

Deeper than Words: Morph-based Alignment for Statistical Machine Translation

Mark Fishel

University of Tartu

Tartu, Estonia

fishel@ut.ee

Abstract

In this paper we introduce a novel approach to alignment for statistical machine translation. The core idea is to align subword units, or morphs, instead of word forms. This results in a joint segmentation and alignment model, aimed to improve translation quality for morphologically rich languages and reduce the size of the required parallel corpora. Here we focus on translating from inflectional languages into languages with simpler morphology, thus segmenting only the input. Using the approach as a preprocessing step produces results below the baseline. On the other hand, emulating the performance of the new model in a joint translation system shows possible potential.

1 Introduction

Word- and phrase-based statistical machine translation ignores possible morphological relatedness of the words. This is more of a problem for inflectional languages – the richer their morphology, the larger the training corpus has to be to cover most of the possible word forms.

A substantial amount of work has been done to account for this problem. In most cases morphological analysis is used to segment the words or otherwise augment the text with morphological information. Also recently an alternative approach of using unsupervised morphology for the same task has been introduced. The problem with all previous work is that all preprocessing is language-specific. The recent advances no longer depend on linguistic tools, but still deduce segmentations that are language-specific, ignoring the bilingual nature of the task at hand.

In our approach deduction of morphology is integrated with SMT training. Instead of searching

for an alignment between words, here we try to align subword units, or morphs. This results in a special case of morphology for the parallel corpus at hand. The approach is bilingual and uses no additional linguistic processing tools, like morphological analysis, or any other resources; the only data used for learning is the parallel corpus itself, which is required by machine translation.

This paper focuses on a one-sided approach, where the morphs of one language are aligned to words of the other one. Such setup is suitable for language pairs where one language is highly inflectional (like Turkish or Finnish) and the other one is not (like English or Chinese). Thus the less inflectional language serves as a guideline for segmenting the other language into morphs.

The approach possesses all the advantages that morphological treatment of words gives to machine translation (reduced out-of-vocabulary rate, smaller corpora required). In comparison with segmenting words with unsupervised morphology here the situation can be modelled more precisely, as the words are only split into morphs where both languages indicate it. Compared to morphological analysis, this can also be more efficient if we assume that linguistic morphology is not necessarily the best way of decomposing words.

The paper is organized as follows. Previous work on morphological improvements for statistical machine translation is briefly reviewed in section 2. The introduced method, which includes the alignment model, its learning and applying are described in section 3. The model is evaluated in context of MT on four language pairs (with English as the target language) in section 4. Section 5 discusses the evaluation results and the model. The paper is concluded in section 6.

2 Related Work

As mentioned in the introduction, a lot of work has been done on using morphological analysis to

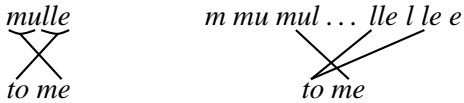


Figure 1: Example of morph-to-word alignment. The Estonian word *mulle* means *to me*. An intuitive alignment is shown on the left. Segmenting the source word results are shown on the right together with an enabled invalid alignment with intersecting morphs.

improve statistical machine translation. In many cases this is done similarly to the present work by segmenting the word forms into morphemes, or just the lemma and ending, or stem and morphological inflection identifier: (Nießen and Ney, 2004), (Badr et al., 2008), (Lee, 2004).

An alternative is unsupervised morphology. In (Virpioja et al., 2007) and (Sereewattana, 2003) both the source and the target language are fully segmented into morphs as a preprocessing step. In (Bojar et al., 2008) the target language is segmented into stems and suffixes. Kirik and Fishel (2008) segment the source language; in addition to both previous segmentation schemes splitting the compounds into simpler parts is also used.

Unlike all previous approaches and similarly to the present work Snyder and Barzilay (2008) describe an unsupervised model of deducing segmentation from parallel corpora.

3 Methodology

Throughout this paper we will assume that the source language is the highly inflectional one and the target language – is not. The method can be easily reversed for the opposite case. Thus the core idea of the introduced approach is to align the source word parts to the target words.

