A Lemmatization Method for Modern Mongolian and its Application to Information Retrieval

Badam-Osor Khaltar **Atsushi Fujii**
Graduate School of Library, Information and Media Studies
University of Tsukuba
1-2 Kasuga Tsukuba, 305-8550, Japan
{khab23, fujii}@slis.tsukuba.ac.jp

Abstract

In Modern Mongolian, a content word can be inflected when concatenated with suffixes. Identifying the original forms of content words is crucial for natural language processing and information retrieval. We propose a lemmatization method for Modern Mongolian and apply our method to indexing for information retrieval. We use technical abstracts to show the effectiveness of our method experimentally.

1 Introduction

The Mongolian language is divided into Traditional Mongolian, which uses the Mongolian alphabet, and Modern Mongolian, which uses the Cyrillic alphabet. In this paper, we focus solely on the latter and use the word "Mongolian" to refer to Modern Mongolian.

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 in Mongolian text is crucial for natural language processing and information retrieval. In information retrieval, the process of normalizing index terms is important, and can be divided into lemmatization and stemming. Lemmatization identifies the original form of an inflected word, whereas stemming identifies a stem, which is not necessarily a word.

Existing search engines, such as Google and Yahoo!, do not perform lemmatization or stemming for indexing Web pages in Mongolian. Therefore, Web pages that include only inflected forms of a query cannot be retrieved.

In this paper, we propose a lemmatization method for Mongolian and apply our method to indexing for information retrieval.

2 Inflection types in Mongolian phrases

Nouns, adjectives, numerals, and verbs can be concatenated with suffixes. Nouns and adjectives are usually concatenated with a sequence of a plural suffix, case suffix, and reflexive possessive suffix. Numerals are concatenated with either a case suffix or a reflexive possessive suffix. Verbs are concatenated with various suffixes, such as an aspect suffix, a participle suffix, and a mood suffix.

Figure 1 shows the inflection types of content words in Mongolian phrases. In (a), there is no inflection in the content word "ном (book)", concatenated with the suffix "ын (the genitive case)". The content words are inflected in (b)-(e).

Type	Example
(a) No inflection	**ном + ын → номын** book + genitive case
(b) Vowel insertion	**ах + д → ахад** brother + dative case
(c) Consonant insertion	**байшин + ийн→ байшингийн** building + genitive case
(d) The letters "ь" or "и" are eliminated, and the vowel converts to "и"	**анги + аас → ангиас** return + ablative case
(e) Vowel elimination	**ажил + аас → ажлаас** work + ablative case

Figure 1: Inflection types of content words in Mongolian phrases.

Loanwords, which can be nouns, adjectives, or verbs in Mongolian, can also be concatenated with suffixes. In this paper, we define a loanword as a word imported from a Western language.

Because loanwords are linguistically different from conventional Mongolian words, the suffix concatenation is also different from that for conventional Mongolian words. Thus, exception rules are required for loanwords.

For example, if the loanword "**станц** (station)" is to be concatenated with a genitive case suffix, "**ын**" should be selected from the five genitive case suffixes (i.e., **ын**, **ийн**, **ы**, **ий**, and **н**) based on the Mongolian grammar. However, because "**станц** (station)" is a loanword, the genitive case "**ийн**" is selected instead of "**ын**", resulting in the noun phrase "**станцийн** (station's)".

Additionally, the inflection (e) in Figure 1 never occurs for noun and adjective loanwords.

3 Related work

Sanduijav et al. (2005) proposed a lemmatization method for noun and verb phrases in Mongolian. They manually produced inflection rules and concatenation rules for nouns and verbs. Then, they automatically produced a dictionary by aligning nouns or verbs with suffixes. Lemmatization for phrases is performed by consulting this dictionary.

Ehara et al. (2004) proposed a morphological analysis method for Mongolian, for which they manually produced rules for inflections and concatenations. However, because the lemmatization methods proposed by Sanduijav et al. (2005) and Ehara et al. (2004) rely on dictionaries, these methods cannot lemmatize new words that are not in dictionaries, such as loanwords and technical terms.

Khaltar et al. (2006) proposed a lemmatization method for Mongolian noun phrases that does not use a noun dictionary. Their method can be used for nouns, adjectives, and numerals, because the suffixes that are concatenated with these are almost the same and the inflection types are also the same. However, they were not aware of the applicability of their method to adjectives and numerals.

