links. Similar trends were observed for the other language pairs/directions.

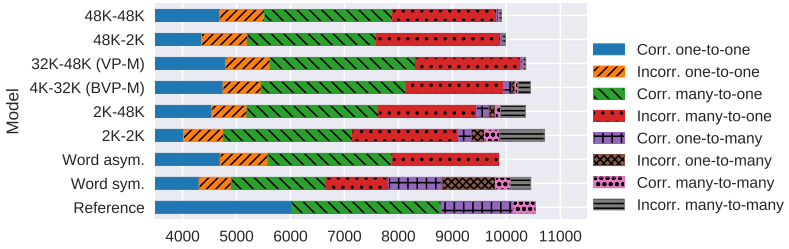

Figure 6: Alignment errors for BPE-based, word-based asymmetrical (Word asym.) and symmetrical alignments (Word sym.) computed by `Fastalign` for English→German.

3.3.2 Aligning rare words

Using subwords affects the overall distribution of units and helps mitigate issues with rare tokens. To measure this effect, we collect rare source words (a word is rare if it occurs once in our training data) and plot their F-scores as a function of target and source vocabulary sizes (see Figure 7). Recall that German has a very large word-based vocabulary size (Table 1). Accordingly, for the German-English direction, we can see a large gain (about +8 points) in F-score when using a reduced German vocabulary size of 32K.

3.4 Improving alignment by voting

As a final experiment, we combine multiple BPE-based alignments using a simple voting procedure.This method is parameterized by the required level of agreement (the percentage of models agreeing on an alignment link). Figure 8 shows that considering the BPE models described above and using an agreement level of 70% improves the F-score by almost 2 points for German→English and Japanese→English. Similar results are obtained for the other language pairs, showing that considering multiple segmentations in alignment can be helpful.

4 Optimizing subword tokenization

In this section, we build on the intuition that pairs of sentences which differ in length are difficult to align (Deguchi et al., 2020), suggesting that subword splitting should be used to make the

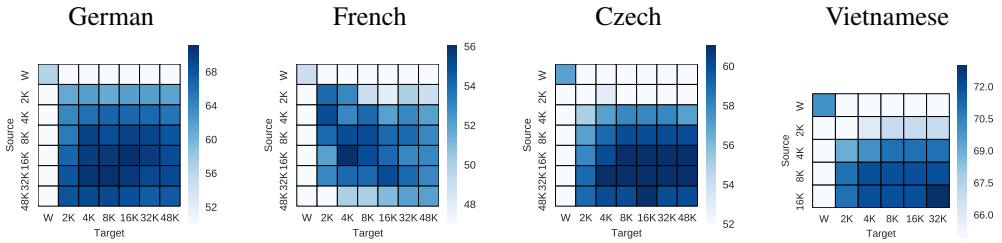

Figure 7: F-scores obtained with `Fastalign` as a function of source and target vocabulary sizes for rare source words in German, French, Czech and Vietnamese, when translating into English. The word-based vocabulary size is denoted W.

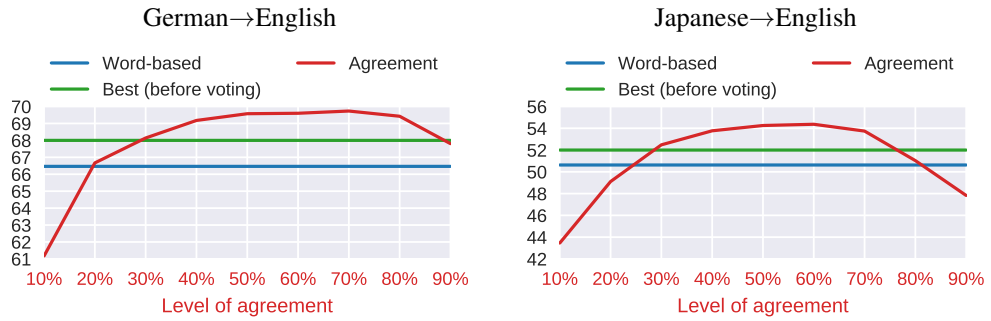

Figure 8: F-score for word-based model (blue line) and for the best BPE-based model (green line). The red curve plots the F-score for each level of agreement for German→English and Japanese→English. For both directions, voting improves the F-score of about 2 pts with a 70% level of agreement.

length of parallel sentences more even. We study global and local ways to achieve this goal.