The task can be reduced to standard word alignment learning techniques, like the IBM models (Brown et al., 1993), by replacing each source language word form with all of its substrings. However here the alignment search space is constrained, unlike the word to word case: the selected morphs cannot intersect and have to cover all of the word forms.

Consider the example on figure 1. The original phrase pair can be intuitively segmented and aligned as shown on the left. However after replacing the source word form with a sequence of

its substrings, invalid alignments are possible, like the one on the right.

The main problem with this constraint is that it disallows summing over all alignments efficiently. Expectation-maximization learning of the lexical probabilities considers the whole distribution of alignments, not just the most probable one, which requires summing over all alignments. In case of the IBM models the alignment of an individual word does not depend on the rest of the alignment. On the contrary, here each morph also depends on what other morphs are selected in the alignment. In case a morph makes the alignment invalid, it has to have a probability of 0; in the former example if an alignment already has entries for *mul* and *le* then $p(f|e)$ must be 0 for any f and e .

Here we approach the problem by estimating the alignment distribution in an unconstrained search space (so that all alignments, including the invalid ones, are included) and only adding the constraint in the search phase; this results in the most probable valid alignment. Thus the estimation task becomes identical with the IBM models, after having the source word forms replaced with the sequence of all their substrings: iteratively update the lexical probabilities with an EM-algorithm, maximizing the log-probability of the data and the posterior probability of the alignments at the same time.

This however causes a different problem to arise, which concerns the lexical probabilities to be estimated. Having no notion of “meaningful” and “non-meaningful” morphs, short substrings of 1-2 letters will have a high occurrence rate, especially with frequent target words (like *the*). As a result those short and frequent morphs dominate in the distribution $p(f|e)$ in case of frequent e ’s, and probability mass of the meaningful f ’s will be pushed down to 0.

On the other hand, if the direction were reversed to $p(e|f)$, absolutely all morphs will stay in the distribution. This means a largely increased parameter space, where both the scarce meaningful morph pairs and the vast amount of all other morphs are all present.

To account for this second problem we employ the idea of joint parameter estimation from (Liang et al., 2006): a lexical pair is considered to be good only if both the source is translated as the target and the target is translated as the source. This is modelled by jointly maximizing the alignment

probabilities: $p_{e|f}(\mathbf{a}|\mathbf{e},\mathbf{f}) \times p_{f|e}(\mathbf{a}|\mathbf{e},\mathbf{f})$. However we do not estimate a symmetrical alignment, since a simple IBM-model style vector alignment satisfies the requirement of aligning the source morphs to the target words.

3.1 Joint Learning for an Asymmetric Alignment

Following (Liang et al., 2006) the objective function to maximize is

$$\sum_{(\mathbf{e},\mathbf{f})} \sum_{\mathbf{a}} p_{f|e}(\mathbf{a}|\mathbf{e},\mathbf{f}) p_{e|f}(\mathbf{a}|\mathbf{e},\mathbf{f}) \log(p(\mathbf{e},\mathbf{a}|\mathbf{f})p(\mathbf{f},\mathbf{a}|\mathbf{e})).$$

Also the parameters have to be valid conditional probabilities, i.e.

$$\forall f : \sum_e p(e|f) = 1, \forall e : \sum_f p(f|e) = 1.$$

Maximization under these constraints is easily achieved with a set of Lagrange’s multipliers, $\{\lambda_f\} \cup \{\lambda_e\}$, since the objective function is easily differentiated.

3.2 Searching for the Most Probable Alignment

The aim of the search is to find an alignment \mathbf{a} for a sentence pair (\mathbf{e},\mathbf{f}) with a maximum joint probability:

$$\hat{\mathbf{a}} = \arg \max_{\mathbf{a}} p(\mathbf{f},\mathbf{a}|\mathbf{e})p(\mathbf{e},\mathbf{a}|\mathbf{f}),$$

and an additional requirement that the alignment has to be valid (i.e. the morphs cannot intersect and have to cover the whole source sentence). The same procedure for word-based alignments is easily solved by maximizing the corresponding single alignment units independently; in our case this is not possible due to the validity requirement, which means that morph probabilities depend on which other morphs have already been included from the same word.

