Stem Translation with Affix-Based Rule Selection for Agglutinative Languages

Zhiyang Wang†, **Yajuan Lü**†, **Meng Sun**†, **Qun Liu**‡†

†Key Laboratory of Intelligent Information Processing
Institute of Computing Technology, Chinese Academy of Sciences
P.O. Box 2704, Beijing 100190, China
{wangzhiyang,lvyajuan,sunmeng,liuqun}@ict.ac.cn
‡Centre for Next Generation Localisation
Faculty of Engineering and Computing, Dublin City University
qliu@computing.dcu.ie

Abstract

Current translation models are mainly designed for languages with limited morphology, which are not readily applicable to agglutinative languages as the difference in the way lexical forms are generated. In this paper, we propose a novel approach for translating agglutinative languages by treating stems and affixes differently. We employ stem as the atomic translation unit to alleviate data spareness. In addition, we associate each stem-granularity translation rule with a distribution of related affixes, and select desirable rules according to the similarity of their affix distributions with given spans to be translated. Experimental results show that our approach significantly improves the translation performance on tasks of translating from three Turkic languages to Chinese.

1 Introduction

Currently, most methods on statistical machine translation (SMT) are developed for translation of languages with limited morphology (e.g., English, Chinese). They assumed that word was the atomic translation unit (ATU), always ignoring the internal morphological structure of word. This assumption can be traced back to the original IBM word-based models (Brown et al., 1993) and several significantly improved models, including phrase-based (Och and Ney, 2004; Koehn et al., 2003), hierarchical (Chiang, 2005) and syntactic (Quirk et al., 2005; Galley et al., 2006; Liu et al., 2006) models. These improved models worked well for translating languages like English with large scale parallel corpora available.

Different from languages with limited morphology, words of agglutinative languages are formed mainly by concatenation of stems and affixes. Generally, a stem can attach with several affixes, thus leading to tens of hundreds of possible inflected variants of lexicons for a single stem. Modeling each lexical form as a separate word will generate high out-of-vocabulary rate for SMT. Theoretically, ways like morphological analysis and increasing bilingual corpora could alleviate the problem of data sparsity, but most agglutinative languages are less-studied and suffer from the problem of resource-scarceness. Therefore, previous research mainly focused on the different inflected variants of the same stem and made various transformation of input by morphological analysis, such as (Lee, 2004; Goldwater and McClosky, 2005; Yang and Kirchhoff, 2006; Habash and Sadat, 2006; Bisazza and Federico, 2009; Wang et al., 2011). These work still assume that the atomic translation unit is word, stem or morpheme, without considering the difference between stems and affixes.

In agglutinative languages, stem is the base part of word not including inflectional affixes. Affix, especially inflectional affix, indicates different grammatical categories such as tense, person, number and case, etc., which is useful for translation rule disambiguation. Therefore, we employ stem as the atomic translation unit and use affix information to guide translation rule selection. Stem-granularity translation rules have much larger coverage and can lower the OOV rate. Affix based rule selection takes advantage of auxiliary syntactic roles of affixes to make a better rule selection. In this way, we can achieve a balance between rule coverage and matching accuracy, and ultimately improve the translation performance.

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 364–369,
Sofia, Bulgaria, August 4-9 2013. ©2013 Association for Computational Linguistics

