Unsupervised Search for The Optimal Segmentation for Statistical Machine Translation

Coşkun Mermer[1,3] and **Ahmet Afşın Akın**[2,3]
[1]Boğaziçi University, Bebek, Istanbul, Turkey
[2]Istanbul Technical University, Sarıyer, Istanbul, Turkey
[3]TÜBİTAK-UEKAE, Gebze, Kocaeli, Turkey
{coskun,ahmetaa}@uekae.tubitak.gov.tr

Abstract

We tackle the previously unaddressed problem of unsupervised determination of the optimal morphological segmentation for statistical machine translation (SMT) and propose a segmentation metric that takes into account both sides of the SMT training corpus. We formulate the objective function as the posterior probability of the training corpus according to a generative segmentation-translation model. We describe how the IBM Model-1 translation likelihood can be computed incrementally between adjacent segmentation states for efficient computation. Submerging the proposed segmentation method in a SMT task from morphologically-rich Turkish to English does not exhibit the expected improvement in translation BLEU scores and confirms the robustness of phrase-based SMT to translation unit combinatorics. A positive outcome of this work is the described modification to the sequential search algorithm of Morfessor (Creutz and Lagus, 2007) that enables arbitrary-fold parallelization of the computation, which unexpectedly improves the translation performance as measured by BLEU.

1 Introduction

In statistical machine translation (SMT), words are normally considered as the building blocks of translation models. However, especially for morphologically complex languages such as Finnish, Turkish, Czech, Arabic etc., it has been shown that using sub-lexical units obtained after morphological preprocessing can improve the machine translation performance over a word-based system (Habash and Sadat, 2006; Oflazer and Durgar El-Kahlout, 2007; Bisazza and Federico, 2009). However, the effect of segmentation on translation performance is indirect and difficult to isolate (Lopez and Resnik, 2006).

The challenge in designing a sub-lexical SMT system is the decision of what segmentation to use. Linguistic morphological analysis is intuitive, but it is language-dependent and could be highly ambiguous. Furthermore, it is not necessarily optimal in that (i) manually engineered segmentation schemes can outperform a straightforward linguistic morphological segmentation, e.g., (Habash and Sadat, 2006), and (ii) it may result in even worse performance than a word-based system, e.g., (Durgar El-Kahlout and Oflazer, 2006).

A SMT system designer has to decide what segmentation is optimal for the translation task at hand. Existing solutions to this problem are predominantly heuristic, language-dependent, and as such are not easily portable to other languages. Another point to consider is that the optimal degree of segmentation might decrease as the amount of training data increases (Lee, 2004; Habash and Sadat, 2006). This brings into question: For the particular language pair and training corpus at hand, what is the optimal (level of) subword segmentation? Therefore, it is desirable to learn the optimal segmentation in an unsupervised manner.

In this work, we extend the method of Creutz and Lagus (2007) so as to maximize the translation posterior in unsupervised segmentation. The learning process is tailored to the particular SMT task via the same parallel corpus that is used in training the statistical translation models.

2 Related Work

Most works in SMT-oriented segmentation are supervised in that they consist of manual experimentation to choose the best among a set of segmentation schemes, and are language(pair)-dependent. For Arabic, Sadat and Habash (2006) present several morphological preprocessing schemes that entail varying degrees of decomposition and com-

Proceedings of the ACL 2010 Student Research Workshop, pages 31–36,
Uppsala, Sweden, 13 July 2010. ©2010 Association for Computational Linguistics

pare the resulting translation performances in an Arabic-to-English task. Shen et al. (2007) use a subset of the morphology and apply only a few simple rules in segmenting words. Durgar El-Kahlout and Oflazer (2006) tackle this problem when translating from English to Turkish, an agglutinative language. They use a morphological analyzer and disambiguation to arrive at morphemes as tokens. However, training the translation models with morphemes actually degrades the translation performance. They outperform the word-based baseline only after some selective morpheme grouping. Bisazza and Federico (2009) adopt an approach similar to the Arabic segmentation studies above, this time in a Turkish-to-English translation setting.

Unsupervised segmentation by itself has garnered considerable attention in the computational linguistics literature (Poon et al., 2009; Snyder and Barzilay, 2008; Dasgupta and Ng, 2007; Creutz and Lagus, 2007; Brent, 1999). However, few works report their performance in a translation task. Virpioja et al. (2007) used Morfessor (Creutz and Lagus, 2007) to segment both sides of the parallel training corpora in translation between Danish, Finnish, and Swedish, but without a consistent improvement in results.

