Postprint: A Translation Rule Selection Method for Morphologically Rich Languages

Authors: Wang Zhiyang, LÜ Yajuan, Sun Meng, Wenbin Jiang, Qun Liu

Date: 2017-03-10T00:00:00+00:00

Abstract

Current machine translation models are primarily designed for morphologically simple languages (e.g., English) and are not well-suited for morphologically rich languages (e.g., Uyghur). In this paper, we propose a novel translation rule selection method for morphologically rich languages by explicitly distinguishing between stems and affixes. We employ stems as the basic translation unit to mitigate the data sparsity problem; additionally, each stem-level translation rule is augmented with an affix distribution. During translation, more appropriate rules are selected by computing the similarity between the affix distribution of the source segment and that of the translation rule. Experimental results on translation from three morphologically rich languages (Uyghur, Kazakh, and Kyrgyz) to Chinese demonstrate that the proposed method significantly improves translation quality.

Full Text

Preamble

Vol. 11 No. 4 Information Technology Letters Vol.11 No.4
A Translation Rule Selection Method for Morphologically Rich Languages
Wang Zhiyang, Lü Yajuan, Sun Meng, Jiang Wenbin, Liu Qun

Abstract

Current machine translation models are designed for languages with simple morphological changes (such as English) and are not well-suited for morphologically rich languages (such as Uyghur). In this paper, we propose a novel translation rule selection method for morphologically rich languages by treating stems and affixes differently. We use stems as the basic translation unit to alleviate data sparsity problems. Additionally, each stem-level translation rule is associated

with an affix distribution. During translation, we select more appropriate translation rules by computing the similarity between the affix distribution of the source segment and that of the translation rule. Translation experiments from three morphologically rich languages (Uyghur, Kazakh, and Kyrgyz) to Chinese demonstrate that this method significantly improves translation quality.

Keywords: machine translation, morphologically rich languages, affix distribution, similarity, dynamic features

1 Introduction

Morphologically rich languages are those with complex and abundant morphological changes. From a morphological perspective, languages can be classified into isolating languages, inflectional languages, agglutinative languages, and polysynthetic languages. In fact, apart from isolating languages and a few inflectional languages, the vast majority of languages are morphologically rich. Among China' s minority languages, Uyghur and Mongolian, as well as the official languages of most neighboring countries, fall into this category.

The most prominent characteristic of morphologically rich languages is their complex morphological changes. We use Uyghur as an example to illustrate the morphological features of such languages. Table 1 lists common morphological change patterns in morphologically rich languages. Inflectional changes refer to the addition of affixes to stems, which alters grammatical functions and simultaneously changes word spelling. For instance, adding the third-person singular suffix "si" to the noun "doppa" (hat) transforms it into "doppasi" (his hat).

Agreement refers to the correspondence between different parts of a sentence or phrase. To maintain consistency with relevant grammatical relations, word forms must be changed accordingly. When expressing "I read the newspaper," the first-person singular suffix "män" must be added after the verb "oqu" (read) to maintain agreement. Additionally, there are compound changes, where two words combine to generate new words with different meanings. For example, the nouns "tax" (stone) and "paqa" (frog) combine to form "taxpaqa," meaning "turtle." Vowel harmony is a common phenomenon in phonographic writing systems. When different syllables combine, certain letters must undergo changes (epenthesis, deletion, weakening, etc.) to conform to pronunciation patterns. This series of rich morphological changes can generate hundreds or even thousands of new word forms from a single stem. Modeling each word form as a separate word leads to severe data sparsity problems, posing a significant challenge to traditional statistical machine translation models.

Although machine translation research has been conducted for many years in various countries, it has primarily focused on languages like English, with relatively few studies targeting morphologically rich languages. In research involving translation from morphologically rich languages to Chinese, most studies have simply applied methods that performed well for English and other lan-

guages. However, due to the unique characteristics of morphologically rich languages, translation results have been unsatisfactory.

Moreover, the most successful statistical machine translation methods require large-scale bilingual parallel corpora for training. For translation between morphologically rich languages and Chinese, the lack of large-scale bilingual parallel corpora makes it difficult for pure statistical methods to achieve ideal results. On the other hand, most morphologically rich languages have relatively weak existing language processing foundations, with limited research and a lack of practical morphological analysis and syntactic parsing tools. Annotated corpora for morphological and syntactic analysis are also extremely limited.

