Realignment from Finer-grained Alignment to Coarser-grained Alignment to Enhance Mongolian-Chinese SMT

Jing Wu Hongxu Hou Congjiao Xie
College of Computer Science
Inner Mongolia University
Hohhot, 010021, China
cshhx@imu.edu.cn

Abstract

The conventional Mongolian-Chinese statistical machine translation (SMT) model uses Mongolian words and Chinese words to practice the system. However, data sparsity, complex Mongolian morphology and Chinese word segmentation (CWS) errors lead to alignment errors and ambiguities. Some other works use finer-grained Mongolian stems and Chinese characters, which suffer from information loss when inducting translation rules. To tackle this, we proposed a method of using finer-grained Mongolian stems and Chinese characters for word alignment, but coarser-grained Mongolian words and Chinese words for translation rule induction (TRI) and decoding. We presented a heuristic technique to transform Chinese character-based alignment to word-based alignment. Experimentally, our method outperformed the baselines: fully finer-grained and fully coarser-grained, in terms of alignment quality and translation performance.

1 Introduction

Mongolian is an agglutinative language and has complex morphology. The current scale of Mongolian-Chinese parallel corpus is very small. These two reasons make data sparsity a very serious problem in Mongolian-Chinese SMT. Using finer-grained Mongolian stems rather than Mongolian words can reveal the word semantics and alleviate data sparsity. On the other hand, CWS is a necessary process to separate Chinese words, because Chinese words are not naturally separated by space (Jiang et al., 2009). CWS can achieve high accuracy, but does not necessarily guarantee better performance of alignment (Chang et al., 2008; Zhang et al., 2008; Xiao et al., 2010). Besides, CWS also brings errors (Xiao et al., 2010). Using of finer-grained Chinese characters, which are separated without using of CWS, can avoid the CWS errors and alleviate data sparsity. However, coarser-grained basic units are proved perform better in translation rule induction (TRI). (Philipp Koehn et al., 2003).

So inspired by the work of (Xi et al., 2011; Xi et al., 2012), we proposed a method that uses different granularity respectively for alignment and TRI. We train a finer-grained alignment using Mongolian stems and Chinese characters. Afterwards, we realign it to Chinese words and Mongolian words alignment for the following TRI and decoding. We design a technique to convert finer-grained alignment to coarser-grained alignment. The conversion can be unambiguous after carefully processing the differences brought by Mongolian word lemmatization and CWS.

In the experiments, our method outperformed the baselines of fully finer-grained and fully coarser-grained, in terms of alignment quality and translation performance. The experiments indicate that using finer-grained basic units for alignment and

29th Pacific Asia Conference on Language, Information and Computation: Posters, pages 89 - 95
Shanghai, China, October 30 - November 1, 2015

coarser-grained basic units for TRI performs better than other granularity combinations.

The rest of the paper is organized as follows: Section 2 explains how our method designed and how can it have good influence on alignment and translation. Section 3 demonstrates the realignment model and analyzes how it works for better alignment. Section 4 describes the evaluations. Section 5 is the conclusion.

2 Design of different Granularity Alignment

The conventional practice of SMT uses Mongolian and Chinese words in the process of word alignment and TRI (Brown et al., 1993). We proposed a method of using finer granularity for word alignment but coarser granularity for TRI to enhance the Mongolian-Chinese SMT system. The process of the method is depicted in Figure 1:

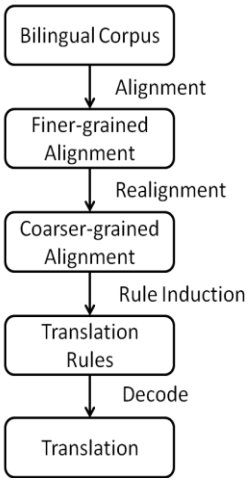


Figure 1. Process of the method

(1) In the first step, we get the finer-grained alignment by using Mongolian stems and Chinese characters as basic units;
(2) In the realignment procedure, we transform finer-grained alignment into coarser-grained alignment through a converting technique;
(3) In the step of TRI and decoding, we use the coarser-grained alignment.