Morfessor, which gives state of the art results in many tests (Kurimo et al., 2009), uses only monolingual information in its objective function. It is conceivable that we can achieve a better segmentation for translation by considering not one but both sides of the parallel corpus. A posssible choice is the post-segmentation alignment accuracy. However, Elming et al. (2009) show that optimizing segmentation with respect to alignment error rate (AER) does not improve and even degrades machine translation performance. Snyder and Barzilay (2008) use bilingual information but the segmentation is learned independently from translation modeling.

In Chang et al. (2008), the granularity of the Chinese word segmentation is optimized by training SMT systems for several values of a granularity bias parameter and it is found that the value that maximizes translation performance (as measured by BLEU) is different than the value that maximizes segmentation accuracy (as measured by precision and recall).

One motivation in morphological preprocessing before translation modeling is "morphology matching" as in Lee (2004) and in the scheme "EN" of Habash and Sadat (2006). In Lee (2004), the goal is to match the lexical granularities of the two languages by starting with a fine-grained segmentation of the Arabic side of the corpus and then merging or deleting Arabic morphemes using alignments with a part-of-speech tagged English corpus. But this method is not completely unsupervised since it requires external linguistic resources in initializing the segmentation with the output of a morphological analyzer and disambiguator. Talbot and Osborne (2006) tackle a special case of morphology matching by identifying redundant distinctions in the morphology of one language compared to another.

3 Method

Maximizing translation performance directly would require SMT training and decoding for each segmentation hypothesis considered, which is computationally infeasible. So we make some conditional independence assumptions using a generative model and decompose the posterior probability $P(M_f|e, f)$. In this notation e and f denote the two sides of a parallel corpus and M_f denotes the segmentation model hypothesized for f. Our approach is an extension of Morfessor (Creutz and Lagus, 2007) so as to include the translation model probability in its cost calculation. Specifically, the segmentation model takes into account the likelihood of both sides of the parallel corpus while searching for the optimal segmentation. The joint likelihood is decomposed into a prior, a monolingual likelihood, and a translation likelihood, as shown in Eq. 1.

$$P(e, f, M_f) = P(M_f)P(f|M_f)P(e|f, M_f) \tag{1}$$

Assuming conditional independence between e and M_f given f, the maximum *a posteriori* (MAP) objective can be written as:

$$\hat{M}_f = \arg\max_{M_f} P(M_f)P(f|M_f)P(e|f) \tag{2}$$

The role of the bilingual component $P(e|f)$ in Eq. 2 can be motivated with a simple example as follows. Consider an occurrence of two phrase pairs in a Turkish-English parallel corpus and the two hypothesized sets of segmentations for the Turkish phrases as in Table 1. Without access to the English side of the corpus, a monolingual segmenter can quite possibly score Seg. #1

	Phrase #1	Phrase #2
Turkish phrase:	anahtar	anahtarım
English phrase:	key	my key
Seg. #1:	anahtar	anahtarı +m
Seg. #2:	anahtar	anahtar +ım

Table 1: Example segmentation hypotheses

higher than Seg. #2 (e.g., due to the high frequency of the observed morph "+m"). On the other hand, a bilingual segmenter is expected to assign a higher alignment probability $P(e|f)$ to Seg. #2 than Seg. #1, because of the aligned words key||anahtar, therefore ranking Seg. #2 higher.

The two monolingual components of Eq. 2 are computed as in Creutz and Lagus (2007). To summarize briefly, the prior $P(M_f)$ is assumed to only depend on the frequencies and lengths of the individual morphs, which are also assumed to be independent. The monolingual likelihood $P(f|M_f)$ is computed as the product of morph probabilities estimated from their frequencies in the corpus.

To compute the bilingual (translation) likelihood $P(e|f)$, we use IBM Model 1 (Brown et al., 1993). Let an aligned sentence pair be represented by (s_e, s_f), which consists of word sequences $s_e = e_1, ..., e_l$ and $s_f = f_1, ..., f_m$. Using a purely notational switch of the corpus labels from here on to be consistent with the SMT literature, where the derivations are in the form of $P(f|e)$, the desired translation probability is given by the expression:

$$P(f|e) = \frac{P(m|e)}{(l+1)^m} \prod_{j=1}^{m} \sum_{i=0}^{l} t(f_j|e_i), \qquad (3)$$

The sentence length probability distribution $P(m|e)$ is assumed to be Poisson with the expected sentence length equal to m.

3.1 Incremental computation of Model-1 likelihood

During search, the translation likelihood $P(e|f)$ needs to be calculated according to Eq. 3 for every hypothesized segmentation.

