Semi-supervised Learning for Mongolian Morphological Segmentation

Zhenxin Yang[1,2], Miao Li[1(✉)], Lei Chen[1], Weihui Zeng[1], Yi Gao[3], and Sha Fu[3]

[1] Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei 230031, China
mli@iim.ac.cn, alan.cl@163.com
[2] University of Science and Technology of China, Hefei 230026, China
{xinzyang, whzeng}@mail.ustc.edu.cn
[3] Yunnan Agricultural Expert System Leading Group Office, Kunming 650000, China
498898209@qq.com, 1769816@qq.com

Abstract. Unlike previous Mongolian morphological segmentation methods based on large labeled training data or complicated rules concluded by linguists, we explore a novel semi-supervised method for a practical application, i.e., statistical machine translation (SMT), based on a low-resource learning setting, in which a small amount of labeled data and large amount of unlabeled data are available. First, a CRF-based supervised learning is exploited to predict morpheme boundaries by using small labeled data. Then, a lexicon-based segmentation model with small labeled data as the heuristic information is used to compensate the weakness in the first step by the abundant unlabeled data. Finally, we present some error correction models to revise segmentation results. Experimental results show that our method can improve the segmentation results compared with the pure supervised learning. Besides, we integrate the morphological segmentation result into Chinese-Mongolian SMT and achieve the satisfactory performance compared with the baseline.

Keywords: Semi-supervised learning; Morphological segmentation; Statistical machine translation; Low-resource language

1 Introduction

Morphological segmentation, which breaks words into the basic syntactic or semantic units, is a key issue in natural language processing, such as machine translation, information retrieval and speech recognition [1]. Morphological segmentation has been a research focus in recent years [2].

Mongolian is a morphological rich minority language which has significant difference compared with Chinese. Mongolian word is generated by connecting stem and none or one or more affixes according to the grammatical order. There are considerable independence between stem and additional ingredient which are just affixed when needed. There are about more than 30000 stems and 297 inflectional affixes in Mongolian [3]. Theoretically speaking, Mongolian word form will be derived in exponential growth. Table 1 illustrates the morphology of Mongolian.

Table 1. Illustration of the morphology of Mongolian.

stem	affix	word	Chinese	English
		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

From table 1, we can conduct that Mongolian is a rich morphology language and morphological segmentation is necessary for Mongolian natural language processing.

Previous work on Mongolian morphological segmentation is based on dictionaries, rules and statistics [4-9]. These proposed methods in previous work have achieved good segmentation results, however, they need lots of annotated training data or complicated rules concluded by linguists. The building of fundamental resource for Mongolian is very time-consuming, which is extremely difficult for researchers who do Mongolian morphological segmentation from scratch.

Difference with above work, we explore a novel and effective semi-supervised morphological segmentation method for a practical application, i.e., statistical machine translation, based on a low-resource learning setting, in which a small amount of labeled data and large amount of unlabeled data are available. We aim to leverage the large unlabeled data to alleviate some drawbacks caused by the lack of the labeled data, and reduce the reliance on the manual annotation.

The framework of the paper is shown in Figure 1.

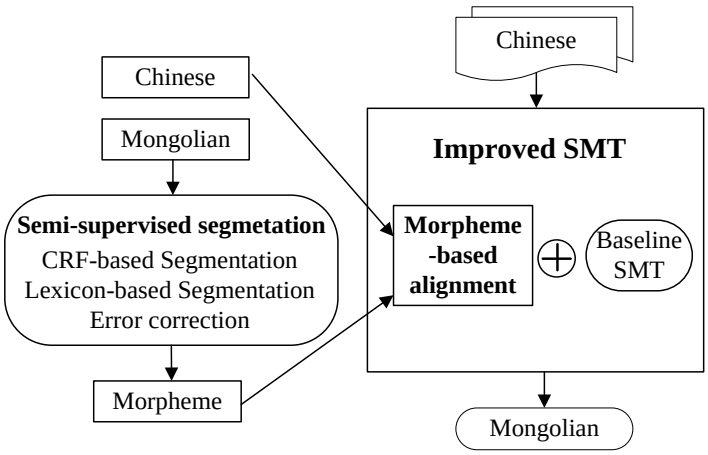

Fig. 1. Framework of the paper.