China is a multi-ethnic country with 56 ethnic groups coexisting in diversity. In addition to the Han majority, ethnic minorities such as Uyghur, Mongolian, and Kazakh have their own written languages, which are widely used within their communities. Among these, Uyghur, Mongolian, and Kazakh are all morphologically rich languages. Among China' s 21 neighboring countries, the official languages of most have relatively rich morphological changes, such as Russian, Japanese, Korean, Indonesian, Malay, and Hindi (with Russian and Korean also being minority languages in China). Among China' s 21 neighboring countries, 16 use morphologically rich languages as official languages either entirely or partially, accounting for as high as 76% of the total. Therefore, research on translation from morphologically rich languages to Chinese has practical significance. Studying machine translation between morphologically rich languages and Chinese can promote multicultural exchanges across regions and strengthen cooperation in various fields such as economy, culture, and education.

In the following sections, we first describe the current state of research on morphologically rich language translation (§2), then introduce in detail our proposed translation rule selection method based on affix disambiguation (§3). After presenting the model, §4 provides a detailed description and analysis of experimental results, and we conclude with a summary and outlook (§5).

2 Related Work

In most natural language processing tasks, words serve as the atomic units of knowledge representation. In statistical machine translation, words are also treated as atomic translation units without considering internal morphological structure. Starting from word-based translation models [1], through subsequent improved phrase models [2], hierarchical phrase models [3], and syntactic models [4-6], this assumption has been maintained. Given sufficiently large bilingual corpora, these improved models achieve good results when translating isolating languages (such as Chinese) and languages with limited morphological changes (such as English). However, for morphologically rich languages, a single stem can be affixed with multiple affixes (prefixes or suffixes), generating hundreds or even thousands of new surface forms. Modeling each word form with the same stem as a separate word leads to severe data sparsity. For example, the

Mongolian verb root "UILED" (do) theoretically has at least 1,700 different inflected forms [7].

There are three different translation granularities for morphologically rich languages. The first is word-level: even word forms with the same stem are modeled as separate words. Using word-level translation can extract more accurate translation rules, but under limited corpus size, data sparsity seriously affects alignment and translation quality. The second is stem-level: the stem is the part of a word after removing inflectional affixes, expressing the basic meaning of the word. Stem-level translation rules have greater coverage, but discarding some useful affixes introduces ambiguity problems. The final granularity is morpheme-level: morphemes are the smallest meaningful units in word formation. Treating each morpheme as a separate translation unit increases the number of elements constituting a sentence, burdening word alignment and translation decoding.

Academic research on morphologically rich language translation (involving languages such as German, Spanish, Arabic, Hindi, Czech, and Finnish) began early. Related research can be divided into three categories.

The first category addresses data sparsity through morphological analysis as preprocessing to improve translation quality. Goldwater and McClosky [8] experimented with various morpheme combination strategies for Czech, improving Czech-to-English translation quality. Popovic and Ney [9] improved Spanish-to-English and Serbian-to-English translation quality under limited data conditions by replacing Spanish adjectives with their roots and all Serbian words with their roots. Habash and Sadat [10] used different morphological segmentation strategies in preprocessing for Arabic-to-English translation. Experiments show that finer morphological segmentation is not always better; an appropriate translation granularity must be determined empirically. Lee [11] introduced bilingual information to select appropriate granularity for representing inputs, balancing morphological variation differences between two languages. Yang and Kirchhoff [12] degraded out-of-vocabulary (OOV) words to their stems for translation. Some studies also addressed compound changes by decomposing compound words [13] to improve translation. Other related work expanded input information, such as using lattice structures [14] and paraphrase [15] for fault-tolerant translation.

The second category fully utilizes morphological and syntactic information, jointly incorporating multiple factors to guide translation. A representative example is the factored model in the open-source translation system Moses [16]. This model generates corresponding part-of-speech and morphological information while generating target translations, then uses higher-order POS N-gram models and factorized morphological N-gram models to optimize target word selection. POS tags, case, and even supertags [17] can be added as factors to improve translation effects. However, for most morphologically rich languages, high-quality processing tools (such as POS taggers and CCG parsers) are currently unavailable. Some work goes further by using syntactic parsing to pre-reorder the source language (morphologically rich language) to better match

target language word order, with representative work including [18][19]. Additionally, to overcome morphological variation differences between languages, Yeniterzi and Oflazer [20] attempted to parse English and restructure it to be more similar to Turkish for English-to-Turkish translation. Ramanathan et al. [21] deeply mined corresponding knowledge from the English side and mapped it to the Hindi side to improve English-to-Hindi translation. Such methods can significantly improve translation quality, but 前提是必须有相应的句法分析工具可供使用。

The third category of research aims to overcome difficulties caused by the scarcity of bilingual parallel corpus resources for most morphologically rich languages. Common international practices involve using resources from similar languages or employing pivot languages for translation [22][23]. However, for most morphologically rich languages, such resources are also extremely scarce. Therefore, borrowing from similar language resources and using pivot languages are not very applicable.