4.1 Global methods for controlling length differences

We first consider two ways to find the vocabulary pair minimizing the average length difference:

- the first one (denoted VP-M) simply picks the vocabulary pair that minimizes this value in the matrix of all vocabulary pairs;
- this solution can be improved by using the following greedy search procedure (VP-GS): we compute the average sequence length difference for a vocabulary pair based on a pre-defined search space radius. If we find a new vocabulary pair producing a smaller average than the current pair, we continue to explore the neighbors of this new pair. We reduce the search space radius ε in the case that no new pair is found.[6] Details are in algorithm 1.[7]

We collect the average F-score, length difference and English vocabulary size for all language pairs and directions (see Table 3). For BPE-based models, minimizing length difference between the source and target sentence outperforms word-based models with a gain of at least 1 point in F-score. This performance is close to the best results found from the matrix of vocabulary pair. Unigram-based models fail to match such performance, but we still observe an

[6] The step size ρ remains the same for the whole procedure.

[7] $f(\alpha, \beta)$ returns the average sequence length difference obtained with vocabularies of size α and β.

Method		F-score		Length difference		# English voc.	
		Fastalign	Eflomal	Fastalign	Eflomal	Fastalign	Eflomal
Word		58.7	64.0	4.23		72K	
BPE	BVP-M	**60.4**	**66.3**	5.0	5.5	9K	16.5K
	VP-M	59.6	66.1	3.56		26K	
	VP-GS	59.8	65.5	3.51		∼21K	
Unigram	BVP-M	57.1	63.8	5.7	5.5	20K	21.6K
	VP-M	56.2	62.9	4.8		17K	
	VP-GS	**58.4**	**64.4**	4.5		∼18K	

Table 3: Average F-score (over language pairs and directions) for global methods of controlling sequence length difference for Fastalign and Eflomal. We also report the best vocabulary pair found in the vocabulary pair matrix (BVP-M).

improvement for the greedy search, which outperforms the word-based models for Eflomal for English-French, English-German, English-Japanese and English-Vietnamese.

Algorithm 1 Finding the vocabulary pair minimizing the average length differences

Require:
α: Source side vocabulary size; β: Target side vocabulary size
ε: search space radius (default = 2000);
ρ: step size (default = 100);
Ensure: $1000 \leq \alpha, \beta \leq 50000$
while $\varepsilon \geq 100$ **do**
for $\nu \in \{\alpha - \varepsilon, \alpha, \alpha + \varepsilon\}$, $\mu \in \{\beta - \varepsilon, \beta, \beta + \varepsilon\}$ **do**
if $f(\nu, \mu) < f(\alpha, \beta)$ **then**
$\alpha = \nu$; $\beta = \mu$; $\epsilon = 2000$
end if
end for
if α and β remain the same **then**
$\varepsilon = \varepsilon - \rho$
end if
end while

4.2 Local methods for controlling the length difference

The methods presented above consider ways to optimize the length difference at the corpus level, using one subword vocabulary that is used across the board. We study here four *local* methods that aim to reduce the length differences *separately for each sentence pair* before training the alignment procedure. With the exception of the first method, they all rely on the unigram algorithm, and use a fixed, predefined, vocabulary size for both languages:

- the first (SP-M) simply picks, among all the considered segmentations of each sentence, the one that minimizes the length difference. When there is more than one minimal segmentation, we select the one for which total source and target lengths is smallest.
- the second[8] (SM1-1VP) relies on the idea of Deguchi et al. (2020): (a) we collect the 10 most likely segmentations for each language using the unigram algorithm; (b) we select the highest probability candidate on both sides, and consider the longer of the two as the

[8]This method and next only apply to unigram, which, contrarily to BPE, is based on a sound probabilistic model.

anchor segmentation; (c) we pair this segmentation with the one, in the other language, that is closest in length and maximally likely. We also consider the case SM5-1VP where we include the top five highest probability in the last step for the training data.

- SSM5-1VP extends the previous idea with more candidates: we sample 10 segmentations using the unigram algorithm for each language, then select the 5 pairs of segmentations that have the smallest length difference, and use it as the training data for the word alignment.
- a last idea (SSM5-GS) uses the same strategy as SSM5-1VP, using the "optimal" pair of vocabulary sizes computed by the greedy search algorithm (Algorithm 1).

We always consider one single pair of segmentations for the test data: we chose the highest probability pair for SM5-1VP and one pair producing the smallest length difference for SSM5-*.

