LETTER

A Morpheme-Based Weighting for Chinese-Mongolian Statistical Machine Translation

Zhenxin YANG[†,††a)], ***Member***, **Miao LI**[†], **Lei CHEN**[†], ***and*** **Kai SUN**[††], ***Nonmembers***

SUMMARY In this paper, a morpheme-based weighting and its integration method are proposed as a smoothing method to alleviate the data sparseness in Chinese-Mongolian statistical machine translation (SMT). Besides, we present source-side reordering as the pre-processing model to verify the extensibility of our method. Experi-mental results show that the morpheme-based weighting can substantially improve the translation quality.

key words: *machine translation, morpheme-based weighting, multiple decoding paths, low-resource language*

1. Introduction

Statistical machine translation (SMT), especially the phrase-based SMT, has developed very fast in the last decade [1]. However, the morphological difference of the language pair and the scarcity of training data limit the performance of SMT system [2]. Mongolian, one of the low-resource minority languages in China, is significant-ly different with Chinese. Generally speaking, Mongolian is a morphologically rich language and its word is composed of stem and affixes flexibly according to the requirement. Affix rather than word is used to represent the grammatical meaning in Mongolian. There are considerable independence between stem and additional ingredient which are just affixed when needed. However, Chinese is an isolated language with no additional ingredient. Table 1 illustrates the morphology of Mongolian based on the same stem and some different affixes.

From Table 1, we can conclude that Mongolian is a rich morphology language and grammatically correct Mongolian word form will be derived in exponential growth. Therefore, only large parallel corpus can contain the vast majority of Mongolian words. However, Mongolian is a low-resource language and the training corpus of the bilingual language pair is scarce [2].

Much of the work on SMT has shown that morphological segmentation could improve the SMT quality because of the sparseness reduction they contributed [3]. Some approach [4] presented the morph-ology in the factored translation model for Chinese-Mongolian SMT and attempted to resolve the problems of selecting word forms in the output sentences. However, the factored model creates additional computational complexity and the translation quality is easily affected by the generation model. Yang et al. [5] proposed a method that adopted morphological information as the features of the maximum entropy based phrase reordering model for Mongolian-Chinese SMT, which alleviated the influence of reordering caused by the data sparseness. Li et al. [6] used morpheme as a pivot language, which translated Chinese into morphemes and retrieved Mongolian sentence from morphemes. This can be regarded as sentence-level pivoting. Similar work based on phrase-level pivoting is explored to enrich translation model [7]. However, pivot-based method ignored the reliable of the specific phrase pairs caused by the data sparseness.

Difference from the above work, we handle the data sparseness in Chinese-Mongolian SMT by making full use of morphological information. A novel and efficient morpheme-based weighting is proposed to evaluate the phrase pair. The framework of our statistical machine translation system is shown in Fig. 1.

First, Mongolian words are segmented into morphemes by our previous work [2]. Then, we explore a smoothing strategy to construct a morpheme-based weighting to es-

Table 1 Illustration of the morphology of Mongolian.

Stem	Affix	Word	Chinese meaning	English meaning
		SVRVGCI	学生	student
	D	SVRVGCID	学生们	students
SVRVGCI	D-VN	SVRVGCID-VN	学生们的	students'
	-YIN	SVRVGCI-YIN	学生的	student's
	-TAI	SVRVGCI-TAI	与学生一起	with student

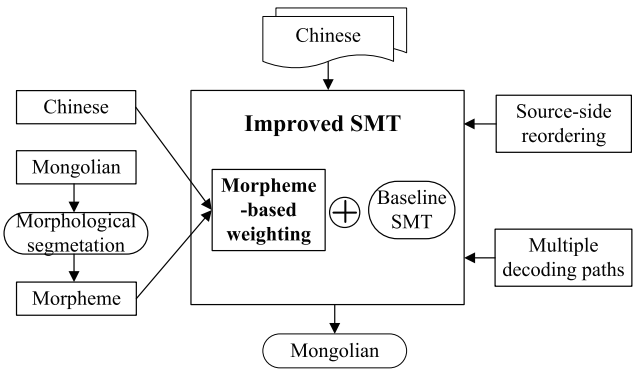

Fig. 1 Framework of improved SMT.

Manuscript received April 14, 2016.
Manuscript revised July 6, 2016.
Manuscript publicized August 19, 2016.

[†]The authors are with Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei 230031, China.
[††]The authors are with University of Science and Technology of China, Hefei 230026, China.
a) E-mail: xinzyang@mail.ustc.edu.cn

DOI: 10.1587/transinf.2016EDL8080

timate translation probabilities. Besides, the morpheme-based weighting is integrated into Chinese-Mongolian baseline SMT flexibly and effectively by multiple decoding paths, which uses two translation tables separately and outputs the translation result with the highest score. Finally, to further verify the effectiveness and extensibility of morpheme-based weighting, source-side reordering model as a pre-processing module is incorporated into the current systems.