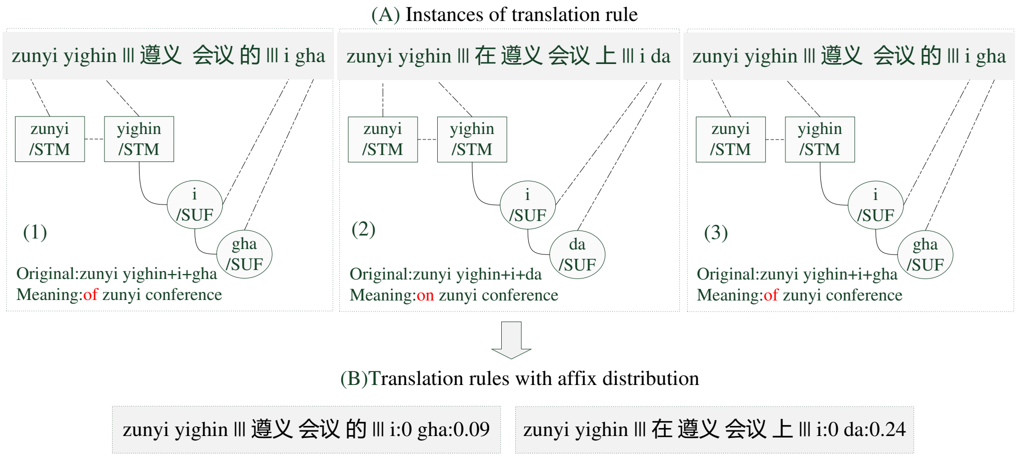
(A) Instances of translation rule
zunyi yighin ||| 遵义 会议 的 ||| i gha
zunyi yighin ||| 在 遵义 会议 上 ||| i da
zunyi yighin ||| 遵义 会议 的 ||| i gha
zunyi /STM
yighin /STM
i /SUF
gha /SUF
da /SUF
(1)
(2)
(3)
Original:zunyi yighin+i+gha
Meaning:of zunyi conference
Original:zunyi yighin+i+da
Meaning:on zunyi conference
Original:zunyi yighin+i+gha
Meaning:of zunyi conference
(B)Translation rules with affix distribution
zunyi yighin ||| 遵义 会议 的 ||| i:0 gha:0.09
zunyi yighin ||| 在 遵义 会议 上 ||| i:0 da:0.24

Figure 1: Translation rule extraction from Uyghur to Chinese. Here tag "/STM" represents stem and "/SUF" means suffix.

2 Affix Based Rule Selection Model

Figure 1 (B) shows two translation rules along with affix distributions. Here a translation rule contains three parts: the source part (on stem level), the target part, and the related affix distribution (represented as a vector). We can see that, although the source part of the two translation rules are identical, their affix distributions are quite different. Affix "gha" in the first rule indicates that something is affiliated to a subject, similar to "of" in English. And "da" in second rule implies location information. Therefore, given a span "zunyi/STM yighin/STM+i/SUF+da/SUF+..." to be translated, we hope to encourage our model to select the second translation rule. We can achieve this by calculating similarity between the affix distributions of the translation rule and the span.

The affix distribution can be obtained by keeping the related affixes for each rule instance during translation rule extraction ((A) in Figure 1). After extracting and scoring stem-granularity rules in a traditional way, we extract stem-granularity rules again by keeping affix information and compute the affix distribution with tf-idf (Salton and Buckley, 1987). Finally, the affix distribution will be added to the previous stem-granularity rules.

2.1 Affix Distribution Estimation

Formally, translation rule instances with the same source part can be treated as a *document collection*[1], so each rule instance in the collection is some kind of *document*. Our goal is to classify the source parts into the target parts on the *document collection* level with the help of affix distribution. Accordingly, we employ vector space model (VSM) to represent affix distribution of each rule instance. In this model, the feature weights are represented by the classic tf-idf (Salton and Buckley, 1987):

$$\mathbf{tf_{i,j}} = \frac{\mathbf{n_{i,j}}}{\sum_{\mathbf{k}} \mathbf{n_{k,j}}} \quad \mathbf{idf_{i,j}} = \mathbf{log}\frac{|\mathbf{D}|}{|\mathbf{j} : \mathbf{a_i} \in \mathbf{r_j}|} \qquad \mathbf{tfidf_{i,j}} = \mathbf{tf_{i,j}} \times \mathbf{idf_{i,j}} \tag{1}$$

where $\mathbf{tfidf_{i,j}}$ is the weight of affix a_i in translation rule instance r_j. $\mathbf{n_{i,j}}$ indicates the number of occurrence of affix a_i in r_j. $|\mathbf{D}|$ is the number of rule instance with the same source part, and $|\mathbf{j} : \mathbf{a_i} \in \mathbf{r_j}|$ is the number of rule instance which contains affix a_i within $|\mathbf{D}|$.

Let's take the suffix "gha" from (A_1) in Figure 1 as an example. We assume that there are only three instances of translation rules extracted from parallel corpus ((A) in Figure 1). We can see that "gha" only appear once in (A_1) and also appear once in whole instances. Therefore, $\mathbf{tf_{gha,(A_1)}}$ is 0.5 and $\mathbf{idf_{gha,(A_1)}}$ is $log(3/2)$. $\mathbf{tfidf_{gha,(A_1)}}$ is the product of $\mathbf{tf_{gha,(A_1)}}$ and $\mathbf{idf_{gha,(A_1)}}$ which is 0.09.

