Log-linear Models for Uyghur Segmentation in Spoken Language Translation

Chenggang Mi[1,2], **Yating Yang**[1,2], **Rui Dong**[1,2,3], **Xi Zhou**[1,2], **Lei Wang**[1,2], **Xiao Li**[1,2], and **Tonghai Jiang**[1,2]
[1]The Xinjiang Technical Institute of Physics & Chemistry of Chinese Academy of Sciences, Urumqi, China
[2]Key laboratory of speech language information processing of Xinjiang, Urumqi, China
[3]University of Chinese Academy of Sciences, Beijing, China
{micg, yangyt, dongrui, zhouxi, wanglei, xiaoli, jth}@ms.xjb.ac.cn

Abstract

To alleviate data sparsity in spoken Uyghur machine translation, we proposed a log-linear based morphological segmentation approach. Instead of learning model only from monolingual annotated corpus, this approach optimizes Uyghur segmentation for spoken translation based on both bilingual and monolingual corpus. Our approach relies on several features such as traditional conditional random field (CRF) feature, bilingual word alignment feature and monolingual suffix-word co-occurrence feature. Experimental results shown that our proposed segmentation model for Uyghur spoken translation achieved 1.6 BLEU score improvements compared with the state-of-the-art baseline.

1 Introduction

Low resource languages like Uyghur usually suffer from data sparsity in related NLP tasks. Due to need a large scale parallel corpus to train the translation model, this situation becomes even worse in Uyghur - Chinese machine translation. To overcome this problem, morphological segmentation is often used to alleviate the data sparsity.

Most approaches on morphological segmentation such as CRF based model rely on annotated data heavily and do not consider the informal situation like spoken language translation (Table 1). Therefore, using a traditional Uyghur morphological segmentation model to segment the corpus for spoken language translation (SLT) cannot expect to achieve a good performance. In this study, we research on Uyghur morphological segmentation for Uyghur-Chinese spoken language translation.[1] We proposed a novel method to optimize morphological segmentation for Uyghur SLT. Our approach based on a log-linear model, several features include CRF feature, bilingually-constrained feature and monolingual co-occurrence feature are derived and feed to the model, the model provide an optimized morphological segmentation results for SLT. Experimental results shown that our proposed approach can achieve 1.6+ BLEU improvements, which outperforms other baselines significantly.

The main contributions of this paper can be summarized as following:

- We propose a log-linear based morphological segmentation model for Uyghur-Chinese spoken language translation, several features are integrated into it to optimize the performance of SLT model.
- Our features include CRF feature, bilingual word alignment feature and monolingual suffix-other words (OW, which means words in current Uyghur sentence except current word) co-occurrence feature, which derived from bilingual corpus and monolingual.
- Through exploring the log-linear based model for spoken Uyghur segmentation, we show that these features: CRF, bilingual word alignment and monolingual suffix-OW co-occurrence are all useful to Uyghur-Chinese spoken language translation.

The rest of this paper is organized as follows: we present the features of Uyghur and morphological segmentation for statistical machine translation (SMT) which are related to our research in section 2; in section 3, we give a detailed introduc-

[1]In this paper, we write Uyghur with the Latin alphabet and Chinese with Pinyin.

Proceedings of Recent Advances in Natural Language Processing, pages 492–500,
Varna, Bulgaria, Sep 4–6 2017.
https://doi.org/10.26615/978-954-452-049-6_065

Uyghur (source)	Chinese (target)
almighanmu ?	hai mei you na ma ? (Have you ever taken it ?)
shu Otkendimu .	you mei you kao shang ne ? (Have you got it ?)
bishim qalaymiqan .	wo nao zi yi pian hun luan . (My mind goes blank .)

Table 1: Examples of spoken Uyghur-Chinese sentence pairs.

tion of our method; experimental settings and results analysis are described in section 4; we finally review related work in section 5 and conclude in section 6.

2 Background

In this section, we first present some features of Uyghur. Then, we give some introductions about morphological segmentation in SMT. Finally, we describe challenges exist in Uyghur segmentation in SLT.