To compute Eq. 3, we need to have at hand the individual morph translation probabilities $t(f_j|e_i)$. These can be estimated using the EM algorithm given by (Brown, 1993), which is guaranteed to converge to a global maximum of the likelihood for Model 1. However, running the EM algorithm to optimization for each considered segmentation model can be computationally expensive, and can result in overtraining. Therefore, in this work we used the likelihood computed after the first EM iteration, which also has the nice property that $P(f|e)$ can be computed incrementally from one segmentation hypothesis to the next.

The incremental updates are derived from the equations for the count collection and probability estimation steps of the EM algorithm as follows. In the count collection step, in the first iteration, we need to compute the fractional counts $c(f_j|e_i)$ (Brown et al., 1993):

$$c(f_j|e_i) = \frac{1}{l+1}(\#f_j)(\#e_i), \qquad (4)$$

where $(\#f_j)$ and $(\#e_i)$ denote the number of occurrences of f_j in s_f and e_i in s_e, respectively.

Let f_k denote the word hypothesized to be segmented. Let the resulting two sub-words be f_p and f_q, any of which may or may not previously exist in the vocabulary. Then, according to Eq. (4), as a result of the segmentation no update is needed for $c(f_j|e_i)$ for $j = 1 \ldots N$, $j \neq p, q$, $i = 1 \ldots M$ (note that f_k no longer exists); and the necessary updates $\Delta c(f_j|e_i)$ for $c(f_j|e_i)$, where $j = p, q$; $i = 1 \ldots M$ are given by:

$$\Delta c(f_j|e_i) = \frac{1}{l+1}(\#f_k)(\#e_i). \qquad (5)$$

Note that Eq. (5) is nothing but the previous count value for the segmented word, $c(f_k|e_i)$. So, all needed in the count collection step is to copy the set of values $c(f_k|e_i)$ to $c(f_p|e_i)$ and $c(f_q|e_i)$, adding if they already exist.

Then in the probability estimation step, the normalization is performed including the newly added fractional counts.

3.2 Parallelization of search

In an iteration of the algorithm, all words are processed in random order, computing for each word the posterior probability of the generative model after each possible binary segmentation (splitting) of the word. If the highest-scoring split increases the posterior probability compared to not splitting, that split is accepted (for all occurrences of the word) and the resulting sub-words are explored recursively for further segmentations. The process is repeated until an iteration no more results in a significant increase in the posterior probability.

The search algorithm of Morfessor is a greedy algorithm where the costs of the next search points

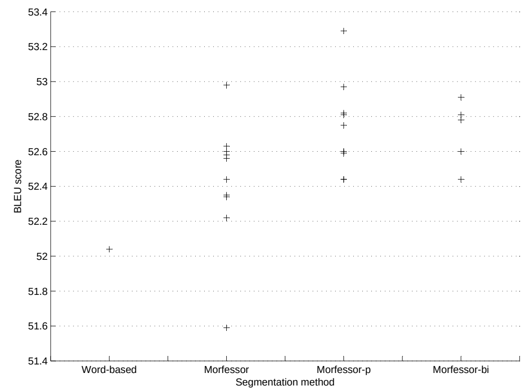

Figure 1: BLEU scores obtained with different segmentation methods. Multiple data points for a system correspond to different random orders in processing the data (Creutz and Lagus, 2007).

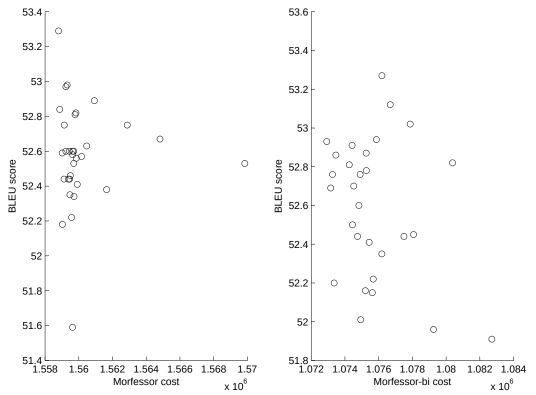

Figure 2: Cost-BLEU plots of Morfessor and Morfessor-bi. Correlation coefficients are -0.005 and -0.279, respectively.

are affected by the decision in the current step. This leads to a sequential search and does not lend itself to parallelization.

We propose a slightly modified search procedure, where the segmentation decisions are stored but not applied until the end of an iteration. In this way, the cost calculations (which is the most time-consuming component) can all be performed independently and in parallel. Since the model is not updated at every decision, the search path can differ from that in the sequential greedy search and hence result in different segmentations.