The method proposed by Khaltar et al. (2006) mistakenly extracts loanwords with endings that are different from conventional Mongolian words. For example, if the phrase "**экологийн** (ecology's)" is lemmatized, the resulting content word will be "**эколог**", which is incorrect. The correct word is "**экологи** (ecology)". This error occurs because the ending "**-ологи** (-ology)" does not appear in conventional Mongolian words.

In addition, Khaltar et al. (2006)'s method applies (e) in Figure 1 to loanwords, whereas inflection (e) never occurs in noun and adjective loanwords.

Lemmatization and stemming are arguably effective for indexing in information retrieval (Hull, 1996; Porter, 1980). Stemmers have been developed for a number of agglutinative languages, including Malay (Tai et al., 2000), Indonesian (Berlian Vega and Bressan, 2001), Finnish (Korenius et al., 2004), Arabic (Larkey et al., 2002), Swedish (Carlberger et al., 2001), Slovene (Popovič and Willett, 1992) and Turkish (Ekmekçioglu et al., 1996).

Xu and Croft (1998) and Melucci and Orio (2003) independently proposed a language-independent method for stemming, which analyzes a corpus in a target language and identifies an equivalent class consisting of an original form, inflected forms, and derivations. However, their method, which cannot identify the original form in each class, cannot be used for natural language applications where word occurrences must be standardized by their original forms.

Finite State Transducers (FSTs) have been applied to lemmatization. Although Karttunen and Beesley (2003) suggested the applicability of FSTs to various languages, no rule has actually been proposed for Mongolian. The rules proposed in this paper can potentially be used for FSTs.

To the best of our knowledge, no attempt has been made to apply lemmatization or stemming to information retrieval for Mongolian. Our research is the first serious effort to address this problem.

4 Methodology

4.1 Overview

In view of the discussion in Section 3, we enhanced the lemmatization method proposed by Khaltar et al. (2006). The strength of this method is that noun dictionaries are not required.

Figure 2 shows the overview of our lemmatization method for Mongolian. Our method consists of two segments, which are identified with dashed lines in Figure 2: "lemmatization for verb phrases" and "lemmatization for noun phrases".