Given a set of $\mathbf{N}$ translation rule instances with the same source and target part, we define the centroid vector $\mathbf{d_r}$ according to the centroid-based classification algorithm (Han and Karypis, 2000),

$$\mathbf{d_r} = \frac{\mathbf{1}}{\mathbf{N}} \sum_{\mathbf{i} \in \mathbf{N}} \mathbf{d_i} \tag{2}$$

[1] We employ concepts from text classification to illustrate how to estimate affix distribution.

Data set	#Sent.	#Type			#Token		
		word	stem	morph	word	stem	morph
UY-CH-Train.	50K	69K	39K	42K	1.2M	1.2M	1.6M
UY-CH-Dev.	0.7K*4	5.9K	4.1K	4.6K	18K	18K	23.5K
UY-CH-Test.	0.7K*1	4.7K	3.3K	3.8K	14K	14K	17.8K
KA-CH-Train.	50K	62K	40K	42K	1.1M	1.1M	1.3M
KA-CH-Dev.	0.7K*4	5.3K	4.2K	4.5K	15K	15K	18K
KA-CH-Test.	0.2K*1	2.6K	2.0K	2.3K	8.6K	8.6K	10.8K
KI-CH-Train.	50K	53K	27K	31K	1.2M	1.2M	1.5M
KI-CH-Dev.	0.5K*4	4.1K	3.1K	3.5K	12K	12K	15K
KI-CH-Test.	0.2K*4	2.2K	1.8K	2.1K	4.7K	4.7K	5.8K

Table 1: Statistics of data sets. $*N$ means the number of reference, $morph$ is short to morpheme. UY, KA, KI, CH represent Uyghur, Kazakh, Kirghiz and Chinese respectively.

$\mathbf{d_r}$ is the final affix distribution.

By comparing the similarity of affix distributions, we are able to decide whether a translation rule is suitable for a span to be translated. In this work, similarity is measured using the cosine distance similarity metric, given by

$$\mathbf{sim}(\mathbf{d_1}, \mathbf{d_2}) = \frac{\mathbf{d_1} \cdot \mathbf{d_2}}{\|\mathbf{d_1}\| \times \|\mathbf{d_2}\|} \qquad (3)$$

where $\mathbf{d_i}$ corresponds to a vector indicating affix distribution, and "·" denotes the inner product of the two vectors.

Therefore, for a specific span to be translated, we first analyze it to get the corresponding stem sequence and related affix distribution represented as a vector. Then the stem sequence is used to search the translation rule table. If the source part is matched, the similarity will be calculated for each candidate translation rule by cosine similarity (as in equation 3). Therefore, in addition to the traditional translation features on stem level, our model also adds the affix similarity score as a dynamic feature into the log-linear model (Och and Ney, 2002).

3 Related Work

Most previous work on agglutinative language translation mainly focus on Turkish and Finnish. Bisazza and Federico (2009) and Mermer and Saraclar (2011) optimized morphological analysis as a pre-processing step to improve the translation between Turkish and English. Yeniterzi and Oflazer (2010) mapped the syntax of the English side to the morphology of the Turkish side with the factored model (Koehn and Hoang, 2007). Yang and Kirchhoff (2006) backed off surface form to stem when translating OOV words of Finnish. Luong and Kan (2010) and Luong et al. (2010) focused on Finnish-English translation through improving word alignment and enhancing phrase table. These works still assumed that the atomic translation unit is word, stem or morpheme, without considering the difference between stems and affixes.

There are also some work that employed the context information to make a better choice of translation rules (Carpuat and Wu, 2007; Chan et al., 2007; He et al., 2008; Cui et al., 2010). all the work employed rich context information, such as POS, syntactic, etc., and experiments were mostly done on less inflectional languages (i.e. Chinese, English) and resourceful languages (i.e. Arabic).

4 Experiments

In this work, we conduct our experiments on three different agglutinative languages, including Uyghur, Kazakh and Kirghiz. All of them are derived from Altaic language family, belonging to Turkic languages, and mostly spoken by people in Central Asia. There are about 24 million people take these languages as mother tongue. All of the tasks are derived from the evaluation of China Workshop of Machine Translation (CWMT)[2]. Table 1 shows the statistics of data sets.