4 Results

We performed *in vivo* testing of the segmentation algorithm on the Turkish side of a Turkish-to-English task. We compared the segmentations produced by Morfessor, Morfessor modified for parallel search (Morfessor-p), and Morfessor with bilingual cost (Morfessor-bi) against the word-based performance. We used the ATR Basic Travel Expression Corpus (BTEC) (Kikui et al., 2006), which contains travel conversation sentences similar to those in phrase-books for tourists traveling abroad. The training corpus contained 19,972 sentences with average sentence length 5.6 and 7.7 words for Turkish and English, respectively. The test corpus consisted of 1,512 sentences with 16 reference translations. We used GIZA++ (Och and Ney, 2003) for post-segmentation token alignments and the Moses toolkit (Koehn et al., 2007) with default parameters for phrase-based translation model generation and decoding. Target language models were trained on the English side of the training corpus using the SRILM toolkit (Stolcke, 2002). The BLEU metric (Papineni et al., 2002) was used for translation evaluation.

Figure 1 compares the translation performance obtained using the described segmentation methods. All segmentation methods generally improve the translation performance (Morfessor and Morfessor-p) compared to the word-based models. However, Morfessor-bi, which utilizes both sides of the parallel corpus in segmenting, does not convincingly outperform the monolingual methods.

In order to investigate whether the proposed bilingual segmentation cost correlates any better than the monolingual segmentation cost of Morfessor, we show several cost-BLEU pairs obtained from the final and intermediate segmentations of Morfessor and Morfessor-bi in Fig. 2. The correlation coefficients show that the proposed bilingual metric is somewhat predictive of the translation performance as measured by BLEU, while the monolingual Morfessor cost metric has almost no correlation. Yet, the strong noise in the BLEU scores (vertical variation in Fig. 2) diminishes the effect of this correlation, which explains the inconsistency of the results in Fig. 1. Indeed, in our experiments even though the total cost kept decreasing at each iteration of the search algorithm, the BLEU scores obtained by those intermediate segmentations fluctuated without any consistent improvement.

Table 2 displays sample segmentations produced by both the monolingual and bilingual segmentation algorithms. We can observe that utilizing the English side of the corpus enabled

Count	Morfessor	Morfessor-bi	English Gloss
7	anahtar	anahtar	(the) key
6	anahtar + ımı	anahtar + ımı	my key (*ACC.*)
5	anahtarla	anahtar + la	with (the) key
4	anahtarı	anahtar + ı	^{1}(the) key (*ACC.*); ^{2}his/her key
3	anahtarı + m	anahtar + ım	my key
3	anahtarı + n	anahtar + ın	^{1}your key; ^{2}of (the) key
1	anahtarı + nız	anahtar + ınız	your (*pl.*) key
1	anahtarı + nı	anahtar + ını	^{1}your key (*ACC.*); ^{2}his/her key (*ACC.*)
1	anahtar + ınızı	anahtar + ınızı	your (*pl.*) key (*ACC.*)
1	oyun + lar	oyunlar	(the) games
2	oyun + ları	oyunlar + ı	^{1}(the) games (*ACC.*); ^{2}his/her games; ^{3}their game(s)
1	oyun + ların	oyunlar + ı + n	^{1}of (the) games; ^{2}your games
1	oyun + larınızı	oyunlar + ı + n + ızı	your (*pl.*) games (*ACC.*)

Table 2: Sample segmentations produced by Morfessor and Morfessor-bi

Morfessor-bi: (i) to consistently identify the root word “anahtar” (top portion), and (ii) to match the English plural word form “games” with the Turkish plural word form “oyunlar” (bottom portion). Monolingual Morfessor is unaware of the target segmentation, and hence it is up to the subsequent translation model training to learn that “oyun” is sometimes translated as “game” and sometimes as “games” in the segmented training corpus.

5 Conclusion

We have presented a method for determining optimal sub-word translation units automatically from a parallel corpus. We have also showed a method of incrementally computing the first iteration parameters of IBM Model-1 between segmentation hypotheses. Being language-independent, the proposed algorithm can be added as a one-time preprocessing step prior to training in a SMT system without requiring any additional data/linguistic resources. The initial experiments presented here show that the translation units learned by the proposed algorithm improves on the word-based baseline in both translation directions.

One avenue for future work is to relax some of the several independence assumptions made in the generative model. For example, independence of consecutive morphs could be relaxed by an HMM model for transitions between morphs (Creutz and Lagus, 2007). Other future work includes optimizing the segmentation of both sides of the corpus and experimenting with other language pairs.