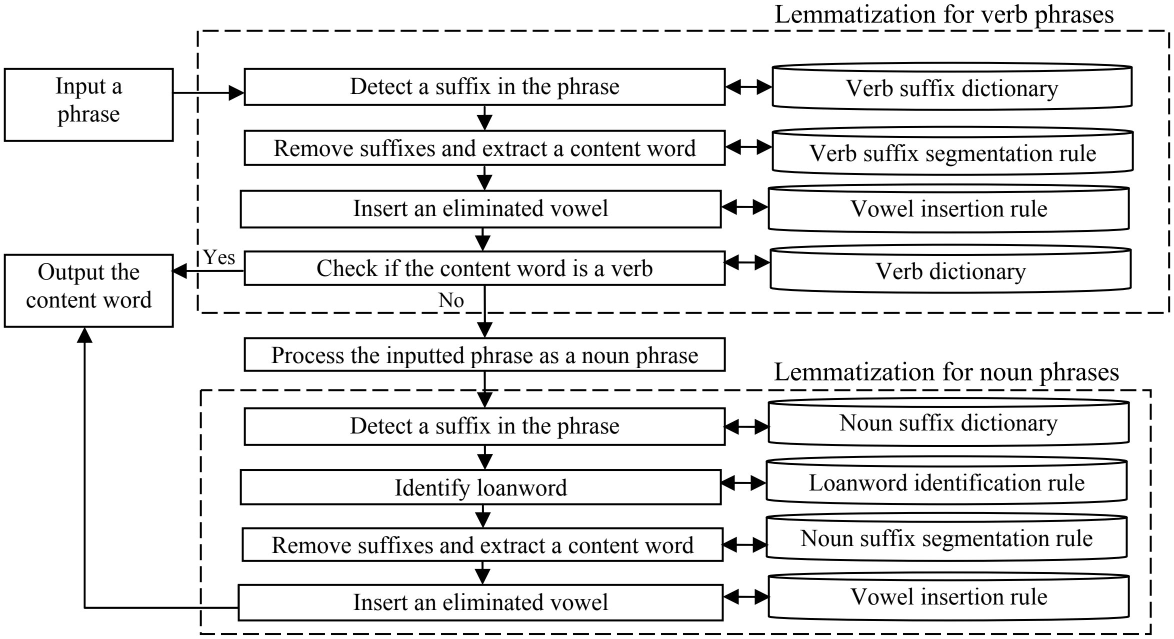

Lemmatization for verb phrases
Input a phrase
Detect a suffix in the phrase
Verb suffix dictionary
Remove suffixes and extract a content word
Verb suffix segmentation rule
Insert an eliminated vowel
Vowel insertion rule
Output the content word
Yes
Check if the content word is a verb
Verb dictionary
No
Process the inputted phrase as a noun phrase
Lemmatization for noun phrases
Detect a suffix in the phrase
Noun suffix dictionary
Identify loanword
Loanword identification rule
Remove suffixes and extract a content word
Noun suffix segmentation rule
Insert an eliminated vowel
Vowel insertion rule

Figure 2: Overview of our lemmatization method for Mongolian.

In Figure 2, we enhanced the method proposed by Khaltar et al. (2006) from three perspectives.

First, we introduced "lemmatization for verb phrases". There is a problem to be solved when we target both noun and verb phrases. There are a number of suffixes that can concatenate with both verbs and nouns, but the inflection type can be different depending on the part of speech. As a result, verb phrases can incorrectly be lemmatized as noun phrases and vice versa.

Because new verbs are not created as frequently as nouns, we predefine a verb dictionary, but do not use a noun dictionary. We first lemmatize an entered phrase as a verb phrase and then check whether the extracted content word is defined in our verb dictionary. If the content word is not defined in our verb dictionary, we lemmatize the input phrase as a noun phrase.

Second, we introduced a "loanword identification rule" in "lemmatization for noun phrases". We identify a loanword phrase before applying a "noun suffix segmentation rule" and "vowel insertion rule". Because segmentation rules are different for conventional Mongolian words and loanwords, we enhance the noun suffix segmentation rule that was originally proposed by Khaltar et al. (2006). Additionally, we do not use the vowel insertion rule, if the entered phrase is detected as a loanword phrase. The reason is that vowel elimination never occurs in noun loanwords.

Third, unlike Khaltar et al. (2006), we targeted adjective and numeral phrases. Because the suffixes concatenated with nouns, adjectives, and numerals are almost the same, the lemmatization method for noun phrases can also be used for adjective and numeral phrases without any modifications. We use "lemmatization for noun phrases" to refer to the lemmatization for noun, adjective, and numeral phrases.

We briefly explain our lemmatization process using Figure 2.

We consult a "verb suffix dictionary" and perform backward partial matching to determine whether a suffix is concatenated at the end of a phrase. If a suffix is detected, we use a "verb suffix segmentation rule" to remove the suffix and extract the content word. This process will be repeated until the residue of the phrase does not match any of the entries in the verb suffix dictionary.

We use a "vowel insertion rule" to check whether vowel elimination occurred in the content word and insert the eliminated vowel.

If the content word is defined in a "verb dictionary", we output the content word as a verb and terminate the lemmatization process. If not, we use the entered phrase and perform lemmatization for noun phrases. We consult a "noun suffix dictionary" to determine whether one or more suffixes are concatenated at the end of the target phrase.

We use a “loanword identification rule” to identify whether the phrase is a loanword phrase. We use a “noun suffix segmentation rule” to remove the suffixes and extract the content word. If the phrase is identified as a loanword phrase we use different segmentation rules.

We use the “vowel insertion rule” which is also used for verb phrases to check whether vowel elimination occurred in the content word and insert the eliminated vowel. However, if the phrase is identified as a loanword phrase, we do not use the vowel insertion rule.

If the target phrase does not match any of the entries in the noun suffix dictionary, we determine that a suffix is not concatenated and we output the phrase as it is.

The inflection types (b)–(d) in Figure 1 are processed by the verb suffix segmentation rule and noun suffix segmentation rule. The inflection (e) in Figure 1 is processed by the vowel insertion rule.

We elaborate on the dictionaries and rules in Sections 4.2–4.8.

4.2 Verb suffix dictionary

We produced a verb suffix dictionary, which consists of 126 suffixes that can concatenate with verbs. These suffixes include aspect suffixes, participle suffixes, and mood suffixes.

Figure 3 shows a fragment of our verb suffix dictionary, in which inflected forms of suffixes are shown in parentheses. All suffixes corresponding to the same suffix type represent the same meaning.