2.1 Introduction of Uyghur

Uyghur is a Turkic language with 10 to 25 million speakers, which is an official language of the Xinjiang Uyghur Autonomous Region of Western China. Various other countries also have Uyghur-speaking communities.

Uyghur is an agglutinative language with not only a very rich but also a productive derivational and inflectional morphology (Table 2). Also, Uyghur displays vowel harmony, lacks noun classes or grammatical gender, and is a left-branching language with subject-object-verb (SOV) word order.

Uyghur (word)	Uyghur (stem + suffix (es))
aliqanimda	**aliqan**+im+da
etrapidikilerni	**etrap**+i+diki+ler+ni
qurulmasining	**qurulma**+si+ning
qalduridu	**qal**+dur+i+d+u

Table 2: Examples of Uyghur word formation.

2.2 Morphological Segmentation in SMT

Data sparsity is one of the enduring problems in SMT. For low-resourced languages like Uyghur, this situation is even worse in related SMT tasks. As one of the most important parts of SMT, the word alignment model try to capture the probability of $p(e|f)$, where f is a word in source language (Uyghur) and e is the target word (Chinese). When translating between two unrelated languages such as Uyghur (morphologically-rich language) and Chinese (morphologically-poor language), disparate morphological systems can intensifies the problem of data sparsity because the large number of word forms created through morphologically productive processes hinders attempts to find concise mappings between concepts.

To alleviate the data sparsity in SMT, morphological analysis methods are proposed. Morphological analysis identifies functional morphemes to be merged into meaning-bearing stems or to be deleted. In Uyghur, functional morphemes typically belong to suffixes.

2.3 Challenges of Uyghur Segmentation in SLT

Unlike the formal news corpus, which is typically written with a clear intention, and moreover has been editorially controlled according to standards of language use; the informal conversation (dialogues) corpus has different intentions and languages use. Therefore, we may face several challenges in morphological segmentation for Uyghur-Chinese spoken translation:

First, the conversion sentence usually very short compared with news corpus; therefore, limited context can be used in translation model learning.

bilina qala . (ke yi kan chu lai ya .)

uxla tExi ? (hai mei shui jiao ba ?)

Second, a large scale of one-to-many and align to NULL alignments (Figure 1) exist in Uyghur-Chinese spoken corpus due to ellipsis in spoken language, which is very harmful to the translation performance.

Third, most of exist approaches on morphological segmentation are trained on formal corpus such as news, law et al., and these models do not perform well on Uyghur spoken corpus (Table 3).

3 Our Method

To overcome these difficulties, we proposed a novel method to optimize morphological segmentation for SLT. Our approach based on a log-linear model, several features include bilingually-constrained feature, monolingual co-occurrence

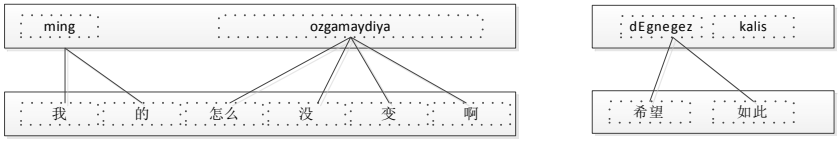

Figure 1: 1-to-many Word Alignments in Spoken Uyghur SMT.

Uyghur (source)	Chinese (translation results)	Chinese (reference)
nim e bol di aka	ge ge de di shen me a	zen me le, da ge.
he shu	jiu shi, shuo shi .	en, jiu shi de.
bol di xudayimgha shvkri	udayimgha shvkri	hao le, xie tian xie di.
chishlimiseng	chishlimiseng	ni bie yao.
he, mubarek bol sun	o, mu ba la ke de shui	en, gong xi ni.

Table 3: Examples of Uyghur sentences, translation results (by previous methods) and reference sentences.

feature are derived and feed to the model, the model provide an optimized morphological segmentation results for SLT.

3.1 Log-linear Model

Log-linear models are widely used in natural language processing (NLP) applications. One of the most important advantages of log-linear models is that they allow a very rich set of features to be used in a model, arguably much richer representations than the other simple estimation techniques (Liu et al., 2005).