In order to find the most probable alignment for a source word form we adapt the forward-backward algorithm to our case. In this algorithm the search is done in two passes – the forward and the backward pass. The aim of the forward pass is to consider each possible alignment without ruling out any based on partial likelihood. In the backward pass the complete likelihood is used to select the most probable alignment.

Consider the example in figure 2, with the Estonian-English pair *mulle : to me*. The model

includes a morph *mull* which matches the beginning of the Estonian word and has a high probability. Even if the English phrase included the word *bubble* corresponding to the Estonian morph *mull*, selecting just that morph would be a mistake, since the model does not have a morph to cover the ending of the Estonian word, thus it would leave no space for completing the alignment.

Since the number of possible segmentations is exponential, a brute force solution of listing all of them would be inefficient. In the forward-backward algorithm it is done via optimizing the search by grouping partial alignments that have a common ending point. For instance, if the English sentence in the former example included the necessary words, the first 3 letters of the Estonian word *mul* could either be kept whole by pairing it with the English *me*, or segmented into *mu* and *l* by pairing these with *my* and *on*. Starting at this point both segmentations would have identical alignments, which means that if two partial segmentations meet, the less probable one can be ruled out.

Thus the search is conducted via dynamical programming. All partial results are saved in the process and later reused when the new partial alignment is built by picking all preceding alignments that could have led to the new one and selecting the one with the highest probability. This way a flow graph of the possible alignments is built. The probability of the new partial alignment is then calculated as the product of its selected predecessor probability and the model probability of the newly added morph and word pair.

The backward pass, as the name suggests, traverses the flow graph from end to beginning. First, the alignment leading to the final node with maximum probability is picked. In the example above that would mean taking the top entering arc *le : to*, since its probability (0.02) is higher than of the bottom arc (*mulle : me*, 0.01). The algorithm then goes along the picked arc to the previous node, where selecting the maximum probability entering arc is repeated. The process is continued until the starting node is reached.

4 Experiments

The previous section described a method of finding the most probable alignment between the source sentence morphs and the target sentence word forms, hence in addition finding the most

<i>f</i>	<i>e</i>	prob.	<i>f</i>	<i>e</i>	prob.
mulle	me	0.01	le	to	0.4
mul	me	0.05	le	on	0.1
mu	my	0.3	l	on	0.3
mull	bubble	0.9

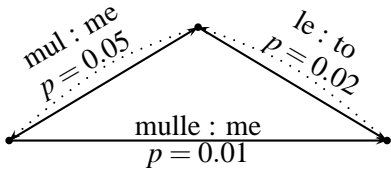


Figure 2: An example of a forward-backward search for a segmentation of the Estonian-English pair *mulle : to me* according to a small alignment model. The flow graph of the forward pass is shown with solid arcs, the selected arcs of the backward pass – with dotted arcs.

probable segmentation of the source sentence. This section presents the results of evaluating the method empirically.

Typical evaluation for an alignment model consists of comparing it to a reference alignment. However these are at most available between word forms, thus excluding our case.

On the other hand, segmentation models are typically compared to a golden standard, which usually is obtained according to linguistic morphology. However the aim of the presented method is to find a possibly better way of segmenting the language. Instead of binding the model to a direct reference we evaluate it only in the context of machine translation, comparing the effects of the method on the translation results.

4.1 Methods of Evaluation

In our opinion there are two intuitive ways of evaluating the presented segmentation and alignment method in context of statistical machine translation.

First, the method can be used as preprocessing to segment the source words into morphs. This however requires a way of applying the model to unseen source text without an available target guideline. Extending the search method defined in the previous section for the current case yields

$$\hat{\mathbf{a}} = \arg \max_{\mathbf{a}, \mathbf{e}} p(\mathbf{f}, \mathbf{a} | \mathbf{e}) \cdot p(\mathbf{e}, \mathbf{a} | \mathbf{f})$$

Consider the same example as on figure 2. Having the same probability table and only the Estonian word *mulle* to segment (with no English

phrase to guide), the search graph expands into the one on figure 3, with the resulting most probable segmentation being *mu-l-le*.