Suffix type	Suffix
Appeal	**гтун, гтүн**
Complete	**чих**
Perfect	**аад (иад), оод (иод), ээд, өөд**
Progressive-perfect	**саар, соор, сээр, сөөр**

Figure 3: Fragment of verb suffix dictionary.

4.3 Verb suffix segmentation rule

For the verb suffix segmentation rule, we produced 179 rules. There are one or more segmentation rules for each of the 126 verb suffixes mentioned in Section 4.2.

Figure 4 shows a fragment of the verb suffix segmentation rule for suffix “**в** (past)”. In the column “Segmentation rule”, the condition of each “if” sentence is a phrase ending. “V” refers to a vowel and “*” refers to any strings. “C9” refers to any of the nine consonants “**ц**”, “**ж**”, “**з**”, “**с**”, “**д**”, “**т**”, “**ш**”, “**ч**”, or “**х**”, and “C7” refers to any of the seven consonants “**м**”, “**г**”, ”**н**”, “**л**”, “**б**”, “**в**”, or “**р**”. If a condition is satisfied, we remove one or more corresponding characters.

Suffix	Segmentation rule
в Past	(i) If (*+ V + V + **в**) Remove **в**
	(ii) If (* + C9 + C7 + V + **в**) Remove V + **в**

Figure 4: Fragment of verb suffix segmentation rule.

For example, because the verb phrase “**шинэчлэв** (renew + past)” satisfies condition (ii), we remove the suffix “**в**” and the preceding vowel “**э**” to extract “**шинэчл**”.

4.4 Verb dictionary

We use the verb dictionary produced by Sanduijav et al. (2005), which includes 1254 verbs.

4.5 Noun suffix dictionary

We use the noun suffix dictionary produced by Khaltar et al. (2006), which contains 35 suffixes that can be concatenated with nouns. These suffixes are postpositional particles. Figure 5 shows a fragment of the dictionary, in which inflected forms of suffixes are shown in parentheses.

4.6 Noun suffix segmentation rule

There are 196 noun suffix segmentation rules, of which 173 were proposed by Khaltar et al. (2006). As we explained in Section 3, these 173 rules often incorrectly lemmatize loanwords with different endings from conventional Mongolian words.

We analyzed the list of English suffixes and found that English suffixes “-ation” and “-ology” are incorrectly lemmatized by Khaltar et al. (2006). In Mongolian, “-ation” is transliterated into “**аци**” or “**яци**” and “-ology” is transliterated into “**ологи**”. Thus, we produced 23 rules for loanwords that end with “**аци**”, “**яци**”, or “**ологи**”.

Figure 6 shows a fragment of our suffix segmentation rule for loanwords. For example, for the loanword phrase “**экологийн** (ecology + genitive)”, we use the segmentation rule for suffix “**ийн** (genitive)” in Figure 6. We remove the suffix “**ийн** (genitive)” and add “**и**” to the end of the content word. As a result, the noun “**экологи** (ecology)” is correctly extracted.

Case	Suffix
Genitive	**н, ы, ын, ий, ийн**
Accusative	**ыг, ийг, г**
Dative	**д, т**
Ablative	**аас (иас), оос (иос), ээс, өөс**

Figure 5: Fragment of noun suffix dictionary.

Suffix	Segmentation rule for loanwords
ийн Genitive	If (* + **логийн**) Remove (**ийн**) , Add (**и**)
ийг Accusative	If (* + **логийг**) Remove (**ийг**), Add (**и**)

Figure 6: Fragment of suffix segmentation rules for loanwords.

4.7 Vowel insertion rule

To insert an eliminated vowel and extract the original form of a content word, we check the last two characters of the content word. If they are both consonants, we determine that a vowel was eliminated. However, a number of Mongolian words end with two consonants inherently and, therefore, Khaltar et al. (2006) referred to a textbook on the Mongolian grammar (Ts, 2002) to produce 12 rules to determine when to insert a vowel between two consecutive consonants. We also use these rules as our vowel insertion rule.

4.8 Loanword identification rule

Khaltar et al. (2006) proposed rules for extracting loanwords from Mongolian corpora. Words that satisfy one of seven conditions are extracted as loanwords. Of the seven conditions, we do not use the condition that extracts a word ending with "consonants + **и**" as a loanword because it was not effective for lemmatization purposes in preliminary study.