Overall, current research on morphologically rich language translation primarily targets languages that are not severely resource-poor, leveraging language processing tools such as morphological and syntactic analyzers to improve translation quality. However, the vast majority of morphologically rich languages have limited bilingual resources and lack corresponding language processing tools.

3 Translation Rule Selection Method Based on Affix Disambiguation

Affixes, especially inflectional affixes, express grammatical meanings such as person, tense, number, and case changes, which play an important role in accurately describing translation rules. Therefore, when extracting stem-level translation rules, we simultaneously preserve corresponding affix information.

3.1 Translation Rule Representation

Figure 1 Figure 1: see original paper shows two examples of Uyghur-to-Chinese translation rules. Translation rule instances (1) and (3) are identical, indicating that the same rule instances were extracted from different bilingual sentence pairs. The affix distribution is represented using the classic Vector Space Model (VSM). From the figure, we can see that although the source sides of these two translation rules are the same, their affix distributions differ significantly. The suffix "gha" in the first rule is a dative case in Uyghur, indicating possession, similar to the English preposition "of." The suffix "da" in the second rule is a locative case, indicating location information. This difference in affixes is directly reflected in the target phrases. Therefore, when the segment to be translated is "zunyi/STM yihin/STM+i/SUF+da/SUF/+⋯" [1], assuming the source stem sequences match, we would prefer the model to select the second

[1]Where STM denotes stem and SUF denotes suffix.

translation rule. We can encourage the selection of more appropriate target rules by computing the similarity between the affix distribution of the segment to be translated and that of candidate translation rules.

3.2 Rule Extraction and Parameter Estimation

The translation rule extraction process is as follows:

1. The source language side is represented as stems (Uyghur), while the target language side remains words (Chinese). Alignment and rule extraction are then performed, ultimately obtaining stem-word translation rules and corresponding probability scores.
2. The source language side is represented as stem+affix combinations, with the target language side as words. Using the stem-word alignment results from step 1 (as mentioned earlier, each word in Uyghur contains only one stem), stem-level rule extraction is performed while preserving corresponding affix information in the rule instances.
3. Based on the rule instances extracted in step 2, the Vector Space Model is used to estimate parameters for affix distributions (detailed below) to obtain the affix distribution for each rule.
4. The translation rules from steps 1 and 3 are merged, primarily by adding affix distributions to the original translation rule table to obtain the final translation rules.

As mentioned, affix distributions are represented in vector form. Here we elaborate on how to obtain the vector representation of affix distributions. We treat translation rules with the same source side as a "document collection," where each translation rule in the "collection" is a "document." Our goal is to use affix distribution information to classify each "document" into its corresponding target phrase. This can be divided into three steps:

First, while extracting stem-level translation rules, corresponding affix information is preserved. In Figure 1(A), from the original Uyghur forms, we can see that the corresponding stem sequences constitute the source side of the translation rules, while the remaining affix sequences and their counts are also retained.

Second, rules with the same source side form a collection. Within this collection, we can use classic TF-IDF [2] to represent the weights of relevant affixes.

Finally, within the same collection, we need to aggregate translation rules with the same target side. Here we use a centroid-based classification algorithm [24] to represent the final affix distribution results:

where N denotes the number of rules with the same target side, and drule is obtained by averaging the affix distributions of the same target side.

[2]Term frequency–inverse document frequency, a commonly used weighting technique in information retrieval and data mining.

Thus, for a segment to be translated, we first obtain its stem sequence and affix distribution (represented as a vector) through morphological analysis. The stem sequence is used to retrieve the translation rule table to obtain translation candidates. When the source stem sequence matches successfully, we then compute the similarity between the affix distribution of the segment to be translated and that of candidate translation rules. In this paper, similarity sim is measured by the cosine of the angle between vectors:

The affix distribution similarity score is added as a dynamic feature to the log-linear model [25] to guide the decoder in selecting more appropriate translation rules.