Mongolian words are formed by stems and suffixes (Hou et.al., 2000). For some examples: when a noun plays different constituents in sentence, like subject or object, the case suffixes added to it are different; a verb adds different inflectional suffixes when it is under different tenses or followed by different nouns; a word has different forms (with the same word stem but different suffixes) when it is in different positions of the sentence. Therefore, Data sparsity is a very serious problem in Mongolian-Chinese SMT because of the complex Mongolian morphology and the small scale parallel corpus. Mongolian stems-based alignment can mitigate this problem, because Mongolian words in different forms but with the same semantic meaning will become one same stem after removing some suffixes. Besides, using Chinese characters for alignment can avoid the errors brought by CWS. Table 1 shows the token distribution of Mongolian words and Mongolian stems in corpus. We can see that the unique tokens in stem-based corpus reduce almost 10% than those in word-based corpus. Table 2 shows the frequency distribution of words and characters of Chinese corpus. The tokens whose frequency is no than 4 has a lower percentage in character-based corpus. We see that the unique tokens in character segment corpus are only one-third of those in word segment corpus. In the fined-grained Chinese corpus, the frequency of 77.88% tokens are equal to or more than 5, while the percentage of word tokens in coarser-grained Chinese corpus is only 38.74%. The above statistical data prove that coarser-grained word alignment suffers from more serious data sparsity than finer-grained word alignment.

	Word	Stem
Total Tokens	37140	29861
Unique Tokens	20859	14340
Percentage (%)	56.16	48.02

Table 1. Unique tokens of Mongolian word and stem

Frequency	Word (%)	Character (%)
1	31.25	9.34
2	14.91	5.85
3	8.84	3.47
4	6.26	3.46
5+	38.74	77.88

Table 2. Frequency distribution of Chinese word and character

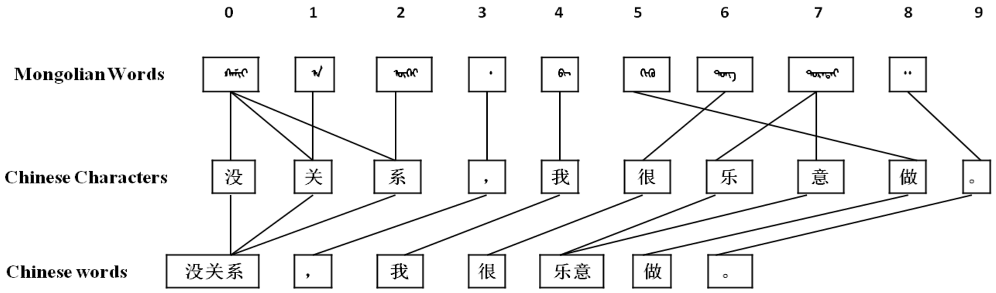


Figure 2. Realignment from finer-grained to coarser-grained

However, comparing to finer-grained tokens, coarser-grained tokens have more complete semantic information. State-of-the-art SMT models achieve excellent results by extracting phrases to induct the translation rules (Philipp Koehn et al., 2003). When the phrase-based translation models try to extract and score the phrases by getting lexical translation table, the probability of words to words can express more semantic information than stems to characters (Deng and Zhou, 2009). Moreover, when we use language model, the position information expressed by Mongolian word suffixes might be ignored by using Mongolian stems. Therefore, we still use coarser-grained units to induct the translation rules.

3 Realignment

The realignment from Mongolian stems to Mongolian words is an easy method of one-to-one mapping because there is no position changing. We build a heuristic model to describe the Chinese realignment. We set e and f as the source (Mongolian) and target (Chinese) sentence in finer-grained alignment. Given finer-grained source sentence (Mongolian) e' and target sentence (Chinese) f', we can get the coarser-grained alignment a by the realignment model as equation (1):

$$\mathrm{P}(\mathrm{a}|\mathrm{e},\mathrm{f}) = \mathrm{P}(\mathrm{a_c}|\mathrm{e}',\mathrm{f}')\mathrm{P}(\mathrm{a}|\mathrm{a_c}) \qquad (1)$$

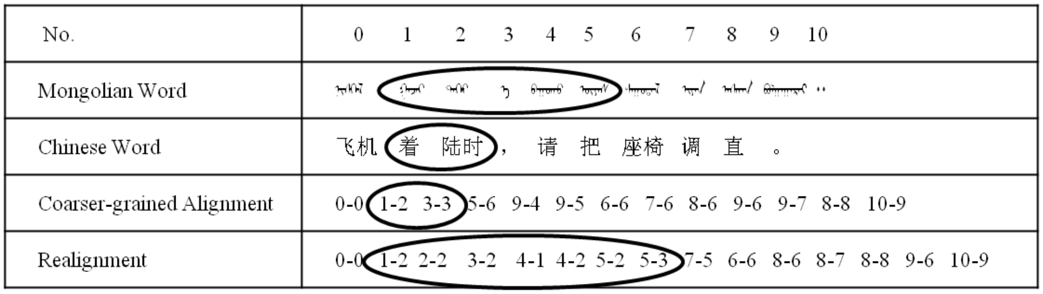
In the model, $\mathrm{a_c}$ is the finer-grained alignment getting from $\mathrm{e}'$and f'. $\mathrm{P}(\mathrm{a_c}|\mathrm{e}',\mathrm{f}')$ is the alignment model used in our alignment training which can be given as log-linear model by (Och and Ney, 2005; Liu et al., 2005) as equation (2).

