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 sparseness. 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.
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):

\[
\text{tf}_{1j} = \frac{n_{1j}}{\sum_k n_{kj}} \quad \text{idf}_{1j} = \log \frac{|D|}{|\{ a_i \in r_j \}|} \quad \text{tfidf}_{1j} = \text{tf}_{1j} \times \text{idf}_{1j}
\]

where \( \text{tfidf}_{1j} \) is the weight of affix \( a_i \) in translation rule instance \( r_j \), \( n_{1j} \) indicates the number of occurrence of affix \( a_i \) in \( r_j \), \(|D|\) is the number of rule instance with the same source part, and \(|\{ a_i \in r_j \}|\) is the number of rule instance which contains affix \( a_i \) within \(|D|\).

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, \( \text{tf}_{\text{gha},(A_1)} \) is 0.5 and \( \text{idf}_{\text{gha},(A_1)} \) is \( \log(3/2) \). \( \text{tfidf}_{\text{gha},(A_1)} \) is the product of \( \text{tf}_{\text{gha},(A_1)} \) and \( \text{idf}_{\text{gha},(A_1)} \) which is 0.09.

Given a set of \( N \) 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),

\[
\mathbf{d}_r = \frac{1}{N} \sum_{i \in N} \mathbf{d}_i
\]
### Table 1: Statistics of data sets.

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

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, Chinese respectively.

\[ \text{sim}(d_1, d_2) = \frac{d_1 \cdot d_2}{\|d_1\| \times \|d_2\|} \]  

where \(d_1\) corresponds to a vector indicating affix distribution, and "\(\cdot\)" 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 Offazer (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


\(^3\)http://www.statmt.org/moses/
Table 2: Translation results from Turkic languages to Chinese. **word**: ATU is surface form, **stem**: ATU is represented stem, **morph**: ATU denotes morpheme, **suffix**: stem translation with affix distribution similarity. BLEU scores in **bold** mean significantly better than the baseline according to (Koehn, 2004) for p-value less than 0.01.

<table>
<thead>
<tr>
<th></th>
<th>UY-CH</th>
<th>KA-CH</th>
<th>KI-CH</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>word</strong></td>
<td>31.74±0.0</td>
<td>28.64±0.0</td>
<td>35.05±0.0</td>
</tr>
<tr>
<td><strong>stem</strong></td>
<td>33.74±2.0</td>
<td>30.14±1.5</td>
<td>35.52±0.47</td>
</tr>
<tr>
<td><strong>morph</strong></td>
<td>32.69±0.95</td>
<td>29.21±0.57</td>
<td>34.97±0.08</td>
</tr>
<tr>
<td><strong>affix</strong></td>
<td>34.34±2.6</td>
<td>30.19±2.27</td>
<td>35.96±0.91</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>UY</th>
<th>Unsup</th>
<th>Sup</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>stem</strong></td>
<td>#Type</td>
<td>39K</td>
<td>21K</td>
</tr>
<tr>
<td></td>
<td>#Token</td>
<td>1.2M</td>
<td>1.2M</td>
</tr>
<tr>
<td><strong>affix</strong></td>
<td>#Type</td>
<td>3.0K</td>
<td>0.3K</td>
</tr>
<tr>
<td></td>
<td>#Token</td>
<td>0.4M</td>
<td>0.7M</td>
</tr>
</tbody>
</table>

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

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.

\(^4\)http://www.cis.hut.fi/projects/morpho/

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-
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