Our contributions can be summarized as follows:

- Since the data sparseness is serious in our machine translation system, we propose a morpheme-based weighting as a smoothing method to evaluate phrase pairs and integrate this model with current baseline SMT flexibly and efficiently.
- Our approach can further improve the translation quality with other modules, such as source-side reordering due to the better extensibility.
- The method in this paper is a general method, besides the Chinese-Mongolian SMT, the method can also be adopted to other morphological low-resource lan-guages, such as Uyghur.

2. Morpheme-Based Weighting

2.1 Motivation

The unsymmetrical morphology of bilingual language pair and the lack of training data make data sparseness of Chinese-Mongolian SMT seriously. Phrase translation probabilities are important components of translation model. However, due to the data sparseness and the morphological richness of Mongolian, the probabilities of some uncommon phrase pair would be unreliable. For example, if both a phrase $\bar{e}$ and $\bar{f}$ appear once in the training data, the translation probabilities $\phi(\bar{f}|\bar{e})$ and $\phi(\bar{e}|\bar{f})$ equal to 1.

Intuitively, the decomposition of large unit into some small units can alleviate the data sparseness, since the small units appear more times than the large unit in the training data. The lexical weighting features estimate the probability of a phrase pair word-by-word, which would suffer from sparseness issues under the low-resource scene. In this paper, we adopt Mongolian morphemes instead of words to estimate phrase pair, reducing the sparseness caused by rich morphology [3]. The morpheme-based weighting is defined at the morpheme level alignment, and the quality of alignment would be improved by morphemes [10]. Finally, we utilize morphological features to estimate the phrase pairs which consist of Chinese words and Mongolian words, incorporating the information of morpheme level and phrase level at the same time.

2.2 Model

The morpheme-based weighting is proposed to estimate the phrase pairs extracted from parallel corpus. Since the log-linear framework allows us to integrate arbitrary features and the useful of the inverse probability is demonstrated in statistical machine translation [8], the morpheme-based weighting is added with both directions, i.e., direct morpheme-based weighting $p_m(\bar{e}|\bar{f})$ and inverse morpheme-based weighting $p_m(\bar{f}|\bar{e})$. In this subsection, we will introduce how to model $p_m(\bar{e}|\bar{f})$. $p_m(\bar{f}|\bar{e})$ is calculated in similar manner.

First, the morpheme-level translation probability distribution $m(t|f)$ is computed from parallel corpus, where the source language is Chinese word sequence and the target language is morpheme sequence. We denote f as a word in source language, denote t as a morpheme in target language. The probability of $m(t|f)$ is calculated as follows:

$$m(t|f) = \frac{Count(f,t)}{\sum_{t'} Count(f,t')} \tag{1}$$

A special source-side "NULL" token is added to each source sentence and aligned to each unaligned target morpheme. $m(t|\text{NULL})$ is calculated for morpheme-based weighting. This special token makes that all target morphemes have aligned points with source sentence. $m(f|t)$ is computed at the same time, which is necessary for calculating inverse morpheme-based weighting.

Given the phrase pair $(\bar{f},\bar{e})$, we can infer the corresponding phrase pair $(\bar{f},\bar{m})$, which $\bar{m}$ is the morpheme sequence of $\bar{e}$. Note that, each $\bar{e}$ may have different $\bar{m}$, since different context may result different morphological segmentations. Besides, during phrase pair extraction, there may be multiple alignments a for phrase pair $(\bar{f},\bar{m})$. If it is observed with more than one alignment pattern or morphological segmentation pattern, we use the most frequent pattern.

The morpheme-based weighting is computed by

$$\begin{aligned} p_m(\bar{e}|\bar{f}) &= p_w(\bar{t}|\bar{f},a) \\ &= \prod_{i=1}^{length(\bar{t})} \frac{1}{|\{j|(i,j)\in a\}|} \sum_{\forall(i,j)\in a} m(t_i|f_j) \end{aligned} \tag{2}$$

The basic idea of our method is that we hope to exploit Mongolian morpheme to alleviate the data sparseness, estimating the phrase pair we extract. If a morpheme has no alignment, its alignment in the source-side is treated as "NULL". Figure 2 provides an example of our morpheme-based weighting.