$$\mathrm{P}(\mathrm{a_c}|\mathrm{e}',\mathrm{f}') = \frac{\exp[\sum_{\mathrm{m}=1}^{\mathrm{M}} \lambda_{\mathrm{m}} \mathrm{h_m}(\mathrm{a_c},\mathrm{e}',\mathrm{f}')]}{\sum_{a_c'} \exp[\sum_{\mathrm{m}=1}^{\mathrm{M}} \lambda_{\mathrm{m}} \mathrm{h_m}(\mathrm{a_c}',\mathrm{e}',\mathrm{f}')]} \qquad (2)$$

The conversion model $\mathrm{P}(\mathrm{a}|\mathrm{a_c})$ can be modeled based on (Zhang, 2003) as equation (3):

$$\mathrm{P}(\mathrm{a}|\mathrm{a_c}) = \mathrm{P}(\mathrm{a},\mathrm{a_c})/\mathrm{P}(\mathrm{a_c}) \qquad (3)$$

It is easy to understand that the transformation from a finer-grained sentence to its coarser-grained sentence is unambiguous. An example of conversion shows in figure 2, we can get the word alignment from Mongolian words to Chinese words by converting Chinese characters into Chinese words. “没关系” is a Chinese word which means “It does not matter”. It is composed of three characters “没”, “关” and “系”. The alignment from Mongolian words to Chinese characters “没”, “关” and “系” is “0-0, 0-1, 0-2, 1-1, 2-2”, the alignment from Chinese characters to Chinese word “没关系” is “0-0, 1-0, 2-0”, so the realignment from Mongolian words to Chinese word is “0-0, 1-0, 2-0”. Another example shows in figure 2 is the alignment from Mongolian word “ ᠪᠠᠶᠠᠷᠲᠠᠢ ” to Chinese word “乐意”, which means “with pleasure”. Comparing with Chinese words, Chinese characters carry more uncertain meaning. “关” is a verb which means “close”, but when it is followed by “系”, which is also a verb and means “tie”, the meaning of “关系” is “relation” and it is a noun. So using Chinese characters as basic unit may induce more interference alignment options. However, the recall score gets higher when we apply Chinese characters to do the alignment. Because we find that when we get the word alignment by realigning



No.	0 1 2 3 4 5 6 7 8 9 10
Mongolian Word	[illegible]
Chinese Word	飞机 着 陆时 ， 请 把 座椅 调 直 。
Coarser-grained Alignment	0-0 1-2 3-3 5-6 9-4 9-5 6-6 7-6 8-6 9-6 9-7 8-8 10-9
Realignment	0-0 1-2 2-2 3-2 4-1 4-2 5-2 5-3 7-5 6-6 8-6 8-7 8-8 9-6 10-9

Figure 3. Comparing coarser-grained alignment with realignment

from Chinese character-based alignment rather than by Chinese words directly, there are fewer invalid alignment options. We believe this problem is avoided by the feature of co-occurrence and distortion used by alignment models which explained in detail by (Xi et al., 2012).

Moreover, we find that our method can mitigate word alignment errors which caused by incorrect CWS result. As shows in figure 3, the correct segmentation should be "飞机 着陆 时" but not "飞机 着 陆时". The correct alignment should make No.1, No.2, No.3, No.4 and No.5 of Mongolian words align to the No.1 and No.2 of Chinese words, but Chinese word-based alignment only gets the right alignment 1-2, but wrongly aligns No.3 of Mongolian word to the Chinese comma as showed in the fourth row of figure 3. In our method, we can get a more precise alignment result based on characters "着", "陆" and "时". The realignment based on a wrong word segmentation result will lead to a wrong word alignment inside the phrase "着 陆时". However, as showed in the fifth row of the figure 3, we find that because of the better character-based alignment, the phrase "着 陆时" as a whole still can be realigned more precisely to its corresponding Mongolian phrase.

In conclusion, due to a more precise Chinese character-based alignment, our realignment based on Chinese word segmentation (even based on a wrong word segmentation result) can get a more precise word alignment result.