For BPE-based models (Figure 9), SP-M only outperforms the word-based model for English-French and English-Vietnamese, and fails to achieve better F-scores than the two global methods. The performance of unigram-based method (assuming vocabularies of sizes 16K-16K) is displayed in Figure 10. They all outperform the baseline (a fixed 16K-16K model) and also the word-based models for French, Japanese and Vietnamese. It also seems that including several segmentation samples for each sentence pair in the training data (as in SSM5-1VP) also helps to improve the performance, resulting in a simple scheme based only on length differences, that consistently outperforms all other unigram-based methods. These results open perspectives for further improving these models, especially for German, Czech and Romanian, for which the 16K-16K setting might be suboptimal. The last method (SSM5-GS) does not succeed in improving SSM5-1VP. Similar observations hold for Eflomal, albeit with better baselines.

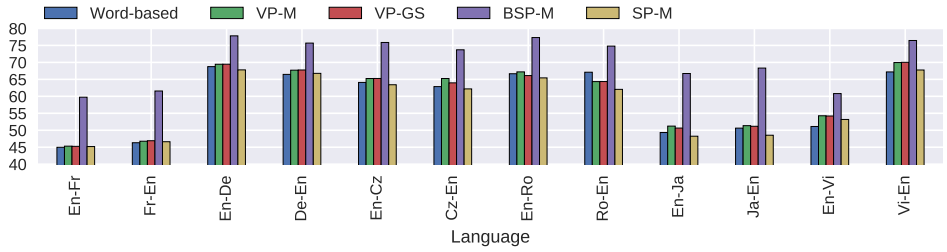

Figure 9: F-scores for BPE-based segmentations. We compare global methods (VP-M and VP-GS) with SP-M and also display scores obtained with best segmentation for each sentence pair (BSP-M), which provides us with an *oracle value*. Alignments are computed by `Fastalign`.

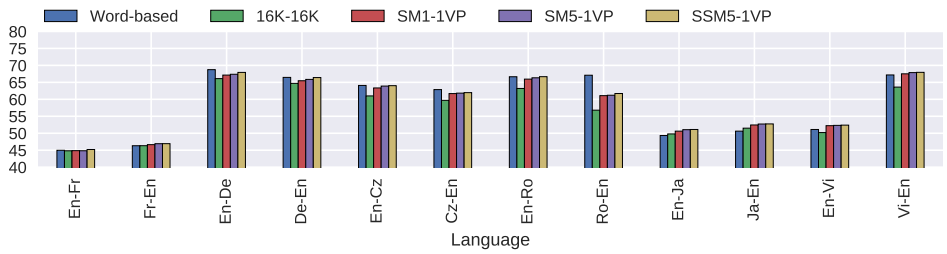

Figure 10: F-scores for unigram-based local strategies; alignments computed by `Fastalign`.

5 Related work

Subword segmentation is introduced in the context of neural translation in (Sennrich et al., 2016), using a reimplementation[9] of the Byte Pair Encoding algorithm of Gage (1994). BPE is a greedy, bottom up algorithm that recursively aggregates frequent bigrams into new symbols, and is thoroughly analyzed in (Gallé, 2019). The main alternative is SentencePiece introduced in (Kudo, 2018; Kudo and Richardson, 2018), which implements a form of variable-length probabilistic unigram model, which can be traced back to (Deligne and Bimbot, 1995).

With BPE/unigram subword tokenization becoming a standard for many applications, several studies have started to investigate more closely the impact on these preprocessing decisions on the final performance. The implementation of SentencePiece[10] reports a large number of MT experiments aimed to compare BPE and unigram in multiple conditions, concluding that both yield comparable BLEU scores across the board when used with a fixed tokenization in words.

The shortcomings of BPE/unigram segmentations have been the subject of several studies, reporting comparisons with (a) linguistic segmentations (Huck et al., 2017; Ataman et al., 2017; Banerjee and Bhattacharyya, 2018; Weller-Di Marco and Fraser, 2020) and (b) alternative preprocessing schemes such as character-based models (eg. in Sennrich (2017); Sajjad et al. (2017); Cherry et al. (2018)). Ding et al. (2019) conduct a systematic exploration considering a large numbers of vocabulary sizes to better understand its impact on NMT performance, comparing several NMT architectures such as shallow/deep-transformer, tiny/shallow/deep-LSTM. Bostrom and Durrett (2020) evaluate the impact of tokenization on language model pre-training. They conclude that tokenization encodes a surprising amount of inductive bias and that LM-based tokenization produces subword units that qualitatively align with morphology much better than those produced by BPE, suggesting that the latter is better than the former for pretrained models.