Note that the symbol "+", which is added by us without any meaning, is used for word recovery. The current phrase pair is "他 的 建议 ||| TEGUN-U JOBLEGE-YI", we find the best corresponding morphological segmentation and the best alignment. In Fig. 2, the morpheme-based weighting is represented as:

$$\begin{aligned} p_m(\bar{e}|\bar{f}) = & m(TEGUN+|\text{他}) \times m(+-U|\text{的}) \\ & \times m(\text{JOBLEGE}+|\text{建议}) \times m(+-\text{YI}|\text{NULL}) \end{aligned}$$

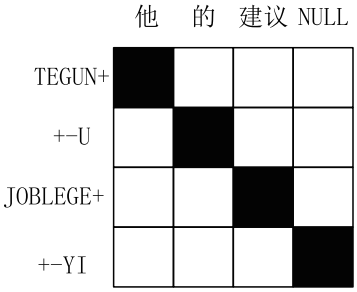

Fig. 2 Illustrative example of morpheme-based weighting.

3. Experiments

3.1 Morpheme-Based Weighting as Smoothing

In this subsection, we evaluate the translation results of method we proposed. Our Chinese-Mongolian parallel corpus, which is the training set for translation system, is obtained from the 5th China Workshop on Machine Translation (CWMT 2009). The statistics of the experimental data are listed in Table 2, where 500×4 means that each source sentence has four reference sentences.

Both Chinese and Mongolian sentences consist of words in our datasets. In order to construct the morpheme-based weighting, Mongolian words in Table 2 are segmented into morphemes by our previous work [2]. For SMT systems, the bidirectional word alignment is generated by GIZA++[†] and grow-diag-final-and heuristic. A 3-gram language model with modified Kneser-Ney smoothing is built by the SRI language modeling toolkit. Stanford parser is employed to parse Chinese sentences. The feature weights of log-linear model are learned by using minimum error rate training. We use toolkit ICTCLAS[††] for Chinese word segmen-tation and the open-source toolkit Moses [8] with its default settings for each translation task. Maximum phrase length is set to 7 when extracting phrase pair.

The baseline is a standard phrase-based SMT system, which we denote as system A. The training data of baseline SMT system consists of Chinese words and Mongolian words. To construct the morpheme-based weighting, we exploit GIZA++ to generate alignment for Chinese word sequences and Mongolian morpheme sequences. In this subsection, we use multiple decoding paths to add direct morpheme-based weighting and inverse morpheme-based weighting to the baseline and we denote this system as system B. Generally speaking, we use two translation tables. One is used in our baseline system, which includes phrase translation probability distributions with both directions and lexical weighting distributions with both directions. The other includes phrase translation probability distributions with both directions and morpheme-based weighting with both directions. Two translation tables are utilized separately by two decoding paths. The translation result with the highest score will be output.

[†]http://code.google.com/p/giza-pp/

[††]http://ictclas.nlpir.org/

Table 2 Statistics of the datasets.

Dataset		Chinese	Mongolian
Training set	sentences	67288	67288
	words	849916	822167
Dev set	sentences	500	500×4
	words	4330	12614
Test set	sentences	500	500×4
	words	4456	12896

Table 3 Translation results of morpheme-based weighting.

System	BLEU(%)
A	20.10
B	**20.73**
C	20.60

Table 4 Source-side reordering rules.

No.	Original rule	Reordering rule
(1)	VP—> VV PP	VP—> PP VV
(2)	VP—> VV NP	VP—> NP VV

Table 5 Translation results of the proposed system.

System	BLEU(%)
A	20.10
D	20.85
E	**21.31**

We report all the results with BLEU [9], which is calculated on Mongolian words. We run each experiment 3 times and get the average BLEU score as the experimental result. Table 3 illustrates translation results. System C is a related work [6].

From Table 3, we can see that morpheme-based weighting is effective for Chinese-Mongolian SMT, achieving 0.63 BLEU points increment over the baseline.

3.2 Source-Side Reordering

To further verify the effectiveness and extensibility of morpheme-based weighting, source-side reordering model as the pre-processing module is integrated into current SMT system to transform the word order of source language to match the order of target language. A phrase structure tree with syntactic information is acquired by Stanford Parser, then we follow our previous work [2] to use the manual rules to reorder the source language. The reordering rules are described in Table 4, where VP denotes verb phrase, PP denotes prepositional phrase, NP denotes noun phrase, VV denotes verb.

We exploit pre-processing model on source-side and morpheme-based weighting on target-side simultaneously. Table 5 illustrates translation results, where system A denotes the baseline SMT system, system D denotes the system which only source-side reordering is integrated into baseline, system E denotes both source-reordering and morpheme-based weighting are applied to the baseline SMT at the same time.