3.3 Selecting Effective Affixes

Affix distributions play a significant role in helping the decoder select more appropriate translation rules. However, affixes are often obtained through monolingual morphological analysis, and the resulting affix set may not be suitable for machine translation. Intuitively, if we consider target language information simultaneously and use bilingual constraints to generate affixes, we may obtain a more suitable affix set for machine translation.

To obtain more useful affixes and discard useless ones (similar to stop word lists in text classification), we propose a discriminative method for obtaining an appropriate affix set.

Given morpheme-level alignment results, we can determine how to handle each affix. If the current affix aligns to the same target word as the previous affix, these two affixes should be merged into one (called "merge"). If the current affix aligns to a different target word than the previous affix, it should be kept separately (called "keep"). When it aligns to null, it should be deleted (called "delete").

In Figure 2 [Figure 2: see original paper], the suffix "gha" aligns to the same target word as the previous suffix "lar," so they should be merged to form a new affix. The suffix "qilidu" aligns to a different target word than previous morphemes and should be kept. The "i" aligns to null and should be deleted directly. That is, within the same word, its constituent affixes can be divided into three categories: merge, keep, and delete. Classification instances can be directly obtained from morpheme-level aligned corpora. To obtain a classification model, we choose Conditional Random Fields (CRF) [26] to train on the instances. CRF is a discriminative probabilistic model that can compute the conditional probability of output state sequences given observation sequences, commonly used for sequence labeling problems. This model does not require the strict independence assumptions of Hidden Markov Models (HMM) [27] and can incorporate arbitrary features. Moreover, it does not suffer from the label bias problem of Maximum Entropy Models [28], as it solves for the globally optimal output state sequence conditional probability given the current observation sequence.

Specifically, the CRF tool used in this experiment is the open-source software CRF++ [3]. Table 2 shows the feature templates used for training the classification model. In addition to neighbor window morpheme features, we also use position information (beginning, middle, end of word, and single-morpheme word) of morphemes within words. Position information is introduced primarily to preserve internal word structure information. Assuming we are considering the suffix "gha" in Figure 2, we use B, M, E, S to denote position information: beginning, middle, end, and single morpheme. To obtain better classification results, this method can be trained iteratively.

4 Experiments

To verify the effectiveness of our proposed method, we conducted translation experiments on three language pairs: Uyghur-Chinese, Kazakh-Chinese, and Kyrgyz-Chinese. Uyghur, Kazakh, and Kyrgyz are minority languages widely used in western China, all belonging to the Turkic language family of the Altaic language system, with exceptionally rich morphological changes. The relevant corpora were obtained from the China Workshop of Machine Translation (CWMT) [4] translation evaluation. Note that since CWMT evaluations are progress tests, we could not obtain the test sets used in the evaluation; the test sets used here were constructed by ourselves. Table 3 shows corpus statistics, where numbers after "*" indicate the number of reference translations. From the table, we can see that after morphological analysis, the vocabulary size of all three morphologically rich languages is significantly reduced, alleviating data sparsity problems.

For language models, we used the SRI language model training tool SRILM [29] to train 5-gram language models based on the target side of the training corpus. The Moses phrase-based system [5] served as the baseline system, with feature weights tuned using Minimum Error Rate Training [30] to maximize word-level BLEU scores [31]. To dynamically incorporate similarity features into the log-linear model, we rebuilt the Moses phrase-based system to dynamically compute affix distribution similarity.

As mentioned earlier, for most morphologically rich languages, corpus and tool resources are relatively scarce, making high-quality morphological analysis tools difficult to obtain. Therefore, we used unsupervised morphological analysis methods for morphological analysis of the languages used, to better verify the language-independent nature of our method. Similar to [32], we also used the unsupervised analysis tool Morfessor [6] developed by the University of Helsinki. To simulate resource-poor languages, we did not perform vowel harmony restoration. Morfessor generates morphological segmentation results based on Minimum Description Length (MDL). Following [33], we classify Morfessor-generated

[3] http://crfpp.sourceforge.net/
[4] http://mt.xmu.edu.cn/cwmt2011/en/index.html
[5] http://www.statmt.org/moses/
[6] http://www.cis.hut.fi/projects/morpho/

"morphemes" (called morphs in the paper, unsupervised minimal segmentation units; different from morphemes in the linguistic sense) into three categories: prefix (PRE), stem (STM), and suffix (SUF), based on which we distinguish stems and affixes. In experiments, we selected the top 5,000 most frequent words in the training corpus to train Morfessor' s segmentation model.

