Enhancing Lemmatization for Mongolian and its Application to Statistical Machine Translation

Odbayar Chimeddorj Atsushi Fujii
Graduate School of Information Science and Engineering, Tokyo Institute of Technology
2-12-1 Ookayama, 152-8552, Japan
chimeddorj.o.aa AT m.titech.ac.jp

ABSTRACT

Lemmatization is crucial in natural language processing and information retrieval especially for highly inflected languages, such as Finnish and Mongolian. The state-of-the-art method of lemmatization for Mongolian does not need a noun dictionary and is scalable, but errors of this method are mainly caused by problems related to part of speech (POS) information. To resolve this problem, we integrate POS tagging and lemmatization for Mongolian. We evaluate the effectiveness of our method and its contribution to statistical machine translation.

KEYWORDS : Morphological segmentation, Lemmatization, Mongolian language, Statistical Machine Translation.

1 Introduction

In Mongolian, two different alphabets are used, Cyrillic and Mongolian. While the Cyrillic alphabet is mainly used in Mongolia, the Mongolian alphabet is mainly used in the Inner Mongolian Autonomous Region of China. Depending on the alphabet used, the writing system is also different in Mongolian. In this paper, we focus only on the Mongolian language that uses the Cyrillic alphabet, which will be termed "Mongolian" hereafter.

In Mongolian, which is an agglutinative language, each sentence is segmented on a phrase-by-phrase basis. A phrase consists of a content word, such as a noun or a verb, and one or more suffixes, such as postpositional participles. A content word can potentially be inflected when concatenated with suffixes.

Identifying the original forms of content words is crucial for natural language processing and information retrieval. In information retrieval, normalizing index terms can involve either lemmatization or stemming. Lemmatization identifies the original form of an inflected word, whereas stemming identifies a stem, which is not necessarily a word. Lemmatization is especially crucial for highly inflected languages, such as Finish and Mongolian. For example, one of the longest phrases in Mongolian "хамтралжуулагдсанаараа" consists of a stem (хам-), four derivational (-т -р -(а)л -ж) and five inflectional (-уул - (а)гд -сан -аар -аа) suffixes. This phrase is translated into 11 words in English as in the following example sentence.

Mongolian:	Тосгоныхон *хамтралжуулагдсанаараа* илүү сайн амьдрах болов.
English:	Village people, in that they were caused to be organized into collective farms, improved their lives.

Proceedings of the 10th Workshop on Asian Language Resources, pages 115–124,
COLING 2012, Mumbai, December 2012.

In this paper, we enhance an existing lemmatization method for Mongolian by using parts of speech annotation and apply our method to statistical machine translation for English to Mongolian.

2 Related work

Ehara et al. (2007) proposed a morphological analysis method for Mongolian-to-Japanese transfer-based machine translation. Ehara et al. manually produced Mongolian morphological inflectional rules, a suffix dictionary, and a lexicon for a morphological analyzer for Japanese. Their method uses these resources and lemmatizes an input phrase to generate its Japanese translation phrase by transferring the morphological structure.

Purev et al. (2005) proposed a method for morphological analysis targeting Mongolian using PC-KIMMO (Antworth, 1990). PC-KIMMO is based on a finite-state two-level morphological description approach (Koskenniemi, 1983). Purev et al. produced 36 two-level morphological rules for Mongolian, and used a lexicon consisting of 29,266 words (6,199 nouns, 18,551 verbs, and 4,516 adjectives) and 223 affixes. The accuracy of Purev et al's method for two novels was 60.5%. Errors were mainly due to out-of-dictionary words and contradictions between manually-written rules.

Sanduijav et al. (2005) proposed a lemmatization method for Mongolian verbs and nouns. This method uses a dictionary that was automatically produced by generating every possible combination of words and suffixes with manually-written morphological rules. Like Purev et al., this method also does not correctly lemmatize out-of-dictionary words.

Khaltar and Fujii (2009) proposed a state-of-the-art lemmatization method for Mongolian, which uses a suffix dictionary and a number of rules for suffix segmentation and vowel insertion. Unlike the above methods, this method does not need a noun dictionary and is therefore scalable. In addition, Khaltar and Fujii showed that their method experimentally outperformed Sanduijav et al. (2005). Therefore, we enhance Khaltar and Fujii's method with parts of speech information, and explain the method in details in following.