5 Experiments

5.1 Evaluation method

We collected 1102 technical abstracts from the "Mongolian IT Park" [1] and used them for experiments. There were 178,448 phrase tokens and 17,709 phrase types in the 1102 technical abstracts. We evaluated the accuracy of our lemmatization method (Section 5.2) and the effectiveness of our method in information retrieval (Section 5.3) experimentally.

5.2 Evaluating lemmatization

Two Mongolian graduate students served as assessors. Neither of the assessors was an author of this paper. The assessors provided the correct answers for lemmatization. The assessors also tagged each word with its part of speech.

The two assessors performed the same task independently. Differences can occur between two assessors on this task. We measured the agreement of the two assessors by the Kappa coefficient, which ranges from 0 to 1. The Kappa coefficients for performing lemmatization and tagging of parts of speech were 0.96 and 0.94, respectively, which represents almost perfect agreement (Landis and Koch, 1977). However, to enhance the objectivity of the evaluation, we used only the phrases for which the two assessors agreed with respect to the part of speech and lemmatization.

We were able to use the noun and verb dictionaries of Sanduijav et al. (2005). Therefore, we compared our lemmatization method with Sanduijav et al. (2005) and Khaltar et al. (2006) in terms of accuracy.

Accuracy is the ratio of the number of phrases correctly lemmatized by the method under evaluation to the total number of target phrases. Here, the target phrases are noun, verb, adjective, and numeral phrases.

Table 1 shows the results of lemmatization. We targeted 15,478 phrase types in the technical abstracts. Our experiment is the largest evaluation for Mongolian lemmatization in the literature. In contrast, Sanduijav et al. (2005) and Khaltar et al. (2006) used only 680 and 1167 phrase types, respectively, for evaluation purposes.

In Table 1, the accuracy of our method for nouns, which were targeted in all three methods, was higher than those of Sanduijav et al. (2005) and Khaltar et al. (2006). Because our method and that of Sanduijav et al. (2005) used the same verb dictionary, the accuracy for verbs is principally the same for both methods. The accuracy for verbs was low, because a number of verbs were not included in the verb dictionary and were mistakenly lemmatized as noun phrases. However, this problem will be solved by enhancing the verb dictionary in the future. In total, the accuracy of our method was higher than those of Sanduijav et al. (2005) and Khaltar et al. (2006).

[1] http://www.itpark.mn/ (October, 2007)

Table 1: Accuracy of lemmatization (%).

	#Phrase types	Sanduijav et al. (2005)	Khaltar et al. (2006)	Our method
Noun	13,016	57.6	87.7	92.5
Verb	1,797	24.5	23.8	24.5
Adjective	609	82.6	83.5	83.9
Numeral	56	41.1	80.4	81.2
Total	15,478	63.2	72.3	78.2

We analyzed the errors caused by our method in Figure 7. In the column "Example", the left side and the right side of an arrow denote an error and the correct answer, respectively.

The error (a) occurred to nouns, adjectives, and numerals, in which the ending of a content word was mistakenly recognized as a suffix and was removed. The error (b) occurred because we did not consider irregular nouns. The error (c) occurred to loanword nouns because the loanword identification rule was not sufficient. The error (d) occurred because we relied on a verb dictionary. The error (e) occurred because a number of nouns were incorrectly lemmatized as verbs.

For the errors (a)-(c), we have not found solutions. The error (d) can be solved by enhancing the verb dictionary in the future. If we are able to use part of speech information, we can solve the error (e). There are a number of automatic methods for tagging parts of speech (Brill, 1997), which have promise for alleviating the error (e).

5.3 Evaluating the effectiveness of lemmatization in information retrieval

We evaluated the effectiveness of lemmatization methods in indexing for information retrieval. No test collection for Mongolian information retrieval is available to the public. We used the 1102 technical abstracts to produce our test collection.

Figure 8 shows an example technical abstract, in which the title is "Advanced Albumin Fusion Technology" in English. Each technical abstract contains one or more keywords. In Figure 8, keywords, such as "**цусны ийлдэс** (blood serum)" and "**эхэс** (placenta)" are annotated.

We used two different types of queries for our evaluation. First, we used each keyword as a query, which we call "keyword query (KQ)". Second, we used each keyword list as a query, which we call "list query (LQ)". The average number for keywords in the keywords list was 6.1. For each query, we used as the relevant documents the abstracts that were annotated with the query keyword in the keywords field. Thus, we were able to avoid the cost of relevance judgments.