4.1 Experimental Results and Analysis

Table 4 shows translation results for the three Turkic languages to Chinese. The "word" method is the baseline system using word-level granularity as the atomic translation unit. The "stem" method uses corresponding stems instead of words during translation. The "morph" method uses morphemes as the minimal translation unit. The "affix" method corresponds to our proposed approach of using stems for translation and affix distributions for rule selection. The "CRF-affix" method shows results after selecting more useful affixes based on the CRF model. Bolded results indicate statistically significant [34] improvements over the baseline system. The table shows that for all three languages, using stems as the minimal translation unit performs better than using words or morphemes, while our proposed method outperforms stem-level translation.

Specifically, in the Uyghur-to-Chinese translation task, the CRF-affix method achieves a 2.9 percentage point BLEU improvement over the baseline system and a 0.9 percentage point improvement over stem-level translation. In Kazakh-to-Chinese translation, improvements are also significant, with 2.6 and 1.1 percentage point BLEU improvements over the baseline and stem translation, respectively. Additionally, after selecting more useful affixes using the CRF model, there is a 0.33 percentage point improvement over the unprocessed version. For Kyrgyz-to-Chinese translation, the improvement is slightly smaller than the first two language pairs but still achieves a 1.22 percentage point improvement. Using the CRF model to select affix sets brings certain translation quality improvements across all three languages, but overall the improvement magnitude is not large. One possible reason is that whether to merge, keep, or delete affixes depends on morpheme-level alignment, especially the alignment results of affixes themselves, which are often unsatisfactory. As future work, we hope to first improve affix alignment quality to obtain more accurate affix classification instances, thereby improving affix set selection results.

Through observation and analysis of translation results, we find that compared with the baseline system, our model generates more fluent translations. Specifically, improvements are mainly manifested in two aspects:

- **Reduced OOV rate**: Since we use stems as atomic translation units, word forms with the same stem are all represented by their stem, effectively alleviating data sparsity problems. In Example 1 of Table 5 , although the word "qutquzishi" does not exist in our training corpus, many word forms with "qutquz" as the stem do exist. Therefore, when using stem-level translation, "qutquzishi" becomes "qutquz," enabling translation. As can

be seen, stem-level translation can significantly reduce the OOV rate.

- **More appropriate word selection**: In the two examples in Table 5, introducing affix distributions for disambiguation can select more appropriate words, generating translation results that better conform to grammar. Example 1 generates the matching measure word "名" (measure word for people), while Example 2 includes the corresponding preposition "向" (to/toward).

4.2 Impact of Morphological Analysis Quality

All the above experiments were conducted on unsupervised morphological analysis results and effectively improved translation quality. Furthermore, we wanted to verify how morphological analysis quality would affect our method. We built a supervised Uyghur morphological analysis tool using the method proposed in [35] and tested its translation effects. Figure 3 [Figure 3: see original paper] compares the results with those from the unsupervised analysis tool Morfessor. It can be seen that after obtaining stems and affixes using the supervised morphological analysis tool, except for morpheme-level results, both stem-level and affix disambiguation module results are better than translation results after unsupervised analysis, with improvements of about 0.2 to 0.38 percentage points. This demonstrates that our proposed method' s translation quality improves as morphological analysis quality improves. Table 6 shows statistical differences between the results of the two morphological analysis methods. The supervised analysis method generates fewer morpheme types, especially affixes—only 1/10 of those generated by Morfessor. The generated affixes are more grammatically meaningful and can better guide translation rule selection.

4.3 Impact of Corpus Size

Additionally, we conducted experiments on a larger-scale Uyghur-Chinese parallel corpus to verify the effectiveness of our method on relatively large-scale corpora. We randomly divided approximately 300,000 government news Uyghur-Chinese sentence pairs into six parts: 50k, 100k, 150k, 200k, 250k, and 300k, to verify how corpus size affects translation quality. The same development and test sets were used, ensuring no overlap with the training sets. Figure 4 [Figure 4: see original paper] shows the translation curves. It can be seen that regardless of corpus size, the stem translation method based on affix disambiguation consistently performs best: the improvement is significant when the corpus is small; as the corpus size increases, the improvement margin slightly decreases. Even so, when using all 300k bilingual sentence pairs, the method still achieves a 0.7 percentage point improvement.