Given a phrase consisting of a content word and one or more suffixes, Khaltar and Fujii's method removes the suffixes and extracts the content word. In addition, the rules are used to identify the original form of the extracted content word. However, because details of the lemmatization process can vary depending on the part of speech (POS) for the target content word, a verb dictionary is used to determine whether the target content word is a verb or not. Because new verbs are created less frequently than nouns, they use a verb dictionary, but not a noun dictionary. Thus, this method is robust against out-of-dictionary words, compared with other existing methods.

However, Khaltar and Fujii's method is associated with three problems. First, their method often misrecognizes an out-of-dictionary verb as a noun and consequently lemmatizes the target phrase incorrectly. Second, their method incorrectly lemmatizes a content word that is associated with more than one POS. For example, a phrase "орон" is either a verb phrase consisting of "ор" (to enter) and "он" (serial verb suffix) or a noun phrase consisting of only a noun "орон" (country), as shown in examples (1) and (2), respectively. We also show an English translation below each sentence.

(1) дотогш ор+он алга болов
Verb+Suffix
(someone) went inside and disappeared
(2) олон орон цөмийн эрчим хүч хэрэглэдэг
Noun
many countries use nuclear energy

For another example, the word "хамгийн" in Mongolian means "most" in English, and its syntactic function is superlative for adjectives and adverbs. Because its lexical structure is same as "хамаг" (whole or all) + "-ийн" (genitive case), "хамгийн" can be misrecognized as a noun concatenated with an inflectional suffix. Third problem is related to phrases that have the same surface form and POS but different meaning and morphological structure. For example, the word "уусан" can be two different inflected verbs depending on the context, as shown below.

(3) уух + сан → уусан
to drink + past tense → drank
Би өчигдөр анх удаа япон ногоон цай уусан
Yesterday, I drank Japanese green tea for the first time.
(4) уусах + н → уусан
to fade/melt + serial verb suffix → faded and [another verb]
Мөс усанд уусан алга болов
Ice melted into water and disappeared.

In the above examples, knowing only the POS of "уусан" is insufficient to segment it correctly even consulting to the verb dictionary because both usages in (3) and (4) are verbs. Therefore, it is necessary to know its inflection from the sentence content.

3 Our method for lemmatization

To resolve the three problems associated with Khaltar and Fujii (2009) described in Section 2, we combine their method and POS tagging. For the first problem, we can use POS information to distinguish nouns and verbs in target phrases. For the second problem, we can identify the POS for an ambiguous word depending on the context and use the corresponding lemmatization process. For the third problem, the POS annotation used in our method includes inflectional structure for verbs and nouns. For example, a POS annotation for a noun phrase is distinguished whether it is inflected or not. If inflected, the POS annotation also carries inflection type such as plural, genitive, and possessive.

In practice, we perform POS tagging for an input sentence and then use Khaltar and Fujii's method to perform lemmatization on a phrase-by-phrase basis. Training the POS tagging needs only POS annotated corpus, instead it does not need lemmatization. Our method consists of three components: POS annotation of input sentence, extracting target phrases with their POS information and lemmatizing target phrase by Khaltar and Fujii's method. The procedure of our lemmatization method is shown in following with a step-by-step example.

Step 1. олон орон цөмийн эрчим хүч хэрэглэдэг
Many countries use nuclear energy

Step 2. олон орон цөмийн эрчим хүч хэрэглэдэг
JJ N NG N N VP

Step 3. олон орон цөмийн эрчим хүч хэрэглэдэг
nuclear+genitive to use+present
N NG N N VP

Step 4. олон орон цөм+ийн эрчим хүч хэрэглэ+дэг
N+genitive V+present

In the above example, a sentence in Mongolian is segmented through three steps. Step 1 is POS tagging on an input sentence. Step 2 is extracting target phrases with their POS information. In Step 3, the target phrases are lemmatized by Khaltar and Fujii's method by consulting with POS information. As shown in the example, two target phrases are identified according to POS annotation: цөмийн and хэрэглэдэг ("nuclear" and "to use" in English, respectively). The example also shows examples of (1) and (2) mentioned in Section 2. The phrase орон ("countries" in English) is incorrectly lemmatized as a verb instead of noun in Khaltar and Fujii's method.