We have a set of possible inputs **X** (morphemes), and a set of possible labels **Y** (R/D). The goal of our task is to model the conditional probability $p(y|x)$. Where for a (morpheme, label) pair $< x, y >$, $x \in \mathbf{X}$ and $y \in \mathbf{Y}$.

In our morphological segmentation task, we have some set **M** of possible morphemes, and a set **T** of possible tags. The set **Y** is simply equal to **T**, and **X** is the set of **M** is the set of contexts of the form $< m_1 m_2 m_3 ... m_n, t_1 t_2 t_3 ... t_{i-1} >$. Where n is the length of the input sentence, $m_j \in \mathbf{M}$, $(j \in \{1..n\})$, $i \in \{1...(n-1)\}$, and $t_j \in \mathbf{T}$ for $j \in \{1...(i-1)\}$.

Accordingly, log-linear model used in our study can be abstractly described as follows. For $m \in \mathbf{M}$, $t \in \mathbf{T}$

$$p(t|m; v) = \frac{exp(v \cdot f(m,t))}{\sum_{t' \in t} exp(v \cdot f(m,t'))} \quad (1)$$

Here, **M** is a set of input morphemes **T** is a set of possible labels; f is a feature function, which maps (m, t) pair to a feature vector $f(m, t)$, and **v** is a parameter vector. Note that the number of features and parameters should be the same in log-linear model.

3.2 Feature Functions

In this paper, we use the CRF as the basic feature in our log-linear model. Moreover, we also use additional information like bilingual word alignment and monolingual suffix-OW co-occurrence as two more features.

3.2.1 CRF Feature

We use a CRF based model to train a Uyghur morphological analyzer. Following (Ruokolainen et al., 2013)'s work, we treat the morphological segmentation as a sequence labeling problem.The CRF based morphological segmentation model can be described as

$$p(y|x; w) \propto \prod_{t=2}^{T} exp(w^T f(y_{t-1}, y_t, x, t)) \quad (2)$$

where x are characters in a word, y means corresponding class to each character. t indexes the characters, T is the length of word, **w** is the parameter vector, and f the vector-valued feature extracting function.

In this paper, we apply the tagging results of CRF model as the basic feature of our proposed approach. To adapt the spoken Uyghur segmentation situation, we extract the tagging probability of stem in each word in our Uyghur spoken corpus. Therefore, the feature function of CRF feature can

be described as

$$h(m,t,fT) = \prod_{i=1}^{l} p_{stem}(t_{stem}|m_i, fT) \quad (3)$$

Where m_i is the ith morpheme, l denotes the number of morphemes. p_stem means the probability of tagging the m_i as "stem" given m_i and a morphological segmentation model **fT**. Which can be calculate as

$$p_{stem}(t_{stem}|m_t) = \frac{N(t_{stem}, m_i)}{N(m_i)} \quad (4)$$

Here, $N(t_{stem}, m_i)$ is the number of times m_i tagging as "stem" and $N(m_i)$ denotes the number of times m_i appeared in training data.

3.2.2 Bilingual Word Alignment Feature

Word alignment is one of the most important topics in SMT, which derive correspondences between words in parallel data as an attempt to explain how translation comes about. Traditional approaches on morphological segmentation for Uyghur SLT suffers from data sparsity seriously. To overcome this problem, we present a bilingual word alignment feature and integrate it into the log-linear model.

Due to the lack of training data and insufficient contextual information in Uyghur-Chinese SLT, we often cannot obtain the correct correspondences of words between Uyghur and Chinese. In word alignment of Uyghur-Chinese spoken corpus, we find that some NULL alignment Chinese words can be aligned to some suffixes of Uyghur words. However, we cannot obtain correspondences between suffixes of Uyghur words and NULL alignment Chinese words directly. In this paper, we first obtain the word alignment results from a large scale Uyghur-Chinese parallel corpus on news domain; one-to-many alignments are extracted and the suffix-word alignment can be obtained accordingly

$$p_{1-to-many}(u) = \frac{N(u, cs)}{N(u)} \quad (5)$$

$N(u, cs)$ is number of times the Uyghur word u correspondding to $n(n > 1)$ Chinese words (1-to-many) in word alignment. $N(u)$ is the number of times Uyghur word u appeared in parallel corpus on news domain.