It is also possible that the probability distributions are not discriminative enough to outweigh the model prior tendencies since the translation probabilities are estimated only crudely (single iteration of Model-1 EM algorithm). A possible candidate solution would be to weigh the translation likelihood more in calculating the overall cost. In fact, this idea could be generalized into a log-linear modeling (e.g., (Poon et al., 2009)) of the various components of the joint corpus likelihood and possibly other features.

Finally, integration of sub-word segmentation with the phrasal lexicon learning process in SMT is desireable (e.g., translation-driven segmentation in Wu (1997)). Hierarchical models (Chiang, 2007) could cover this gap and provide a means to seamlessly integrate sub-word segmentation with statistical machine translation.

Acknowledgements

The authors would like to thank Murat Saraçlar for valuable discussions and guidance in this work, and the anonymous reviewers for very useful comments and suggestions. Murat Saraçlar is supported by the TÜBA-GEBİP award.

References

Arianna Bisazza and Marcello Federico. 2009. Morphological Pre-Processing for Turkish to English Statistical Machine Translation. In *Proc. of the International Workshop on Spoken Language Translation*, pages 129–135, Tokyo, Japan.

M.R. Brent. 1999. An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, 34(1):71–105.

P.F. Brown, V.J. Della Pietra, S.A. Della Pietra, and R.L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Pi-Chuan Chang, Michel Galley, and Christopher D. Manning. 2008. Optimizing Chinese word segmentation for machine translation performance. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 224–232, Columbus, Ohio.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

M. Creutz and K. Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing*, 4(1):1–34.

Sajib Dasgupta and Vincent Ng. 2007. High-performance, language-independent morphological segmentation. In *Proceedings of HLT-NAACL*, pages 155–163, Rochester, New York.

İlknur Durgar El-Kahlout and Kemal Oflazer. 2006. Initial explorations in English to Turkish statistical machine translation. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 7–14, New York City, New York, USA.

Jakob Elming, Nizar Habash, and Josep M. Crego. 2009. Combination of statistical word alignments based on multiple preprocessing schemes. In Cyrill Goutte, Nicola Cancedda, Marc Dymetman, and George Foster, editors, *Learning Machine Translation*, chapter 5, pages 93–110. MIT Press.

Nizar Habash and Fatiha Sadat. 2006. Arabic preprocessing schemes for statistical machine translation. In *Proc. of the HLT-NAACL, Companion Volume: Short Papers*, pages 49–52, New York City, USA.

G. Kikui, S. Yamamoto, T. Takezawa, and E. Sumita. 2006. Comparative study on corpora for speech translation. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5):1674–1682.

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 the 45th Annual Meeting of the Association for Computational Linguistics, Companion Volume: Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic.

M. Kurimo, S. Virpioja, V.T. Turunen, G.W. Blackwood, and W. Byrne. 2009. Overview and Results of Morpho Challenge 2009. In *Working notes of the CLEF workshop*.

Young-Suk Lee. 2004. Morphological analysis for statistical machine translation. In *Proceedings of HLT-NAACL, Companion Volume: Short Papers*, pages 57–60, Boston, Massachusetts, USA.

Adam Lopez and Philip Resnik. 2006. Word-based alignment, phrase-based translation: What's the link? In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA-06)*, pages 90–99.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Kemal Oflazer and İlknur Durgar El-Kahlout. 2007. Exploring different representational units in English-to-Turkish statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 25–32, Prague, Czech Republic.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA.

Hoifung Poon, Colin Cherry, and Kristina Toutanova. 2009. Unsupervised morphological segmentation with log-linear models. In *Proceedings of HLT-NAACL*, pages 209–217, Boulder, Colorado.

Fatiha Sadat and Nizar Habash. 2006. Combination of Arabic preprocessing schemes for statistical machine translation. In *Proc. of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 1–8, Sydney, Australia.

Wade Shen, Brian Delaney, and Tim Anderson. 2007. The MIT-LL/AFRL IWSLT-2007 MT system. In *Proc. of the International Workshop on Spoken Language Translation*, Trento, Italy.

Benjamin Snyder and Regina Barzilay. 2008. Unsupervised multilingual learning for morphological segmentation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: HLT*, pages 737–745, Columbus, Ohio.

A. Stolcke. 2002. SRILM-an extensible language modeling toolkit. In *Seventh International Conference on Spoken Language Processing*, volume 3.

David Talbot and Miles Osborne. 2006. Modelling lexical redundancy for machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 969–976, Sydney, Australia.

S. Virpioja, J.J. Väyrynen, M. Creutz, and M. Sadeniemi. 2007. Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. In *Machine Translation Summit XI*, pages 491–498, Copenhagen, Denmark.

D. Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–403.