Semi-supervised learning includes three steps by small amount of labeled data and abundant amount of unlabeled data. First, we investigate a CRF-based supervised learning to predict morpheme boundaries via small amount of labeled data. Then, the abundant unlabeled data is exploited to compensate the weakness of the CRF-based

segmentation. Finally, we present some error correction models to revise segmentation results.

As the proposed method is for statistical machine translation, we integrate morphemes into Chinese-Mongolian SMT by the combination of word alignment and morpheme-based alignment. The translation results demonstrate that morphological segmentation based on small amount of labeled data can help to achieve a satisfactory translation performance.

2 Semi-supervised Morphological Segmentation

2.1 CRF-based Segmentation

Conditional random fields [10] is a probabilistic models to segment and label sequence data, which can avoid label bias problem. Assume X is the the random variable over data sequence to be labeled and Y is the random variable over corresponding label sequence. Let $G=(V,E)$ denotes an undirected graphical model, where $v \in V$ represents a random variable Y_v, and the edge $e \in E$ represents probabilistic dependencies among random variables. $P(Y|X)$ is a conditional random fields if the following formula is established for any vertex v:

$$P(Y_v \mid X, Y_w, w \neq v) = P(Y_v \mid X, Y_w, w \sim v) \tag{1}$$

where $w \sim v$ means w and v are neighbors in G and $w \neq v$ means all vertices except v.

In this paper, we simplify the CRF into linear chain conditional random fields by assuming X and Y have the same graphical structure, since we regard Mongolian morphological segmentation as sequence labeling problem.

Formally, given $X=x$ (characters in a word) and $Y=y$ (classes corresponding to characters), the probability $P(y|x)$ is written as:

$$P(y \mid x) = \frac{1}{Z(x)} \exp \sum_{k=1}^{K} w_k f_k(y, x) \tag{2}$$

$$Z(x) = \sum_{y} \exp \sum_{k=1}^{K} w_k f_k(y, x) \tag{3}$$

In the subsection, only small amount of labeled data is exploited to accomplish pure supervised segmentation. It should be noted that there exists stem lemmatization in morphological segmentation. For example, Mongolian word “BAYIG_A” has a stem “BAI” and an affix “G_A”. Since labeled data is small, we just ignore the stem lemmatization in a low-resource setting. The performance of SMT with morphological information demonstrates the effectiveness of semi-supervised segmentation although we ignore the stem lemmatization. In this paper, “BAYIG_A” will be segmented to “BAYI” and “G_A”. The tag set used in conditional random fields is shown in Table 2.

Table 2. The tag set used in CRF.

Tag name	Meaning
S_B	first letter of stem
S_O	other letter of stem
A_B	first letter of affix
A_E	last letter of affix
A_O	letter of affix not belong to A_B and A_E
A_ONLY	only one letter in affix

Feature extraction is the key issue for conditional random fields. We describe the position *t* of the word *x* using all left and right substrings up to a maximum length. Furthermore, we extract interval letter relationship to alleviate the long-distance dependence.

2.2 Lexicon-based Segmentation

The performance of CRF-based model will be limited due to the small amount of labeled data. In this subsection, we explore a lexicon-based segmentation model which utilizes large unlabeled data to overcome the weakness of CRF-based model. The key idea is that we hope to learn valuable knowledge from large unlabeled data given small labeled data as heuristic information.

We follow the work [11], which is a generative probabilistic model, in this subsection. The model parameters θ encode a morph lexicon, which includes the properties of the morphs, such as their string representations. Each morph *m* in the lexicon has a probability of occurring in a word. The probabilities are assumed to be independent. The model uses a prior $P(\theta)$ derived using the Minimum Description Length (MDL) principle. During model learning, θ is optimized to maximize the posterior probability:

$$\theta^{MAP} = \arg\max_{\theta} P(\theta \mid D_w) = \arg\max_{\theta} P(\theta)P(D_w \mid \theta) \tag{4}$$

where D_w includes words in the training data.

The cost function is expressed as:

$$L(\theta, z, D_w) = -\ln P(\theta) - \ln P(D_w \mid z, \theta) \tag{5}$$

We process one word at a time, and the segmentation that minimizes the cost function with the optimal model parameters is selected:

$$z_j^{(t+1)} = \arg\min_{z_j} \{\min_{\theta} L(\theta, z^{(t)}, D_w)\} \tag{6}$$

Then, the parameters are updated:

$$\theta^{(t+1)} = \arg\min_{\theta}\{L(\theta, z^{(t+1)}, D_w)\} \tag{7}$$

We repeatedly exploit small amount of labeled data as the heuristic information to train morph lexicon with the large unlabeled data, which perform Expectation-Maximization (EM) using the Viterbi algorithm on the morphological segmentation.