The work of Deguchi et al. (2020) is our main inspiration, and explore ways to optimize the subword segmentation, using, as we do, sampling techniques and length-based heuristics to chose the most appropriate target for each source, and observing gains in translation performance.

6 Conclusion and outlook

In this work, we have studied the interaction between word alignment and word segmentation based on two algorithms (BPE and unigram) and multiple word aligners. Using smaller units notably mitigate issues with rare/unknown words; shorter units also help to retrieve more correct links for non-canonical (one-to-many, many-to-one) alignment links. Based on these observations, we have thoroughly analyzed the variation of alignment scores with respect to vocabulary sizes, showing that the word-based segmentation was less than optimal. We have finally explored various ways to actively optimize the subword tokenization; promising results in this direction have been obtained with the unigram algorithm, owing to its ability to generate multiple high-probability segmentations. We have notably found that adjusting length differences in source and target was a reasonable heuristic to progress towards better joint tokenizations, even though (a) the relationship between length difference and alignment quality was not as clear as one may have wished; (b) inconsistencies have been observed between unigram and BPE. In the future, we will continue to explore inexpensive ways to identify promising joint segmentations and improve the alignment between subword units.

Acknowledgements

This work has been made possible thanks to the Saclay-IA computing platform.

[9] https://github.com/rsennrich/subword-nmt

[10] https://github.com/google/sentencepiece/blob/master/doc/experiments.md

References

Alkhouli, T., Bretschner, G., and Ney, H. (2018). On the alignment problem in multi-head attention-based neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 177–185, Brussels, Belgium. Association for Computational Linguistics.

Alkhouli, T. and Ney, H. (2017). Biasing attention-based recurrent neural networks using external alignment information. In *Proceedings of the Second Conference on Machine Translation*, pages 108–117, Copenhagen, Denmark. Association for Computational Linguistics.

Ataman, D., Negri, M., Turchi, M., and Federico, M. (01 Jun. 2017). Linguistically motivated vocabulary reduction for neural machine translation from turkish to english. *The Prague Bulletin of Mathematical Linguistics*, 108(1):331 – 342.

Banerjee, T. and Bhattacharyya, P. (2018). Meaningless yet meaningful: Morphology grounded subword-level NMT. In *Proceedings of the Second Workshop on Subword/Character LEvel Models*, pages 55–60, New Orleans. Association for Computational Linguistics.

Bostrom, K. and Durrett, G. (2020). Byte pair encoding is suboptimal for language model pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4617–4624, Online. Association for Computational Linguistics.

Chen, C., Sun, M., and Liu, Y. (2021). Mask-align: Self-supervised neural word alignment. *CoRR*, abs/2012.07162.

Chen, Y., Liu, Y., Chen, G., Jiang, X., and Liu, Q. (2020). Accurate word alignment induction from neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 566–576, Online. Association for Computational Linguistics.

Cherry, C., Foster, G., Bapna, A., Firat, O., and Macherey, W. (2018). Revisiting character-based neural machine translation with capacity and compression. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, EMNLP'18, pages 4295–4305, Brussels, Belgium. Association for Computational Linguistics.

Deguchi, H., Utiyama, M., Tamura, A., Ninomiya, T., and Sumita, E. (2020). Bilingual subword segmentation for neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4287–4297, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Deligne, S. and Bimbot, F. (1995). Language modeling by variable length sequences: Theoretical formulation and evaluation of multigrams. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 1, pages 169–172.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Ding, S., Renduchintala, A., and Duh, K. (2019). A call for prudent choice of subword merge operations in neural machine translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 204–213, Dublin, Ireland. European Association for Machine Translation.

Dou, Z.-Y. and Neubig, G. (2021). Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. Association for Computational Linguistics.

Dyer, C., Chahuneau, V., and Smith, N. A. (2013). A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Gage, P. (1994). A new algorithm for data compression. *Computer Users Journal*, 12(2):23–38.

Gallé, M. (2019). Investigating the effectiveness of BPE: The power of shorter sequences. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1375–1381, Hong Kong, China. Association for Computational Linguistics.