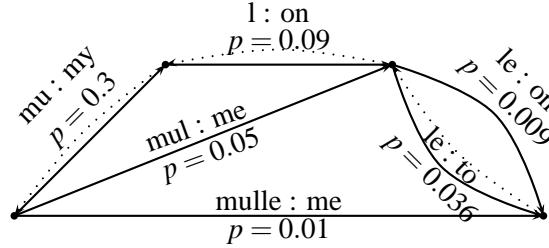


Figure 3: Unguided segmentation of the Estonian word “mulle”.

Although feasible in practice, this application method has a serious drawback, namely the methods of segmenting the source sentences during training and applying are different. As the results will show, this drawback greatly influences the resulting performance of the translation system.

An alternative way of using the model is to integrate it with the reordering and other models, replacing the word- or phrase-based lexical modelling. Here we emulate this setup by processing the test corpus with the target language corpus part kept and used as a guideline. Although the setup is not useful in practice (since no guideline is provided for the unseen source text), it could assess the performance of the introduced alignment model as part of a composite system.

Thus we perform two kinds of evaluation – guided and unguided; the first one uses the target part of the test corpus a guideline for alignment and segmentation, and the latter only uses it for evaluating the translation, performing segmentation on the source part solely.

4.2 Used Corpora and Tools

To evaluate the performance of the introduced model we used the OpenSubtitles corpus, described in (Tiedemann, 2007). We picked four language pairs with English as the target and Estonian, German, Norwegian and Polish as the source languages; the choice was motivated by the languages having variably rich morphology and being from different language families.

Standard preprocessing was applied to all corpora: the whole text was converted to lower-case, numbers and punctuation were put separately (preserving XML entities), sentence pairs with either sentence empty, longer than 100 words or the ratio of the lengths of the two sentences in words

Translation direction	Baseline		Unguided segm.		Guided segm.	
	BLEU	OOV	BLEU	OOV	BLEU	OOV
Estonian-English	0.189	10.6%	0.118	17.4%	0.166	2.4%
German-English	0.169	6.6%	0.088	13.3%	0.142	1.1%
Norwegian-English	0.204	5.0%	0.135	9.6%	0.189	1.2%
Polish-English	0.216	8.7%	0.118	18.1%	0.177	2.0%

Table 1: Experiment results: the score and the out-of-vocabulary rate of the translations. In the baseline the text is translated with no segmentation. In unguided segmentation the target part of the test corpus is only used for scoring, in the guided segmentation – also as a guideline for segmenting the source part.

exceeding 9 were excluded. Finally 2000 random sentence pairs were excluded from the training corpus for held-out validation. This resulted in the following number of sentence pairs in the training corpora: 68.4 thousand for German, 66.7 thousand for Norwegian, 49.9 thousand for Polish and 27.5 thousand for Estonian.

We used the Moses toolkit (Koehn et al., 2007) in our experiments. Translation quality is evaluated with the BLEU score (Papineni et al., 2001); in addition we kept track of the out-of-vocabulary (OOV) rate in the translation hypotheses.

4.3 Results

The resulting BLEU scores and OOV rates of the translations are presented in table 1. Three translation setups are covered: the baseline, the guided and unguided segmentation evaluations. The baseline was obtained by training a translation system with an out-of-the-box Moses toolkit, without any segmentation.

As expected the results of the unguided segmentation are way below the baseline for all four language pairs. In addition the OOV rates are considerably higher in all cases. This supports the suggestion that using different methods of segmenting during learning and applying is not efficient.

The scores of the guided segmentation are also below the expectations as well as the baseline. On the contrary, the OOV rates are heavily reduced, especially for German and Norwegian. However due to the lower scores this is not necessarily a benefit, as translating more morphs/words wrongly does not improve translation.

On the other hand all guided segmentation results are higher than in the unguided case. This strongly suggests that guidelines help to improve the segmentation. Although in practice there is no guideline while translating unseen text, we follow with a discussion of how this could be useful.

5 Discussion

The results of both guided and unguided segmentation indicate a flaw in the model. This can only mean that the estimation phase is to blame, as all other steps were directly defined to optimize the objective.