The lexicon-based model proposed in this subsection is exploited to segment the Mongolian words which is error from the CRF-based model. We have an inflectional affix dictionary which contains 297 inflectional affixes. If the affix of the word segmented by the CRF-based model is not found in the inflectional affix dictionary, we explore the lexicon-based model to segment its word form.

2.3 Error Correction

We find the error in CRF-based segmentation and lexicon-based segmentation according to inflectional affix dictionary and employ the lexicon model of word and its affix, the reverse maximum match with 1-gram model of affix to correct the mistake. We search the most possible affix for the error segmentation result based on the lexicon model. If the affix is not found in lexicon model, we use reverse maximum match method to obtain the final segmentation result.

The lexicon model of word and its affix is

$$L =< w,[< a_1, c_1 > \cdots < a_i, c_i > \cdots] > \tag{8}$$

where w is Mongolian word, a_i is the possible affix and c_i is the count of a_i in the training data. We can get the statistical vocabulary V={(w_i,[<a_{ik}, c_{ik}>…])…}.

We find the most possible affix for the error word and affix pair w_i/a_i by the following decision rule

$$\hat{a}_i = \{a_{ij} \mid c_{ij} \geq \max\{c_{ik}\},\ j \neq k\} \tag{9}$$

If w_i is not found in lexicon model, we reverse to search w_i and collect all affixes that match the 1-gram model of affix to find the maximum length affix as the final segmentation affix.

3 Experiments

3.1 Morphological Segmentation

In this subsection, we will report the data set, evaluation metric and experimental results. We extract 800 high frequency words from Mongolian monolingual corpus and the 800 Mongolian word types are segmented by the linguists manually. Note that a Mongolian word may have different segmentations due to the different context and we choose the most frequent one. We only use 800 high frequency words as small labeled data in this paper. And it is different with much related work [4, 5, 7, 9, 12],

which exploit large amount of labeled sentences. The test set contains 1000 sentences. 29.4% word tokens in the test set are unknown words.

A Mongolian word is segmented correctly when both stem and affix is correct. Since semi-supervised segmentation in the paper is used to improve the performance of Chinese-Mongolian SMT, we utilize stem of Mongolian word in the word alignment step and ignore the boundaries between different affixes. In this paper, we assess the segmentation results with stem-level accuracy as evaluation indicators.

Experimental results are shown in Table 3.

Table 3. Stem-level accuracy of segmentation.

Semi-supervised learning	Accuracy(%)
CRF-based Segmentation	80.05
+Lexicon-based Segmentation	87.58
+Error Correction	**90.03**

From Table 3, we can conclude that the accuracy of our semi-supervised morphological segmentation, which uses CRF-based segmentation, lexicon-based segmentation and error correction, could reach 90.03% although 29.4% word tokens in the test set are unknown words. The segmentation accuracy improves significantly when we utilize lexicon-based approach for the errors in CRF-based segmentation, demonstrating that large unlabeled data is useful for Mongolian morphological segmentation.

There are large amount of labeled sentences needed in related work, while semi-supervised learning we proposed only need 800 high frequency words annotation rather than labeled sentences. The comparative experiments are not conducted because much related work need labeled sentences, which capture the context information for n-grams.

3.2 Improved SMT with Morpheme

Our Chinese-Mongolian parallel corpus is obtained from the 5th China Workshop on Machine Translation (CWMT 2009). The statistics of the experimental data are listed in Table 4, where 500×4 means that each source sentence has four reference sentences.

Table 4. Statistics of all 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

In this subsection, we integrate Mongolian morpheme information into Chinese-Mongolian SMT. Generally speaking, we use Mongolian stem rather than word to generate word alignment matrix with Chinese word sequence. The word alignment matrix that contains morpheme information is used to replace the word alignment result of baseline SMT. Besides, we explore the two alignment results combination to further improve the translation performance.

The baseline is a standard phrase-based statistical machine translation system. We conduct two group experiments to verify the effectiveness of our method.