Accordingly, we define the bilingual word alignment feature function as

$$h(m,t,u,c,a) = \sum_{i=1}^{l_u} \sum_{j=1}^{n_i} \frac{p_{1-to-many}(u_i)}{n_{ij}} \quad (6)$$

Where l_u is the length of current Uyghur sentence, n_i is the number of suffixes of Uyghur word u_i.

3.2.3 Monolingual Stem-Suffixes Co-occurrence Feature

Uyghur is an agglutinative language. Nouns are inflected for number and case. Verbs are conjugated for tense, voice, aspect and mood. In Uyghur, we find that suffixes of a Uyghur word often appeared with other words (OW), and some of these words are omitted in informal corpus (such as conversion). Accordingly, OW word can express the same meaning with suffixes. Inspired by this, we can reserve certain suffix (es) as a single word if the OWs are not available. We can achieve this by counting the co-occurrence of stem-OW from a monolingual corpus in news domain.

Compare with bilingual parallel corpora, monolingual corpora are easy to build. We first segment a large scale of Uyghur monolingual corpus by a pre-trained morphological analyzer. Then, we obtain the stem-OW co-occurrence probability by counting the relative frequency:

$$p_{sow}(s|ow) = \frac{N_{co}(s, ow)}{N(ow)} \quad (7)$$

Here, $N_{co}(s, ow)$ is the frequency that a suffix is appeared with an OW in Uyghur monolingual corpus, $N(ow)$ is the frequency of one OW appeared.

Therefore, the feature function of stem-OW co-occurrence can be defined as

$$h(m,t,Mono) = \sum_{i=1}^{l} co\_occur(m_{ow(i)}|m_{s(i)}) \quad (8)$$

Where

$$co\_occur(m_{ow(i)}|m_{s(i)}, Mono) = p_{sow}(s|ow) \quad (9)$$

3.3 Model Training

In this study, we train the log-linear model with maximum-likelihood estimation (MLE), which is commonly used in machine learning. We assume that we have a training set, examples (m_i, t_i) are included in it, where m_i belongs to **M**, and t_i belongs to **T**.

For any example, we can calculate the log conditional probability as

$$L(v) = \sum_{i=1}^{n} \log p(t^{(i)}|m^{(j)}; v) \quad (10)$$

To prevent the log-linear model from overfitting the training data, we follow the common solution that modifies the objective function to include a regularization term. Regularization will prevent overfitting when we have a lot of features. Therefore, the function of the parameters **v** can be defined as

$$L'(v) = \sum_{i=1}^{n} \log p(t^{(i)}|m^{(j)}; v) - \frac{\lambda}{2} \sum_{k} v_k^2 \quad (11)$$

3.4 Decoding

By using the MLE training criteria, the optimal model parameters **v** can be estimated. Given the optimized parameters **v** and an input **M**, decoding with the log-linear model defined in section 3.1 can be described as follows:

$$T' = \arg\max_{T} p(T|M, v) \quad (12)$$

Where **M** is a sequence of morphemes, and **T** means a sequence of labels tagged by our model. **v** is a feature vector which is optimized during model training.

4 Experiments

In this section, we measure the effect of morphological segmentation on Uyghur-Chinese spoken language translation performance.

4.1 Data and settings

For spoken machine translation experiments, we train our model by Moses[2] on a Uyghur-Chinese parallel corpus of approximately 30K sentences, which consists mainly of documents in daily life from QQ, Weixin, etc. Develop set and test set are all collected from the same resources, which include 1K and 1.5K sentences, respectively (Table 4). We train a 5-gram language model on the Sougou corpus using the SRILM[3] with modified Kneser-Ney Smooth algorithm. We use the minimum error rate training (MERT) to optimize the feature weights on the develop set. We evaluate the performance of our model with BLEU.