Manual evaluation of some translations showed that only in some rare cases (like typing mistakes in the input) the segmentation improved the translation. It was found after inspecting the learned parameters and resulting segmentations that most morphs remained unaligned despite the presence of the target sentence in the guided case. The learning parameters included many pairs of small morphs aligned to frequent target words.

This shows that in the current case joint learning did not fulfill the requirements and most meaningful morph pairs were pushed out of the corresponding probability distributions by shorter less meaningful more frequent morph pairs.

One other problem is the corpus – apart from typing mistakes and alignment errors, due to the small size of the training set the test set includes a lot of unseen word forms.

Coming back to the guided approach, in reality unseen source text naturally is not augmented with the target text. However in a joint system the models themselves provide a kind of a guideline for each other by binding together different aspects of translation. In other words the search space of the lexical model is constrained by disallowing lexical choices that cause bad segmentations and ungrammatical/badly ordered output. Thus the first step to enable the guided approach in practice would be to combine the introduced model with models of reordering and language correctness in a joint learning system.

An additional possible source of error is that morphs are typically more ambiguous in their

meaning than word forms and the word and sentence context of the morph is required to determine the meaning. This issue could be addressed by adapting phrase-based translation models, so that instead of single morphs, morph sequences can be matched with the target sentence.

6 Conclusions

This paper introduced the approach of aligning word parts, or morphs, instead of word forms. This results in automatic unsupervised segmentation of the word forms, which is beneficial for morphologically rich languages. The approach requires no additional resources or tools. We described an asymmetric jointly learned alignment used in the approach together with algorithms for learning the lexical parameters and searching for the most probable alignment of a sentence pair.

The model was evaluated as a preprocessing step for a translation system. The source text was aligned to the target text, thus resulting in its segmentation. The segmented text is then used to train the translation model.

The scores of the enhanced translation are below the baseline, which indicates a flaw in the learning of the alignment parameters. On the other hand using the target test set as a guideline for alignment improves the scores.

The results suggest a lot of future work. From the technical point of view the model has to be optimized to enable learning on larger data sets. The alignment model and the learning of the parameters has to be replaced in order to find better morph/word pairs and their distributions. Finally, the approach can be integrated into a joint translation system to possibly lead to much better results and higher translation quality.

References

Ibrahim Badr, Rabih Zbib, and James Glass. 2008. Segmentation for english-to-arabic statistical machine translation. In *Proceedings of ACL'08*, pages 153–156. Columbus, Ohio.

Ondřej Bojar, Pavel Straňák, and Daniel Zeman. 2008. English-Hindi Translation in 21 Days. In *Proceedings of the ICON-2008 NLP Tools Contest*. Pune, India.

Peter F. Brown, Stephen Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation:

Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Harri Kirik and Mark Fishel. 2008. Modelling linguistic phenomena with unsupervised morphology for improving statistical machine translation. In *Proceedings of the Workshop on Unsupervised Methods in NLP, SLTC'08*. Stockholm, Sweden.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL'07*, volume 45, page 2. Prague, Czech Republic.

Young-Suk Lee. 2004. Morphological analysis for statistical machine translation. In *Proceedings of NAACL-HLT'04*, pages 57–60. Boston, Massachusetts, USA.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *Proceedings of NAACL-HLT'06*, pages 104–111.

Sonja Nießen and Hermann Ney. 2004. Statistical machine translation with scarce resources using morpho-syntactic information. *Computational linguistics*, 30(2):181–204.

Kishore Papieni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL'01*, pages 311–318. Philadelphia, PA, USA.

Siriwan Sereewattana. 2003. *Unsupervised segmentation for statistical machine translation*. Ph.D. thesis, University of Edinburgh.

Benjamin Snyder and Regina Barzilay. 2008. Unsupervised multilingual learning for morphological segmentation. In *Proceedings of ACL'08*, pages 737–745. Columbus, Ohio.

Jörg Tiedemann. 2007. Building a multilingual parallel subtitle corpus. In *Proceedings of CLIN 17*. Leuven, Belgium.

Sami Virpioja, Jaakko J. Väyrynen, Mathias Creutz, and Markus Sadeniemi. 2007. Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. In *Proceedings of MT Summit XI*, pages 491–498. Copenhagen, Denmark.