We employ GIZA++ and grow-diag-final-and [13] heuristic to generate the bidirectional word alignment. A 3-gram language model with modified Kneser-Ney smoothing [14] is built by the SRI language modeling toolkit [15]. We use Stanford parser [16] to parse Chinese sentences. The log-linear model feature weights are learned by using minimum error rate training [17]. Besides, we report all the results with BLEU [18]. We use toolkit ICTCLAS for Chinese word segmentation. Maximum phrase length is set to 7 when extracting phrase pair. We run each experiment 3 times and get the average BLEU score as the experimental result.

Table 5 illustrates translation results, where "A" denotes the standard phrase-based system, "B" denotes that we use Mongolian stem rather than word to generate word alignment matrix with Chinese word sequence. The word alignment matrix which contains morpheme information is used to replace the alignment result of system "A". "C" denotes we combine the alignment results of both "A" and "B" at the same time.

Table 5. Translation results with morphological information.

System	BLEU(%)
A	20.10
B	20.58
C	**20.91**

From Table 5, we can conclude that morphological information segmented by semi-supervised learning improves the performance of SMT significantly.

Example 1	
Source sentence:	我 喝 点 茶 吧 。
English translation:	I drink some tea.
System A:	BI JIGAHAN CAI **VVGV** .
System C:	BI JIGAHAN CAI **VVGVY_A** .
Ref0:	BI JIGAHAN CAI VVGVY_A .
Ref1:	BI CAI VVGVY_A .
Ref2:	BI CAI VVGVHV SANAGATAI .
Ref3:	BI CAI VVGVMAR BAYIN_A .

Fig. 2. Comparison examples between the baseline and our proposed method

In order to have a better intuition about the performance improvement, we compare translation result between system "A" and system "C". Figure 2 illustrates the transla-

tion results, where “Source sentence” means Chinese sentence to be translated into Mongolian, “English translation” denotes the corresponding English translation for better understanding, “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 example is the selection of the correct morphology. “System A” predicts the verb “喝” of the source sentence as “VVGV”. Although “VVGV” is the translation of “喝”, “Y_A” is affix that represents the first person. In the source sentence, since “我” is the first person, “VVGVY_A” not only means “ 喝” but also represents grammatical meaning. Hence, “System C” can generate correct morphology of Mongolian.

4 Related Work

Mongolian morphological segmentation has been a research focus recently. Generally speaking, previous work are based on dictionaries, rules and statistics [4-9].

(Nasanurtu, 1997) [19] proposed the method of combining rule-based and dictionary to accomplish Mongolian word segmentation. The construction of the rules and dictionary need significantly efforts by linguists. (Hou, et al., 2009) [6] used rules to segment Mongolian words and applied Mongolian statistical language model to eliminate the ambiguity in the process. However, the rules may also be conflicted with each other.

Statistical methods are the dominant approaches. (He, et al., 2012) exploited a HMM-based approach [8] for Mongolian morphological segmentation. Besides, CRF-based model [7, 9] achieved outstanding performance. Some approach [12] combined the statistical machine translation method and the minimum constituent context cost model to accomplish Mongolian morphological segmentation, which handled in-vocabulary and out-of-vocabulary Mongolian words well respectively. However, these approaches did not pay much attention on word formation characteristics of the morphology.

(Jiang, et al., 2011) [4, 5] proposed a directed graph model for Mongolian lexical analysis. This model described the lexical analysis result as a directed graph and used three kinds of transition or generation probabilities. The approach predicted the best segmented and tagged candidate for each word according to the context.

The methods mentioned above need lots of annotated training data or complicated rules concluded by linguists, and the construction process is significantly time-consuming.

Difference from above work, we explore a novel and effective method which use large unlabeled data to compensate the weakness of the lack of labeled data. Our work focus on the practical application, i.e., statistical machine translation, hence we pay more attention to the performance of machine translation system. We explore the semi-supervised learning for Mongolian morphological segmentation. To our knowledge, it is the first time for Mongolian to apply semi-supervised morphological segmentation.

5 Conclusion and Future Work

The paper proposes a method, which makes full use of large unlabeled data and small labeled data, to segment Mongolian words into morphemes. We investigate a CRF-based supervised learning to predict morpheme boundaries via small amount of labeled data. Besides, the abundant unlabeled data is exploited to compensate the weakness of the small amount of labeled data. Furthermore, some error correction models are exploited to revise segmentation results. The experimental results on morphological segmentation and Chinese-Mongolian SMT demonstrate the effectiveness of the proposed approach.