4 Experiments

We implement Moses as our basic SMT system and built it as follows: alignment performed by GIZA++ (Och and Ney, 2003). A phrase-based MT decoder similar to the work of (Koehn et al., 2007) was used with the decoding weights optimized by MERT (Och, 2003). We use a 3-gram language model. Mongolian language resources and Mongolian processing tools are scarce. CWMT'2009 (Zhao et al., 2012) was used for the experiments. It is a small training set when compares to major language training set because as a small language, public Mongolian and Chinese parallel corpus is limit. The lemmatization tool we used is the same as (Yu and Hou, 2011; Hou et.al., 2009). Table 3 shows the data set in detail. Mo is the abbreviation of Mongolian and Ch is the abbreviation of Chinese.

	Train	Dev	Test
Bilingual sentence pairs	66808	1000	1000
Scale	18.3MB	214KB	213KB
Total Mo words/stems	869168	11239	11134
Total Ch words	846574	8765	8697
Total Ch characters	1096551	12569	12526

Table 3. Data set

We manually aligned 100 pairs of bilingual sentence to evaluate the alignment performance including precision, recall, F-score and AER (David et al., 2003). As table 4 shows, after using finer-grained stem-based as basic units: precision has been increased from 62.75% to 63.82%; recall has been increased from 75.91% to 83.47% and improved significantly by using Chinese characters; AER has been reduced 2.74% from 39.44 to 38.36. These evaluations prove that our method of using finer-grained for alignment enhances the quality of SMT alignment and reduce the AER. The good performance in alignment partly because of the process of data sparsity we argued in section 2 and partly because of the good realignment we discussed in section 3.

Mongolian	Chinese	Precision	Recall	F-score	AER
word	word	62.75	75.91	69.33	39.44
stem	word	62.94	77.39	70.17	38.83
word	character	63.71	82.25	72.98	38.89
stem	character	63.82	83.47	73.65	38.36

Table 4. Alignment evaluation of Precision, Recall and F-score

To evaluate the translation performance of tour method, we do experiments on all kinds of grammatical components including: fully coarser-grained, different grained units for alignment and TRI. We also evaluate the influence on using finer-grained and coarser-grained units on source or target language. In the experiments of translation, we set conventional Mongolian-Chinese SMT system as baseline 1. We also set baseline 2, baseline 3 and baseline 4 which use finer-grained for both alignment and TRI to compare with our systems.

From table 5 we can see that all our three systems outperform the baseline 1. The comparison between our systems and the baseline 1 shows that using finer-grained basic units in alignment outperforms the conventional Mongolian-Chinese SMT. The BLEU of System 3 is higher than system 1 and system 2, which proves that using finer-grained for both source language and target language achieve better performance than using it on one language.

		Alignment	TRI	BLEU
Baseline1	Mo	word	word	21.88
	Ch	word	word	
System 1	Mo	stem	word	22.15
	Ch	word	word	
System 2	Mo	word	word	23.36
	Ch	character	word	
System 3	Mo	stem	word	23.49
	Ch	character	word	

Table 5. Translation evaluation of proposed systems and Baseline 1.

In the comparison of table 6, baseline 2 uses finer-grained basic units for Mongolian alignment and TRI, while system 1 uses finer-grained basic units only for Mongolian alignment but not TRI. System 1 outperformed Baseline 2 indicates that using coarser-grained Chinese units for TRI is more proper and applying our method to source language of Mongolian is successful.

		Alignment	TRI	BLEU
Baseline 2	Mo	stem	stem	21.97
	Ch	word	word	
System 1	Mo	stem	word	22.15
	Ch	word	word	

Table 6. Compare our System 1 with Baseline 2.

In the comparison of table 7, baseline 3 uses finer-grained basic units for Chinese alignment and TRI, while system 2 uses finer-grained basic units only for Chinese alignment but not TRI. System 3 outperformed Baseline 4 indicates that using coarser-grained Chinese units for TRI is more proper and applying our method to target language of Chinese is successful.

		Alignment	TRI	BLEU
Baseline 3	Mo	word	word	23.19
	Ch	character	character	
System 2	Mo	word	word	23.36
	Ch	character	word	

Table 7. Compare our System 2 with Baseline 3.

In the comparison of table 8, baseline 4 uses finer-grained basic units for both Mongolian and Chinese alignment and TRI, while system 3 uses finer-grained basic units only for Mongolian and Chinese alignment but not TRI. System 1 outperformed Baseline 2 indicates that using coarser-grained units in both Chinese and Mongolian for TRI is more proper and our method is successful in the evaluation.

		Alignment	TRI	BLEU
Baseline 4	Mo	stem	stem	22.73
	Ch	character	character	
System 3	Mo	stem	word	23.49
	Ch	character	word	

Table 8. Compare our System 3 with Baseline 4.

These comparisons of table 5 to table 8 proved that:

(1) Using finer-grained for alignment performed better then coarser-grained (table 5) because finer-grained basic units can enhance the alignment quality (table 4).