From Table 5, it can be noted that morpheme-

Example 1	
Source sentence:	你 晚上 干什么 ？
English translation:	What are you going to do tonight ?
Baseline:	TA 0R0I YAGV HIDEG BVI ?
System E:	TA 0R0I YAGV HIHU BVI ?
Ref0:	CI 0R0I YAGV HIHU BVI ?
Ref1:	TA 0R0 YAGV HIHU BVI ?
Ref2:	TA 0R0I YAGV HIHU YVM BVI ?
Ref3:	TA 0R0I DAGAN YAGV HIHU-BER BAYIN_A ?
Example 2	
Source sentence:	你 是 美国 人 吗 ?
English translation:	Are you an American ?
Baseline:	CI CINI AMeRIKACVD VV ?
System E:	CI AMeRIKA HOMON UU ?
Ref0:	CI AMeRIKA HUMUN UU ?
Ref1:	TA AMeRIKA-YIN HOMON UU ?
Ref2:	TA AMeRIKA HOMON UU ?
Ref3:	TA AMeRIKA HOMON MON UU ?

Fig. 3 Comparison examples between baseline and system E.

based weighting are effective and extensible for Chinese-Mongolian SMT. The highest BLEU score is 21.31%, which achieves 1.21 points increment over the baseline system, demonstrating that the combination of morphological information and syntax information can further improve the translation quality under the situation of linguistic difference and data sparseness.

3.3 Analysis of Results

In order to have a better intuition about the performance improvement, we compare baseline with system E. The translation results of all SMT systems are Mongolian word sequences. Figure 3 illustrates the translation results, where "English translation" denotes the corresponding English translation of source sentence, "Ref0" to "Ref3" denotes source sentence is translated by four Mongolian linguistic experts independently since the correct answer of translation result is not unique.

The first example is an illustration of tense choice. The difference between baseline and system E is "HIDEG" and "HIHU". "HI" is the stem which means "干什么" and it can be connected different affixes to express different tenses. "DEG" represents present tense while "HU" represents future tense. In the first example, source sentence expresses future tense, so system E can help translation system to select proper tense.

The second example is the selection of singular and plural. "CVD" is additional ingredient of plural noun. "AMeRIKA" means an American, which is the correct translation of "美国 人". However, "AMeRIKACVD" means Americans while its Chinese representation is "美国 人 们". Hence, system E is better than baseline.

4. Conclusion

The paper makes full use of morphological information to handle the problem in Chinese-Mongolian SMT caused by the morphological difference and the data sparseness. Experimental results show the effective-ness of our method. Besides, we take advantage of both source and target linguistic information to further enhance the performance of Chinese-Mongolian SMT. In future, we will verify the method for more low-resource morphological rich languages.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under No. 61572462, No. 61502445.

References

[1] M. Tu, Y. Zhou, and C. Zong, "Exploring diverse features for statistical machine translation model pruning," IEEE/ACM TASLP, vol.23, no.11, pp.1847–1857, 2015.
[2] L. Chen, M. Li, J. Zhang, Z. Zhu, and Z. Yang, "A statistical method for translating Chinese into under-resourced minority languages," Proc. 10th CWMT, vol.493, pp.49–60, 2014.
[3] M.R. Costa-Jussà, and M. Farrus, "Statistical machine translation enhancements through linguistic levels: A survey," ACM Computing Surveys (CSUR), vol.46, no.3, pp.1–28, 2014.
[4] P. Yang, J. Zhang, M. Li, Wudabala, and Y. Xue, "Morphology-processing in Chinese-Mongolian statistical machine trans-lation," Journal of Chinese Information Processing, vol.23, pp.50–57, 2009.
[5] Z. Yang, M. Li, Z. Zhu, L. Chen, L. Wei, and S. Wang, "A maximum entropy based reordering model for Mongolian-Chinese SMT with morpho-logical information," Proc. IALP, pp.175–178, 2014.
[6] W. Li, L. Chen, Wudabala, and M. Li, "Chained machine translation using morphemes as pivot language," Proc. COLING, pp.169–177, 2010.
[7] Z. Yang, M. Li, L. Chen, L. Wei, J. Wu, and S. Chen, "Constructing morpheme-based translation model for Mongolian-Chinese SMT," Proc. IALP, pp.25–28, 2015.
[8] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open source toolkit for statistical machine translation," Proc. ACL, pp.177–180, 2007.
[9] K. Papineni, S. Roukos, T. Ward, and W.-J Zhu, "BLEU: a method for automatic evaluation of machine translation," Proc. ACL, pp.311–318, 2002.
[10] M. Salameh, C. Cherry, G. Kondrak, "Lattice desegmentation for statistical machine translation," Proc. ACL, pp.100–110, 2014.