Garg, S., Peitz, S., Nallasamy, U., and Paulik, M. (2019). Jointly learning to align and translate with transformer models. In *Proc. IJCNLP-EMNLP*, Hong Kong, China.

Huck, M., Riess, S., and Fraser, A. (2017). Target-side word segmentation strategies for neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 56–67, Copenhagen, Denmark. Association for Computational Linguistics.

Jalili Sabet, M., Dufter, P., Yvon, F., and Schütze, H. (2020). SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Koehn, P. (2010). *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA, 1st edition.

Kudo, T. (2018). Subword regularization: Improving neural network translation models with multiple subword candidates. In Gurevych, I. and Miyao, Y., editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 66–75. Association for Computational Linguistics.

Kudo, T. and Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Legrand, J., Auli, M., and Collobert, R. (2016). Neural network-based word alignment through score aggregation. In *Proceedings of the First Conference on Machine Translation*, pages 66–73, Berlin, Germany. Association for Computational Linguistics.

Liu, C., Liu, Y., Sun, M., Luan, H., and Yu, H. (2015). Generalized agreement for bidirectional word alignment. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1828–1836, Lisbon, Portugal. Association for Computational Linguistics.

Luong, M.-T. and Manning, C. D. (2015). Stanford neural machine translation systems for spoken language domain. In *International Workshop on Spoken Language Translation*, Da Nang, Vietnam.

Mareček, D. (2016). Czech-English manual word alignment. Technical report, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Mihalcea, R. and Pedersen, T. (2003). An evaluation exercise for word alignment. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond - Volume 3*, HLT-NAACL-PARALLEL '03, pages 1–10, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nagata, M., Chousa, K., and Nishino, M. (2020). A supervised word alignment method based on cross-language span prediction using multilingual BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 555–565, Online. Association for Computational Linguistics.

Neubig, G. (2011). The Kyoto free translation task. http://www.phontron.com/kftt.

Ngo Ho, A. K. (2021). *Generative Probabilistic Alignment Models for Words and Subwords : a Systematic Exploration of the Limits and Potentials of Neural Parametrizations*. PhD thesis, Université Paris-Saclay.

Ngo-Ho, A.-K. and Yvon, F. (2019). Neural Baselines for Word Alignments. In *Proc. IWSLT*, Hong-Kong, China.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 160–167, Stroudsburg, PA, USA. Association for Computational Linguistics.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Comput. Linguistics*, 29(1):19–51.

Östling, R. and Tiedemann, J. (2016). Efficient word alignment with Markov Chain Monte Carlo. *Prague Bulletin of Mathematical Linguistics*, 106:125–146.

Pham, M. Q., Crego, J., Senellart, J., and Yvon, F. (2018). Fixing translation divergences in parallel corpora for neural MT. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2967–2973, Brussels, Belgium.

Sajjad, H., Dalvi, F., Durrani, N., Abdelali, A., Belinkov, Y., and Vogel, S. (2017). Challenging language-dependent segmentation for Arabic: An application to machine translation and part-of-speech tagging. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 601–607, Vancouver, Canada. Association for Computational Linguistics.

Sennrich, R. (2017). How grammatical is character-level neural machine translation? assessing MT quality with contrastive translation pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 376–382. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016). 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)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Specia, L., Scarton, C., Paetzold, G. H., and Hirst, G. (2018). *Quality Estimation for Machine Translation*. Morgan & Claypool Publishers.

Stahlberg, F., Saunders, D., and Byrne, B. (2018). An operation sequence model for explainable neural machine translation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 175–186, Brussels, Belgium. Association for Computational Linguistics.

Tiedemann, J. (2011). *Bitext Alignment*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In Chair), N. C. C., Choukri, K., Declerck, T., Dogan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Tomeh, N., Allauzen, A., and Yvon, F. (2014). Maximum-entropy word alignment and posterior-based phrase extraction for machine translation. *Machine Translation*, 28(1):19–56.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

Wang, H. and Lepage, Y. (2016). Yet another symmetrical and real-time word alignment method: Hierarchical sub-sentential alignment using f-measure. In Park, J. C. and Chung, J., editors, *Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation, PACLIC 30, Seoul, Korea, October 28 - October 30, 2016*. ACL.

Weller-Di Marco, M. and Fraser, A. (2020). Modeling word formation in English–German neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4227–4232. Association for Computational Linguistics.