4 Experiments

4.1 Overview

We conducted two separate experiments to evaluate our lemmatization method for Mongolian. In Section 4.2, our method is evaluated on lemmatizing verb and noun phrases, and the result is compared to the Khaltar and Fujii's method. In Section 4.3, we evaluate the effectiveness of Khaltar and Fujii's and our methods in statistical machine translation (SMT).

For POS tagging purpose, we used a statistical POS tagger "TnT" (Thorsten, 2000) and a 5 M word Mongolian corpus, in which each word is manually annotated with its POS tag and inflectional structure (Jaimai and Chimeddorj, 2008), for training purposes. This corpus consists of common domains such as laws, novels and news.

4.2 Evaluating lemmatization accuracy

In the evaluation of lemmatization, we used the same test data as in Khaltar and Fujii (2009), which consists of 183 newspaper articles (hereafter "News") and 1,467 technical abstracts (hereafter "Tech") for Mongolian. Furthermore, we targeted on the noun and verb phrases of the test data due to the most inflectional POS in Mongolian and the NLP and IR application. The amount of the targeted phrases is shown in Table 1.

Test data	Noun phrase		Verb phrase	
	In types	In tokens	In types	In tokens
News	5,201	14,538	5,086	11,723
Tech	15,982	73,625	4,797	37,477
Total	21,899	86,554	9,880	49,200

TABLE 1 – Target phrase types and tokens for the experiment.

As shown in Table 1, we targeted on 31,779 types of phrases of which 21,899 are noun phrases and 9,880 are verb phrases, respectively.

First, the test data was tagged with the TnT. We found that the accuracy for POS tagging was 93.8%. Second, the phrases shown in Table 4 were extracted from News and Tech with their POS tags, and each of them was given to the lemmatization method with its POS information. Finally, the result of the lemmatization was compared with human assessed correct answers. The total accuracy of Khaltar and Fujii's method was 73.4% while that of our method was 86.6%, as shown in Table 2. Furthermore, the accuracy of lemmatization for the Mongolian was improved substantially for verb phrases and slightly for noun phrases by using POS information.

Test data	Khaltar and Fujii			Our method		
	Nouns	Verbs	Total	Nouns	Verbs	Total
News	85.5%	54.2%	70.0%	89.5%	84.4%	86.9%
Tech	84.8%	43.3%	75.2%	86.1%	88.0%	86.5%
Total	84.9%	48.9%	73.4%	86.9%	86.1%	86.6%

TABLE 2 – Accuracy of lemmatization by phrase types.

As shown in Table 2, the accuracy of the lemmatization on the noun and verb phrase types is improved for the both domain (News by 16.9% and Tech by 11.3%). In addition, we evaluated the performance of our method on the total tokens of the test data (Table 3). As a result, the total improvements are 9.2% on News and 10.6% on Tech.

Test data	Khaltar and Fujii			Our method		
	Nouns	Verbs	Total	Nouns	Verbs	Total
News	87.1%	75.1%	81.7%	91.0%	90.8%	90.9%
Tech	96.0%	60.1%	83.8%	96.9%	84.0%	92.5%
Total	94.5%	63.6%	82.9%	96.9%	85.6%	93.5%

TABLE 3 – Accuracy of lemmatization by phrase tokens.

As shown in Tables 2 and 3, the results of our method are higher than that of Khaltar and Fujii's method in the both of phrase types and phrase tokens.

We manually analyzed the errors in our method, and found seven types of errors in lemmatizing noun phrases (Table 4), and six types of errors in lemmatizing verb phrases (Table 5), respectively.

Error (# in News/Tech)	Examples	Correct
(a) Incorrect suffix removal (221/831)	дайнд → дай noun + dative in the war	дайн war
(b) Incorrect vowel insertion (139/719)	үнийг → үн noun + accusative price	үнэ price
(c) Soft sign insertion (9/198)	сургуулийн → сургуули noun + genitive of school	сургууль school
(d) Irregular plural suffix (108/116)	охид → охид noun + plural girls	охин girl
(e) Special possessive suffix (84/218)	ахынхаа → ахын noun + genitive + possessive my brother's	ах brother
(f) POS ambiguity (21/77)	орноос → ор noun + ablative from country	орон country
(h) Incorrect POS tagging (63/268)	угаар → уг noun smoke	угаар smoke

TABLE 4 – Errors of our method for noun phrases.