In morphological segmentation experiments, we train a traditional model based on CRF with 12K annotated corpus. We extract the bilingually-constrained features on a Uyghur-Chinese parallel corpus on news domain of about 40K sentences, and monolingual co-occurrence features from the Uyghur parts of the same corpus (Table 5). We use a log-linear model to integrate these features together, which is the framework of our proposed method.

corpus	**size of corpus (sentence pairs)**		
	training set	**dev set**	**test set**
Spoken	30K	1K	1.5K
News	40K	/	/

Table 4: Statistics of data in SMT

4.2 Results and Discussion

Table 6 gives performance of spoken Uyghur morphological segmentation for machine translation using different models. It should be noted that morphological segmentation mentioned in this paper is different from the traditional way. In our approach, we assumed that we already have the segmentation results according to a morphological analyzer; our task is to make a decision that whether to reserve or delete the suffixes of a certain Uyghur word to make a better translation performance. Log-linear model with only the CRF feature do not yield satisfactory results, while our proposed model (log-linear model with CRF, bilingual word alignment and monolingual stem-suffix co-occurrence features) perform significantly better at predicting tags of suffixes on the spoken Uyghur corpus. A single CRF, bilingual word alignment or monolingual stem-suffix co-occurrence feature cannot fully capture bilingual relationship and contexts of current Uyghur word; therefore, performance of these models (LL+XXs) is much lower that our proposed approach.

Table 7& 8 present results on spoken Uyghur-Chinese machine translation with different translation models (Table 7) and segmentation models integrated with different features (Table 8).

From Table 7, we observe that compare with word based model, models based on morphological segmentation achieved better results. That is because segmentation can alleviate data sparsity in model training, especially for low-resource and morphologically rich languages. Although factored translation model can use more linguistic in-

[2]http://www.statmt.org/moses/

[3]http://www.speech.sri.com/projects/srilm/

corpus	size of corpus (sentences/tokens)		
	training set	**dev set**	**test set**
morphseg0	12K/1.2M	/	/
morphseg1	5K/0.5M, 5K/0.6M	0.5K/5K, 0.5K/6K	0.8K/8K, 0.8K/8K
bilingualWA	40K/1M, 40K/1.2M	/	/
monoCO	40K/1M	/	/

Table 5: Statistics of data used in morphological segmentation

models	performance (%)		
	recall	**precision**	**f1**
LL+CRF	76.24	78.51	77.36
LL+bilingualWA	75.69	78.20	76.92
LL+monoCO	74.32	77.94	76.09
LL+bilingualWA+monoCO	78.56	78.62	78.59
LL+CRF+bilingualWA	78.70	79.02	78.86
LL+CRF+monoCO	78.93	80.59	79.75
LL+CRF+bilingualWA+monoCO	**82.40**	**85.37**	**83.86**

Table 6: Performance of morphological segmentation for spoken Uyghur translation using different models

models	BLEU	
	test set	**dev set**
Word based model	15.05	16.81
Factored model	17.18	18.59
Stem based model	18.30	17.52
Stem based model (Ours)	**19.92(+1.62)**	**18.70(+0.11)**

Table 7: Test set and Dev set performance for Uyghur-Chinese spoken machine translation results by models using different translation units.

features	BLEU	
	test set	**dev set**
CRF	18.30	17.52
bilingualWA	17.28	17.03
monoCoOccur	17.14	16.50
bilingualWA+monoCoOccur	18.29	18.50
CRF+bilingualWA	17.80	17.42
CRF+monoCoOccur	17.68	17.01
CRF+bilingualWA+monoCoOccur(Ours)	**19.92(+1.62)**	**18.70(+0.20)**

Table 8: Test set and Dev set performance for Uyghur-Chinese spoken machine translation (BLEU) results by using our segmentation model with different features.

formation, it can't alleviate the data sparsity effectively in spoken Uyghur-Chinese machine translation. Therefore, stem-based models both outperform factored based translation model in test set. Our proposed approach integrated both bilingual information and monolingual information into the morphological segmentation model, that's why it achieves best translation results in both test and dev sets among four translation models. We also found that the performance of dev set of factored model outperforms stem based model, a possible reason is that stem based model lost some information when Uyghur words segment incorrectly.