In summary, we make the following contributions.

(1) We explore the semi-supervised learning based on a low-resource learning setting, in which a small amount of labeled data and large amount of unlabeled data are available.
(2) Our work reduces the reliance on the manual annotation for Mongolian morphological segmentation.
(3) The method in this paper is a general method, besides Mongolian morphological segmentation, the method can also be adopted to other morphological low-resource languages, such as Uyghur.

In future, we will focus on the semi-supervised morphological segmentation with POS information. Besides, we will verify the method for more low-resource morphological rich languages.

Acknowledgement. This work is supported by the National Natural Science Foundation of China under No. 61572462, No. 61502445, the National Key Technology R&D Program under No. 2014BAD10B03.

Reference

1. Ruokolainen, T., Kohonen, O., Sirts, K., Gr¨onroos, S.A., Kurimo, M., Virpioja, S.: A comparative study on minimally supervised morphological segmentation. Computational Linguistics p. 42(1) (2016)
2. Ahlberg, M., Forsberg, M., Hulden, M.: Semi-supervised learning of morphological paradigms and lexicons. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. pp. 569–578 (2014)
3. Yang, P., Zhang, J., Li, M., Wudabala, Xue, Y.: Morphology-Processing in Chinese-Mongolian Statistical Machine Translation (In Chinese). Journal of Chinese Information Processing 23(1), 50–57 (2009)
4. Jiang, W., Wu, J., Chang, Q., Nasan-urtu, Liu, Q., Zhao, L.: Directed graph model for Mongolian lexical analysis (In Chinese). Journal of Chinese Information Processing 25(5), 94–100 (2011)

5. Jiang, W., Wu, J., Wuriliga, Nasan-urtu, Liu, Q.: Discriminative stem-affix segmentation for directed-graph-based Mongolian lexical analyzer (In Chinese). Journal of Chinese Information Processing 25(4), 30–34 (2011)
6. Hou, H., Liu, Q., Nasanurtu, Murengaowa, Li, J.: Mongolian word segmentation based on statistical language model. Pattern Recognition and Artificial Intelligence (In Chinese) 22(1), 109–112 (2009)
7. Zhao, W., Hou, H., Cong, W., Song, M.: Research on conditional random fields based Mongolian word segmentation (In Chinese). Journal of Chinese Information Processing 24(5), 31–35 (2010)
8. He, M., Li, M., Chen, L.: Mongolian Morphological Segmentation with Hidden Markov Model. In: IALP. pp. 117–120 (2012)
9. Liu, H., Li, M., Zhang, J., Chen, L.: Morpheme Segmentation Using Bilingual Features. In: IALP. pp. 209–212 (2012)
10. Lafferty, J., McCallum, A., Pereira, F.C.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proc. of ICML (2001)
11. Creutz, M., Lagus, K.: Unsupervised models for morpheme segmentation and morphology learning. ACM Transactions on Speech and Language Processing (TSLP) 4(1), 3 (2007)
12. Li, W., Chen, L., Wudabala, Li, M.: Chained machine translation using morphemes as pivot language. In: COLING 2010 workshop: ALR. pp. 169–177 (2010)
13. Koehn, P., Och, F.J., Marcu, D.: Statistical phrase-based translation. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. pp. 48–54. Association for Computational Linguistics (2003)
14. Chen, S.F., Goodman, J.: An empirical study of smoothing techniques for language modeling. In: Proceedings of the 34th annual meeting on Association for Computational Linguistics. pp. 310–318. Association for Computational Linguistics (1996)
15. Stolcke, A.: SRILM - an extensible language modeling toolkit. In: Proc. Intl. Conf. on Spoken Language Processing. pp. 901–904 (2002)
16. Levy, R., Manning, C.: Is it harder to parse chinese, or the chinese treebank? In: Proceedings of ACL. pp. 439–446 (2003)
17. Och, F.J.: Minimum error rate training in statistical machine translation. In: ACL. pp. 160–167 (2003)
18. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: ACL. pp. 311–318 (2002)
19. Nasanurtu: An automatic segmentation system for the root, stem, sufix of the mongolian. Journal of Inner Mongolia University (In Chinese) 29(2), 53–57 (1997)