Reasons of errors	#Errors	Example
(a) Word ending is the same as a suffix.	274	**сорт → сор** sort
(b) Noun plural tense is irregular.	244	**амьтан → амьт** animal
(c) Noun loanword ends with two consonants.	94	**динозавр → динозавар** dinosaur
(d) Verb does not exist in our verb dictionary.	689	**кодло → кодлох** to code
(e) Word corresponds to multiple part of speech.	853	**орон → ор** country inter

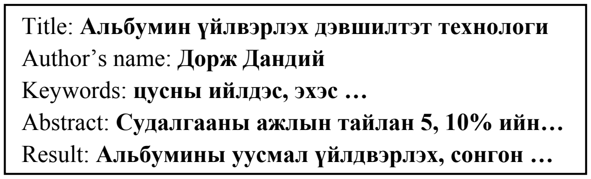
Figure 7: Errors of our lemmatization method.

The target documents are the 1102 technical abstracts, from which we extracted content words in the title, abstract, and result fields as index terms. However, we did not use the keywords field for indexing purposes. We used Okapi BM25 (Robertson et al., 1995) as the retrieval model.

We used the lemmatization methods in Table 2 to extract content words and compared the Mean Average Precision (MAP) of each method using KQ and LQ. MAP has commonly been used to evaluate the effectiveness of information retrieval. Because there were many queries for which the average precision was zero in all methods, we discarded those queries. There were 686 remaining KQs and 273 remaining LQs.

The average number of relevant documents for each query was 2.1. Although this number is small, the number of queries is large. Therefore, our evaluation result can be stable, as in evaluations for question answering (Voorhees and Tice, 2000).

We can derive the following points from Table 2. First, to clarify the effectiveness of the lemmatization in information retrieval, we compare "no lemmatization" with the other methods. Any lemmatization method improved the MAP for both KQ and LQ. Thus, lemmatization was effective for information retrieval in Mongolian. Second, we compare the MAP of our method with those of Sanduijav et al. (2005) and Khaltar et al. (2006). Our method was more effective than the method of Sanduijav et al. (2005) for both KQ and LQ. However, the difference between Khaltar et al. (2006) and our method was small for KQ and our method



Title: **Альбумин үйлвэрлэх дэвшилтэт технологи**

Author's name: **Дорж Дандий**

Keywords: **цусны ийлдэс, эхэс …**

Abstract: **Судалгааны ажлын тайлан 5, 10% ийн…**

Result: **Альбумины уусмал үйлдвэрлэх, сонгон …**

Figure 8: Example of technical abstract.

Table 2: MAP of lemmatization methods.

	Keyword query	List query
No lemmatization	0.2312	0.2766
Sanduijav et al. (2005)	0.2882	0.2834
Khaltar et al. (2006)	0.3134	0.3127
Our method	0.3149	0.3114
Correct lemmatization	0.3268	0.3187

was less effective than Khaltar et al.(2006) for LQ. This is because although we enhanced the lemmatization for verbs, adjectives, numerals, and loanwords, the effects were overshadowed by a large number of queries comprising conventional Mongolian nouns. Finally, our method did not outperform the method using the correct lemmatization.

We used the paired t-test for statistical testing, which investigates whether the difference in performance is meaningful or simply because of chance (Keen, 1992). Table 3 shows the results, in which "<" and "<<" indicate that the difference of two results was significant at the 5% and 1% levels, respectively, and "—" indicates that the difference of two results was not significant.

Looking at Table 3, the differences between no lemmatization and any lemmatization method, such as Sanduijav et al. (2005), Khaltar et al. (2006), our method, and correct lemmatization, were statistically significant in MAP for KQ. However, because the MAP value of no lemmatization was improved for LQ, the differences between no lemmatization and the lemmatization methods were less significant than those for KQ. The difference between Sanduijav et al. (2005) and our method was statistically significant in MAP for both KQ and LQ. However, the difference between Khaltar et al. (2006) and our method was not significant in MAP for both KQ and LQ. Although, the difference between our method and correct lemmatization was statistically significant in MAP for KQ, the difference was not significant in MAP for LQ.

Table 3: t-test result of the differences between lemmatization methods.