In Table 8, we list translation results using different features in log-linear model. The CRF model is just the same as the stem based model. A single bilingualWA or monoCoOccur model has a relative lower performance compared with CRF model that is because CRF model learned from annotated corpus, both bilingualWA and monoCoOccur models learn features unsurpervisedly. With CRF features, the performance of bilingualWA and monoCoOccur both improved. Our proposed log-linear based morphological segmentation model achieved best translation results among seven models, one possible reason is that our method integrates manually annotated information, bilingual word alignment information and monolingual suffix-word co-occurrence information; therefore, it can optimize the morphological segmentation of Uyghur in spoken translation situation.

5 Related Work

Many researchers have focused on morphological segmentation in past few years. These methods can be classified into three categories: unsupervised approaches (Goldsmith, 2001) (Creutz and Lagus, 2002) (Creutz et al., 2007) (Poon et al., 2009) (Abudukelimu et al., 2017), supervised approaches (Sirts and Goldwater, 2013) (Ruokolainen et al., 2013) and semi-supervised approaches (Kohonen et al., 2010) (Ruokolainen et al., 2014) (Tursun et al., 2016). Other people also put their efforts on morphological segmentation for SMT, such as (Lee, 2004) (Grönroos et al., 2016) (Mermer and Saraclar, 2011) (Sereewattana, 2003) (Clifton, 2010) (Bisazza and Federico, 2009) (Al-Haj and Lavie, 2012) (Rasooli et al., 2013) (Mi et al., 2015).

Our proposed approach is different from previous works. One of the most important reasons is that most of above studies are focused on formal corpora, such as news, government documents et al., and our study mainly put efforts on spoken language. Moreover, our method integrates supervised feature (CRF feature) and unsupervised features (bilingual word alignment feature and monolingual suffix-word co-occurrence feature) into the log-linear model.

6 Conclusion

In this paper, we present a bilingually - constrained based Uyghur segmentation method to optimize the performance of Uyghur-Chinese spoken translation. Our approach aims to maintain some useful suffixes of Uyghur words to overcome the information loss and data sparsity exist in spoken translation. The proposed method consists of four parts: 1) Uyghur segmentation based on a CRFs model; 2) bilingual alignment features collection from the GIZA++; 3) monolingual co-occurrence features extract from a large Uyghur corpus; and 4) training a log-linear based Uyghur segmentation model, under the bilingual alignment features and monolingual co-occurrence features. The experimental results show that the proposed model can achieve significant BLEU (+1.6) improvements over several baselines in Uyghur-Chinese spoken translation.

In our future work, we plan to integrate linguistic information such as part-of-speech, syntax into our proposed segmentation approach.

Acknowledgments

We sincerely thank the anonymous reviewers for their thorough reviewing and valuable suggestions. This work is supported by the West Light Foundation of The Chinese Academy of Sciences under Grant No.2015-XBQN-B-10, the Xinjiang Key Laboratory Fund under Grant No. 2015KL031, the Xinjiang Science and Technology Major Project under Grant No.2016A03007-3 and the Natural Science Foundation of Xinjiang under Grant No.2015211B034.

References

Halidanmu Abudukelimu, Yong Cheng, Yang Liu, and Maosong Sun. 2017. Uyghur morphological segmentation with bidirectional gru neural networks. *Journal of Tsinghua*

University(Science and Technology) 57(1):1. https://doi.org/10.16511/j.cnki.qhdxxb.2017.21.001.

Hassan Al-Haj and Alon Lavie. 2012. The impact of arabic morphological segmentation on broad-coverage english-to-arabic statistical machine translation. *Machine Translation* 26(1):3–24.

Arianna Bisazza and Marcello Federico. 2009. Morphological pre-processing for turkish to english statistical machine translation. In *Proceedings of IWSLT 2009*. Tokyo, Japan, IWSLT 2009, pages 129–145.

Ann Clifton. 2010. *Unsupervised morphological segmentation for statistical machine translation*. Simon Fraser University.

Mathias Creutz, Teemu Hirsimäki, Mikko Kurimo, Antti Puurula, Janne Pylkkönen, Vesa Siivola, Matti Varjokallio, Ebru Arisoy, Murat Saraçlar, and Andreas Stolcke. 2007. Morph-based speech recognition and modeling of out-of-vocabulary words across languages. *ACM Trans. Speech Lang. Process.* 5(1):3:1–3:29. https://doi.org/10.1145/1322391.1322394.