Error (# in News/Tech)	Example	Correct
(i) Incorrect suffix removal (302/236)	ярьжээ → ярьж verb + past told (to tell)	ярь to tell
(j) Incorrect vowel insertion (86/69)	идэвхжисэн → идэвхэж verb + past activated	идэвхж to active
(k) Soft sign insertion (9/7)	дэвшиж→дэвшь verb + serial verb suffix advanced	дэвш to advance
(l) Ignored by POS tagging (198/148)	авчихлаа → авчих verb + past perfect + past have just taken	ав to take
(m) POS ambiguity (11/13)	үрж → үр verb multiply	үрж multiply
(n) Incorrect POS tagging (100/83)	уудагийг → уудаг verb + present + accusative case that it drinks	уу to drink

TABLE 5 – Errors of our method for verb phrases.

As shown in Table 4, the most dominant errors in the noun phrase lemmatization are (a) and (b). Error (a) is a suffix homonym problem. In Mongolian, many suffixes are similar in their surface form, but different in their meaning or its opposite (similar in their meaning and different in their surface form). For resolving Error (a), only POS and syntactical function (such as cases, plural, etc) information is insufficient. It needs more detail lexical information to recognize the suffix boundaries. Error (b) is caused by the contradiction among the vowel insertion rules and the irregular concatenation form as well. In addition, Error (c) is similar to the Error (b). For correct vowel insertion needs more linguistic analysis for appropriate rule descriptions. Errors (d) and (e) can be resolved by simple heuristics. Error (d) needs a dictionary for irregular nouns while Error (e) can be solved by extending the segmentation rule. In the previous method, the segmentation rule did not consider the special possessive suffix.

Errors (f) and (h) are related to the POS tagging process. Although some cases of POS ambiguity (mentioned in the section 2) are solved in this work, there are other more ambiguous phrases, which the POS tagging in this work is not enough to resolve. Furthermore, the incorrect POS tagged phrases lead to the inappropriate lemmatization process as causing the error (h).

As shown in Table 5, errors from (i) to (k) are the same problems as in the noun errors. The errors from (l) to (n) are related to the POS tagging. Errors (m) and (n) are also the same errors in the noun lemmatization while Error (l) is caused by that the POS tagging used in this work ignores some inflectional functions of verbs. As a result, such verb suffixes are not removed.

4.3 Evaluating the contribution of lemmatization to SMT

In this experiment, we evaluated the effectiveness of our lemmatization method for English-Mongolian (En-Mn) phrase-based SMT. Khaltar and Fujii's method was also evaluated for comparison. We used Moses (Koehn et al., 2007) with the standard configuration and GIZA++ (Och et al., 2003) with the grow-diag-final-and heuristic for word-alignment. Our parallel data set was collected from web sites (http://www.legalinfo.mn/ and http://mongolia.usembassy.gov/), and consists of law and news domains. Example En-Mn sentence pairs in our data are shown below.

En1: Occupational safety and health measures shall not involve any expenditure for the workers .
Mn1: Хөдөлмөрийн аюулгүй байдал , эрүүл ахуйн арга хэмжээтэй холбогдох аливаа зардлыг ажилчид хариуцахгүй .

En2: Agriculture even holds a key to delivering new forms of clean energy .
Mn2: Үүнтэй зэрэгцээд хөдөө аж ахуй нь цэвэр эрчим хүчний шинэ төрлийг бий болгоход ч голлох үүрэг гүйцэтгэж байна .

The numbers of sentence pairs for training a translation model, tuning parameters, and testing were 24 K, 2 K, and 500, respectively. We used SRILM (Stolcke et al., 2011) and a 5-gram word language model in Mongolian was produced from 106 K sentences in Mongolian.

We compared two types of SMT methods for English-Mongolian: an SMT with lemmatization for noun and verb phrases in Mongolian (WL) and an SMT without lemmatization (WOL). We used BLEU (Papineniet al., 2002) for evaluation purposes. While translations in Mongolian produced by WL were lemmatized inherently, translations by WOL and reference translations were not lemmatized. To compare BLEU values for WOL and WL strictly, we segmented the

translations by WOL and the reference translations using the same lemmatization method as WL. Table 6 shows BLEU values for different SMT methods.

WOL1	38.74
Khaltar and Fujii	38.43
WOL2	39.11
Our method	40.48

TABLE 6 – BLEU values for different SMT methods.