For the language model, we use the SRI Language Modeling Toolkit (Stolcke, 2002) to train a 5-gram model with the target side of training corpus. And phrase-based Moses[3] is used as our

[2] http://mt.xmu.edu.cn/cwmt2011/en/index.html.

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

	UY-CH	KA-CH	KI-CH
word	$31.74_{+0.0}$	$28.64_{+0.0}$	$35.05_{+0.0}$
stem	$\mathbf{33.74}_{\mathbf{+2.0}}$	$\mathbf{30.14}_{\mathbf{+1.5}}$	$35.52_{+0.47}$
morph	$32.69_{+0.95}$	$29.21_{+0.57}$	$34.97_{-0.08}$
affix	$\mathbf{34.34}_{\mathbf{+2.6}}$	$\mathbf{30.19}_{\mathbf{+2.27}}$	$35.96_{+0.91}$

Table 2: Translation results from Turkic languages to Chinese. **word**: ATU is surface form, **stem**: ATU is represented stem, **morph**: ATU denotes morpheme, **affix**: stem translation with affix distribution similarity. BLEU scores in **bold** means significantly better than the baseline according to (Koehn, 2004) for p-value less than 0.01.

baseline SMT system. The decoding weights are optimized with MERT (Och, 2003) to maximum word-level BLEU scores (Papineni et al., 2002).

4.1 Using Unsupervised Morphological Analyzer

As most agglutinative languages are resource-poor, we employ unsupervised learning method to obtain the morphological structure. Following the approach in (Virpioja et al., 2007), we employ the Morfessor[4] Categories-MAP algorithm (Creutz and Lagus, 2005). It applies a hierarchical model with three categories (prefix, stem, and suffix) in an unsupervised way. From Table 1 we can see that vocabulary sizes of the three languages are reduced obviously after unsupervised morphological analysis.

Table 2 shows the translation results. All the three translation tasks achieve obvious improvements with the proposed model, which always performs better than only employ **word**, **stem** and **morph**. For the Uyghur to Chinese translation (UY-CH) task in Table 2, performances after unsupervised morphological analysis are always better than the baseline. And we gain up to +2.6 BLEU points improvements with **affix** compared to the baseline. For the Kazakh to Chinese translation (KA-CH) task, the improvements are also significant. We achieve +2.27 and +0.77 improvements compared to the baseline and **stem**, respectively. As for the Kirghiz to Chinese translation (KI-CH) task, improvements seem relative small compared to the other two language pairs. However, it also gains +0.91 BLEU points over the baseline.

UY		Unsup	Sup
stem	#Type	39K	21K
	#Token	1.2M	1.2M
affix	#Type	3.0K	0.3K
	#Token	0.4M	0.7M

Table 3: Statistics of training corpus after unsupervised(Unsup) and supervised(Sup) morphological analysis.

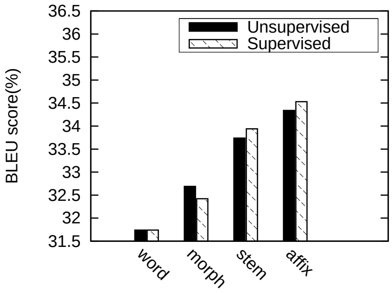

Figure 2: Uyghur to Chinese translation results after unsupervised and supervised analysis.

4.2 Using Supervised Morphological Analyzer

Taking it further, we also want to see the effect of supervised analysis on our model. A generative statistical model of morphological analysis for Uyghur was developed according to (Mairehaba et al., 2012). Table 3 shows the difference of statistics of training corpus after supervised and unsupervised analysis. Supervised method generates fewer type of stems and affixes than the unsupervised approach. As we can see from Figure 2, except for the **morph** method, **stem** and **affix** based approaches perform better after supervised analysis. The results show that our approach can obtain even better translation performance if better morphological analyzers are available. Supervised morphological analysis generates more meaningful morphemes, which lead to better disambiguation of translation rules.

5 Conclusions and Future Work

In this paper we propose a novel framework for agglutinative language translation by treating stem and affix differently. We employ the stem sequence as the main part for training and decoding. Besides, we associate each stem-granularity translation rule with an affix distribution, which could be used to make better translation decisions by calculating the affix distribution similarity be-

[4] http://www.cis.hut.fi/projects/morpho/

tween the rule and the instance to be translated. We conduct our model on three different language pairs, all of which substantially improved the translation performance. The procedure is totally language-independent, and we expect that other language pairs could benefit from our approach.