Mathias Creutz and Krista Lagus. 2002. Unsupervised discovery of morphemes. In *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning - Volume 6*. Association for Computational Linguistics, MPL '02, pages 21–30. https://doi.org/10.3115/1118647.1118650.

John Goldsmith. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics* 27(2):153–198. https://doi.org/10.1162/089120101750300490.

Stig-Arne Grönroos, Sami Virpioja, and Mikko Kurimo. 2016. *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, Association for Computational Linguistics, Berlin, Germany, chapter Hybrid Morphological Segmentation for Phrase-Based Machine Translation, pages 289–295. https://doi.org/10.18653/v1/W16-2312.

Oskar Kohonen, Sami Virpioja, and Krista Lagus. 2010. Semi-supervised learning of concatenative morphology. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*. Association for Computational Linguistics, Uppsala, Sweden, SIGMORPHON '10, pages 78–86. http://dl.acm.org/citation.cfm?id=1870478.1870488.

Young-Suk Lee. 2004. Morphological analysis for statistical machine translation. In *Proceedings of HLT-NAACL 2004: Short Papers*. Association for Computational Linguistics, Boston, Massachusetts, HLT-NAACL-Short '04, pages 57–60. http://dl.acm.org/citation.cfm?id=1613984.1613999.

Yang Liu, Qun Liu, and Shouxun Lin. 2005. Log-linear models for word alignment. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, Ann Arbor, Michigan, ACL '05, pages 459–466. https://doi.org/10.3115/1219840.1219897.

Coskun Mermer and Murat Saraclar. 2011. Unsupervised turkish morphological segmentation for statistical machine translation. In *Workshop of MT and Morphologically-rich Languages.*.

Chenggang Mi, Yating Yang, Rui Dong, Xi Zhou, Lei Wang, Xiao Li, Tonghai Jiang, and Turghun Osman. 2015. Optimized uyghur segmentation for statistical machine translation. In *Proceedings of the 20th International Conference on Applications of Natural Language to Information Systems, NLDB 2015*. Springer International Publishing, Passau, Germany, pages 395–398.

Hoifung Poon, Colin Cherry, and Kristina Toutanova. 2009. Unsupervised morphological segmentation with log-linear models. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Boulder, Colorado, NAACL '09, pages 209–217. http://dl.acm.org/citation.cfm?id=1620754.1620785.

Sadegh Mohammad Rasooli, Ahmed El Kholy, and Nizar Habash. 2013. Orthographic and morphological processing for persian-to-english statistical machine translation. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*. Asian Federation of Natural Language Processing, Nagoya, Japan, pages 1047–1051. http://aclweb.org/anthology/I13-1144.

Teemu Ruokolainen, Oskar Kohonen, Sami Virpioja, and Mikko Kurimo. 2013. *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, Association for Computational Linguistics, Sofia, Bulgaria, chapter Supervised Morphological Segmentation in a Low-Resource Learning Setting using Conditional Random Fields, pages 29–37. http://aclweb.org/anthology/W13-3504.

Teemu Ruokolainen, Oskar Kohonen, Sami Virpioja, and mikko kurimo. 2014. Painless semi-supervised morphological segmentation using conditional random fields. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*. Association for Computational Linguistics, Gothenburg, Sweden, pages 84–89. https://doi.org/10.3115/v1/E14-4017.

Siriwan Sereewattana. 2003. *Unsupervised segmentation for statistical machine translation*. University of Edinburgh.

Kairit Sirts and Sharon Goldwater. 2013. Minimally-supervised morphological segmentation using adaptor grammars. *Transactions of the Association of Computational Linguistics* 1:255–266. http://aclweb.org/anthology/Q13-1021.

Eziz Tursun, Debasis Ganguly, Turghun Osman, Ya-Ting Yang, Ghalip Abdukerim, Jun-Lin Zhou, and Qun Liu. 2016. A semisupervised tag-transition-based markovian model for uyghur morphology analysis. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 16(2):8:1–8:23. https://doi.org/10.1145/2968410.