	Keyword query	List query
No lemmatization vs. Correct lemmatization	<<	<
No lemmatization vs. Sanduijav et al. (2005)	<<	—
No lemmatization vs. Khaltar et al. (2006)	<<	<
No lemmatization vs. Our method	<<	<
Sanduijav et al. (2005) vs. Our method	<<	<
Khaltar et al. (2006) vs. Our method	—	—
Our method vs. Correct lemmatization	<	—

6 Conclusion

In Modern Mongolian, 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 this paper, we proposed a lemmatization method for Modern Mongolian. We enhanced the lemmatization method proposed by Khaltar et al. (2006). We targeted nouns, verbs, adjectives, and numerals. We also improved the lemmatization for loanwords.

We evaluated our lemmatization method experimentally. The accuracy of our method was higher than those of existing methods. We also applied our lemmatization method to information retrieval and improved the retrieval accuracy.

Future work includes using a part of speech tagger because the part of speech information is effective for lemmatization.

References

Vinsensius Berlian Vega S N and Stéphane Bressan. 2001. Indexing the Indonesian Web: Language identification and miscellaneous issues. *Tenth International World Wide Web Conference, Hong Kong.*

Eric Brill. 1997. Natural Language Processing Using Very Large Corpora. Kluwer Academic Press.

Johan Carlberger, Hercules Dalianis, Martin Hassel, and Ola Knutsson. 2001. Improving Precision in Information Retrieval for Swedish using Stemming. *Proceedings of NODALIDA '01 - 13th Nordic Conference on Computational Linguistics.*

Terumasa Ehara, Suzushi Hayata, and Nobuyuki Kimura. 2004. Mongolian morphological analysis using ChaSen. *Proceedings of the 10th Annual Meeting of the Association for Natural Language Processing,* pp. 709-712. (In Japanese).

Çuna F. Ekmekçioglu, Michael F. Lynch, and Peter Willett. 1996. Stemming and n-gram matching for term conflation in Turkish texts. *Information Research News,* Vol. 7, No. 1, pp. 2-6.

David A. Hull. 1996. Stemming algorithms – a case study for detailed evaluation. *Journal of the American Society for Information Science and Technology*, Vol. 47, No. 1, pp. 70-84.

Lauri Karttunen and Kenneth R. Beesley. 2003. Finite State Morphology. *CSLI Publications*. Stanford.

Micheal E. Keen. 1992. Presenting results of experimental retrieval comparisons. *Information Processing and Management*, Vol. 28, No. 4, pp. 491-502.

Badam-Osor Khaltar, Atsushi Fujii, and Tetsuya Ishikawa. 2006. Extracting loanwords from Mongolian corpora and producing a Japanese-Mongolian bilingual dictionary. *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pp. 657-664.

Tuomo Korenius, Jorma Laurikkala, Kalervo Järvelin, and Martti Juhola. 2004. Stemming and Lemmatization in the Clustering of Finnish Text Documents. *Proceedings of the thirteenth Association for Computing Machinery international conference on Information and knowledge management.* pp. 625-633.

Richard J. Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, Vol. 33, No. 1, pp. 159-174.

Leah S. Larkey, Lisa Ballesteros, and Margaret E. Connel. 2002. Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-occurrence Analysis. *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval,* pp. 275-282.

Massimo Melucci and Nicola Orio. 2003. A Novel Method for Stemmer Generation Based on Hidden Markov Models. *Proceedings of the twelfth international conference on Information and knowledge management*, pp. 131-138.

Mirko Popovič and Peter Willett. 1992. The effectiveness of stemming for natural-language access to Slovene textual data. *Journal of the American Society for Information Science and Technology*, Vol. 43, No. 5, pp. 384-390.

Martin F. Porter. 1980. An algorithm for suffix stripping. *Program*, Vol. 14, No. 3, pp. 130-137.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1995. Okapi at TREC-3. *Proceedings of the Third Text REtrieval Conference, NIST Special Publication* 500-226. pp. 109-126.

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) .

Sock Y. Tai, Cheng O. Ong, and Noor A. Abdullah. 2000. On designing an automated Malaysian stemmer for the Malay language. *Proceedings of the fifth international workshop on information retrieval with Asian languages, Hong Kong,* pp. 207-208.

Bayarmaa Ts. 2002. Mongolian grammar for grades I-IV. (In Mongolian).

Ellen M. Voorhees and Dawn M. Tice. 2000. Building a Question Answering Test Collection. *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* pp. 200-207.

Jinxi Xu and Bruce W. Croft. 1998. Corpus-based stemming using co-occurrence of word variants. *ACM Transactions on Information Systems,* Vol. 16, No. 1, pp. 61-81.