Acknowledgments

The authors were supported by 863 State Key Project (No. 2011AA01A207), and National Key Technology R&D Program (No. 2012BAH39B03), Key Project of Knowledge Innovation Program of Chinese Academy of Sciences (No. KGZD-EW-501). Qun Liu's work is partially supported by Science Foundation Ireland (Grant No.07/CE/I1142) as part of the CNGL at Dublin City University. We would like to thank the anonymous reviewers for their insightful comments and those who helped to modify the paper.

References

Arianna Bisazza and Marcello Federico. 2009. Morphological pre-processing for Turkish to English statistical machine translation. In *Proceedings of IWSLT*, pages 129–135.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.*, 19(2):263–311.

Marine Carpuat and Dekai Wu. 2007. Improving statistical machine translation using word sense disambiguation. In *Proceedings of EMNLP-CoNLL*, pages 61–72.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. Word sense disambiguation improves statistical machine translation. In *Proceedings of ACL*, pages 33–40.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of ACL*, pages 263–270.

Mathias Creutz and Krista Lagus. 2005. Inducing the morphological lexicon of a natural language from unannotated text. In *Proceedings of AKRR*, pages 106–113.

Lei Cui, Dongdong Zhang, Mu Li, Ming Zhou, and Tiejun Zhao. 2010. A joint rule selection model for hierarchical phrase-based translation. In *Proceedings of ACL, Short Papers*, pages 6–11.

Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer. 2006. Scalable inference and training of context-rich syntactic translation models. In *Proceedings of COLING/ACL*, pages 961–968.

Sharon Goldwater and David McClosky. 2005. Improving statistical MT through morphological analysis. In *Proceedings of HLT-EMNLP*, pages 676–683.

Nizar Habash and Fatiha Sadat. 2006. Arabic preprocessing schemes for statistical machine translation. In *Proceedings of NAACL, Short Papers*, pages 49–52.

Eui-Hong Sam Han and George Karypis. 2000. Centroid-based document classification: analysis experimental results. In *Proceedings of PKDD*, pages 424–431.

Zhongjun He, Qun Liu, and Shouxun Lin. 2008. Improving statistical machine translation using lexicalized rule selection. In *Proceedings of COLING*, pages 321–328.

Philipp Koehn and Hieu Hoang. 2007. Factored translation models. In *Proceedings of EMNLP-CoNLL*, pages 868–876.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of NAACL*, pages 48–54.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP*, pages 388–395.

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

Yang Liu, Qun Liu, and Shouxun Lin. 2006. Tree-to-string alignment template for statistical machine translation. In *Proceedings of COLING-ACL*, pages 609–616.

Minh-Thang Luong and Min-Yen Kan. 2010. Enhancing morphological alignment for translating highly inflected languages. In *Proceedings of COLING*, pages 743–751.

Minh-Thang Luong, Preslav Nakov, and Min-Yen Kan. 2010. A hybrid morpheme-word representation for machine translation of morphologically rich languages. In *Proceedings of EMNLP*, pages 148–157.

Aili Mairehaba, Wenbin Jiang, Zhiyang Wang, Yibulayin Tuergen, and Qun Liu. 2012. Directed graph model of Uyghur morphological analysis. *Journal of Software*, 23(12):3115–3129.

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

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of ACL*, pages 295–302.

Franz Josef Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Comput. Linguist.*, pages 417–449.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of ACL*, pages 160–167.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318.

Chris Quirk, Arul Menezes, and Colin Cherry. 2005. Dependency treelet translation: syntactically informed phrasal SMT. In *Proceedings of ACL*, pages 271–279.

Gerard Salton and Chris Buckley. 1987. Term weighting approaches in automatic text retrieval. Technical report.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proceedings of ICSLP*, pages 311–318.

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*, pages 491–498.

Zhiyang Wang, Yajuan Lü, and Qun Liu. 2011. Multi-granularity word alignment and decoding for agglutinative language translation. In *Proceedings of MT SUMMIT*, pages 360–367.

Mei Yang and Katrin Kirchhoff. 2006. Phrase-based backoff models for machine translation of highly inflected languages. In *Proceedings of EACL*, pages 1017–1020.

Reyyan Yeniterzi and Kemal Oflazer. 2010. Syntax-to-morphology mapping in factored phrase-based statistical machine translation from English to Turkish. In *Proceedings of ACL*, pages 454–464.