5 Conclusion and Outlook

In this paper, we propose a novel translation rule selection method for morphologically rich languages by treating stems and affixes differently. Throughout

the translation pipeline, we use stems as atomic translation units. Additionally, each stem-level rule is associated with a corresponding affix distribution. By computing the similarity between the affix distribution of the segment to be translated and that of translation rules, we help the decoder select more appropriate translation rules. Experiments on three different morphologically rich languages demonstrate that this method significantly improves translation quality, especially when bilingual corpora are relatively scarce.

This paper represents the first work to study morphologically rich language translation by differentiating stems and affixes. The method is independent of specific language pairs. We plan to verify our results on more morphologically rich languages in the future and improve translation quality. Furthermore, stems here are similar to content words, while affixes correspond to function words. By this analogy, our method should also be applicable to translating languages with less rich morphological changes, such as English.

References

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

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

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

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

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

[6] 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.

[7] 那顺乌日图，刘群，巴达玛放德斯尔. 面向机器翻译的蒙古语生成. 全国第六届计算语言学联合学术会议论文集, 2001.

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

[9] Maja Popovic and Hermann Ney. Statistical machine translation with a small amount of bilingual training data. In LREC workshop on Minority Language, pages 25–29, 2006.

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

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

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

[13] Philipp Koehn and Kevin Knight. Empirical methods for compound splitting. Proceedings of EACL, pages 187–193, 2003.

[14] Christopher Dyer, Smaranda Muresan, and Philip Resnik. Generalizing word lattice translation. In Proceedings of ACL: HLT, pages 1012–1020, 2008.

[15] Preslav Nakov and Hwee Tou Ng. Translating from morphologically complex languages: A paraphrase-based approach. In Proceedings of ACL: HLT, pages 1298–1307, 2011.

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

[17] Alexandra Birch, Miles Osborne, and Philipp Koehn. CCG supertags in factored statistical machine translation. In Proceedings of StatMT, pages 9–16, 2007.

[18] Peng Xu, Jaeho Kang, Michael Ringgaard, and Franz Josef Och. Using a dependency parser to improve smt for subject-object-verb languages. In Proceedings of NAACL, pages 245–253, 2009.

[19] Dmitriy Genzel. Automatically learning source-side reordering rules for large scale machine translation. In Proceedings of COLING, pages 376–384, 2010.

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

[21] Ananthakrishnan Ramanathan, Hansraj Choudhary, Avishek Ghosh, and Pushpak Bhattacharyya. Case markers and morphology: Addressing the crux of the fluency problem in English-Hindi SMT. In Proceedings of ACL, pages 800–808, 2009.

[22] Preslav Nakov and Hwee Tou Ng. Improved statistical machine translation for resource-poor languages using related resource-rich languages. In Proceedings of EMNLP, pages 1358–1367, 2009.

[23] Pidong Wang, Preslav Nakov, and Hwee Tou Ng. Source language adaptation for resource-poor machine translation. In Proceedings of EMNLP, pages 286–296, 2012.

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

[25] Franz Josef Och and Hermann Ney. Improved statistical alignment models. In Proceedings of ACL, pages 440–447, 2000.

[26] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of ICML, pages 282–289, 2001.

[27] Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In Proceedings of IEEE, pages 257–286, 1989.

[28] Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. A maximum entropy approach to natural language processing. Comput. Linguist., 22(1):39–71, 1996.

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

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

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

[32] Sami Virpioja, Jaakko J. Väyrynen, Mathias Creutz, and Markus Sadeniemi. Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. In Proceedings of MT SUMMIT, pages 491–498, 2007.

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

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

[35] 麦热哈巴. 艾力，姜文斌，王志洋，吐尔根. 依布拉音，刘群. 维吾尔语词法分析的有向图模型. 软件学报, 23(12):3115–3129, 2012.

Authors:

Wang Zhiyang: Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Ph.D. student, wangzhiyang@ict.ac.cn

Lü Yajuan: Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Associate Professor

Sun Meng: Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Master' s student

Jiang Wenbin: Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Assistant Professor

Liu Qun: Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Professor

Note: Figure translations are in progress. See original paper for figures.