(2) Using coarser-grained for TRI, which means using finer-grained only for alignment rather than

using them though the whole translation process is better (table 8), because stems and characters are too finer to induct the translation rules.

(3)Using our method of finer-grained for alignment and coarser-grained for TRI improved the conventional SMT system and outperformed other grammatical components (table 5 and table 8);

(4) Using our method only in one side of source language or target language also performed better (table 6 and table 7).

5 Conclusion

We presented a method of using finer-grained Mongolian stems and Chinese characters as basic units for alignment, but coarser-grained Mongolian and Chinese words for TRI. Our method outperforms four baselines, mitigates the data sparsity and enhances the alignment quality and translation performance. Through the experiments we find some conclusions as follows: applying finer-grained units can perform a better word alignment result; Using finer-grained basic units for alignment, but coarser-grained for TRI can be a more efficient way than fully finer-grained or fully coarser-grained; using our method for both source language and target language can achieve better performance than using it for either source or target language. We do the same experiments on the Chinese-Mongolian SMT system and get the same conclusion. The experiments indicate that using finer-grained basic units for alignment and coarser-grained basic units for TRI performs better than other granularity combination. We also find that using Chinese characters contribute more than using Mongolian stems in Chinese-Mongolian SMT, which partly because of the errors brought by lemmatization. If we can combine more features (Elming and Habash, 2007) to do the realignment, and have a higher accuracy tool of lemmatization, our method can be better.

Acknowledgments

We thank PACLIC29 reviewers. This work is supported by the National Natural Science Foundation of China (No. 61362028) and the Postgraduate Scientific Research Innovation Foundation of Inner Mongolia (No. 1402020201).

Reference

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Peitra, Robert L. Mercer. 1993. The Mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263-311.

Pichuan Chang, Michel Galley, and Christopher D. Manning. 2008. Optimizing Chinese word segmentation for machine translation performance. *In proceedings of third workshop on SMT*, pages 224-232.

Yonggang Deng and Bowen Zhou. 2009. Optimizing word alignment combination for phrase table training. *In Proceedings* of *the ACL and the 4th IJCNLP of the AFNLP*, pages 229–232

Vilar, David, Maja Popovic, and Hermann Ney. 2006. AER: Do we need to “improve” our alignments? *In Proceedings of the International Workshop on Spoken Language Translation*, pages 205–212,

Jakob Elming and Nizar Habash. 2007. Combination of statistical word alignments based on multiple preprocessing schemes. *In Proceedings of the Association for Computational Linguistics*, pages 25-28.

Wenbin Jiang, Liang Huang, and Qun Liu. 2009. Automatic adaptation of annotation standards: Chinese word segmentation and POS tagging - a case study. *In Proceedings of ACL and the 4th IJCNLP of the AFNLP*, pages 522-530.

Hongxu Hou, Qun Liu, Nasanurtu. 2009. Mongolian word segmentation based on statistical language model. *Pattern Recognition and Artificial Intelligence*, 22(1): 108-112.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. *In ACL*

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. *In Proceeding of NAACL.*

Yang Liu, Qun Liu, and Shouxun Lin. 2005. Log-linear models for word alignment. *In ACL ’05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 459–466.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. *In Proceedings of the Association for Computational Linguistics*, pages 440-447.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1): 19-51.

Ning Xi, Guangchao Tang, Boyuan Li, and Yinggong Zhao. 2011. Word alignment combination over multiple word segmentation. *In Proceedings of the ACL2011 Student Session*, pages 1-5.

Ning Xi, Guangchao Tang, Xinyu Dai, Shujian Huang, Jiajun Chen. 2012. Enhancing Statistical Machine Translation with Character Alignment. *In Proceedings of the ACL2012*, pages 285-290.

Xinyan Xiao, Yang Liu, Young-Sook Hwang, Qun Liu, and Shouxun Lin. 2010. Joint tokenization and translation. *In Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1200-1208.

Ming Yu and Hongxu Hou. 2011. Researching of Mongolian word segmentation system based on dictionary, rules and language model. *Inner Mongolian University*.

Ruiqiang Zhang, Keiji Yasuda, and Eiichiro Sumita. 2008. Improved statistical machine translation by multiple Chinese word segmentation. *In Proceedings of the Third Workshop on Statistical Machine Translation*, pages 216-223.

Hongmei Zhao, Yajuan Lv, Guosheng Ben, Yun Huang, Qun Liu. 2012. Summary on CWMT2011 MT Translation Evaluation. *Journal of Chinese Information Processing*, 26(1): 22-30.