In Table 6, there are four SMT methods. Two of them are WLs ("Khaltar and Fujii" and "Our method") and the remaining methods are WOL1 and WOL2. While our lemmatization method was used in "Our method" and the output of WOL2, Khaltar and Fujii's method was used in "Khaltar and Fujii" and the output of WOL1. Looking at Table 6, the BLEU value for Khaltar and Fujii's method was smaller than that for WOL1. In other words, Khaltar and Fujii's method was not effective in terms of SMT. However, the BLEU value for our method was greater than that for WOL2. In addition, we performed a statistical testing (Koehn, 2004) and found that the difference between our method and WOL2 in BLEU was statistically significant with the 95% confidence level. We can conclude that our lemmatization for Mongolian was effective for English-Mongolian SMT.

Conclusion

In this paper, we proposed a lemmatization method, which identifies the original form of the content word in a Cyrillic Mongolian phrase. Although the state-of-the-art method does not need a noun dictionary and is therefore scalable, this method incorrectly lemmatizes out-of-dictionary verbs and words associated with more than one part of speech (POS). To resolve this problem, our method first performs statistical POS tagging for an input phrase and then performs the lemmatization. To evaluate the effectiveness of our method, we targeted noun and verb phrases in newspaper articles and technical abstracts. Experimental results showed that our method substantially improved the accuracy of the state-of-the-art lemmatization method. We also applied our lemmatization method to English-Mongolian SMT and showed that our lemmatization method improved BLEU values for SMT experimentally.

Future work includes improving lemmatization rules for special noun possessive suffixes and a dictionary for irregular plural nouns. In addition, more linguistic analysis is necessary for statistically resolving the vowel insertion and the suffix homonym problems. Further research is necessary to obtain more improvement over English-Mongolian SMT. It needs to determine the effective phrases for the segmentation of Mongolian.

References

Thorsten Brants (2000). *TnT – A Statistical Part-of-Speech Tagger*, In Proceedings of the Sixth Applied Natural Language Processing Conference.

Terumasa Ehara, Suzushi Hayata, and Nobuyuki Kimura (2007). *Mongolian to Japanese Machine Translation System – Focused on Translation Selection*, In Proceedings of the 2nd International Symposium on Information and Language Processing.

Antworth Evan (1990). *PC-KIMMO: A Two-level Processor for Morphological Analysis*, Summer Institute of Linguistics, Inc.

Purev Jaimai and Odbayar Chimeddorj (2008). *POS Tagging for Mongolian, In Proceedings of Sixth Workshop on Asian Language Processing*, IJCNLP2008.

Purev Jaimai, Tsolmon Zundui, Altangerel Chagnaa, and Cheol-Young Ock (2005). *PC-KIMMO-based Description of Mongolian Morphology*, In Proceeding of International Journal of Information Processing Systems, Vol. 1, No.1.

Badam-Osor Khaltar and Atsushi Fujii (2009). *A Lemmatization method for Mongolian and its application to indexing for information retrieval*, Information Processing & Management, Vol. 45, No.4, pp.438-451.

Philipp Koehn (2007). *Statistical Significance Tests for Machine Translation Evaluation*. In the Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain.

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 Proceedings of the 45th Annual Meeting of the ACL, Demonstration session, Prague, Czech Republic.

Koskenniemi Kimmo (1983). *Two-level morphology: a general computational model for word-form recognition and production*, Publication No. 11. Helsinki: University of Helsinki Department of General Linguistics.

Franz Josef Och, Hermann Ney (2003). *A Systematic Comparison of Various Statistical Alignment Models*, Computational Linguistics, volume 29, number 1, pp. 19-51.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. (2002). *Bleu: a method for automatic evaluation of machine translation*. In Proceedings of ACL2002, pages 311–318, Philadelphia, Pennsylvania.

Enkhbayar Sanduijav, Takehito Utsuro, and Satoshi Sato (2005). *Mongolian phrase generation and morphological analysis based on phonological and morphological constraints*, Journal of Natural Language Processing, Vol. 12, No. 5, pp.185-205. (In Japanese).

Andreas Stolcke, Jing Zheng, Wen Wang and Victor Abrash (2011). *SRILM at Sixteen: Update and Outlook*. In Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop, Waikoloa, Hawaii.