Extracting loanwords from Mongolian corpora and producing a Japanese-Mongolian bilingual dictionary

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

Atsushi Fujii
Graduate School of Library, Information and Media Studies
University of Tsukuba
1-2 Kasuga Tsukuba, 305-8550
Japan
fujii@slis.tsukuba.ac.jp

Tetsuya Ishikawa
The Historiographical Institute
The University of Tokyo
3-1 Hongo 7-chome, Bunkyo-ku
Tokyo, 133-0033
Japan
ishikawa@hi.u-tokyo.ac.jp

Abstract

This paper proposes methods for extracting loanwords from Cyrillic Mongolian corpora and producing a Japanese–Mongolian bilingual dictionary. We extract loanwords from Mongolian corpora using our own handcrafted rules. To complement the rule-based extraction, we also extract words in Mongolian corpora that are phonetically similar to Japanese Katakana words as loanwords. In addition, we correspond the extracted loanwords to Japanese words and produce a bilingual dictionary. We propose a stemming method for Mongolian to extract loanwords correctly. We verify the effectiveness of our methods experimentally.

1 Introduction

Reflecting the rapid growth in science and technology, new words and technical terms are being progressively created, and these words and terms are often transliterated when imported as loanwords in another language.

Loanwords are often not included in dictionaries, and decrease the quality of natural language processing, information retrieval, machine translation, and speech recognition. At the same time, compiling dictionaries is expensive, because it relies on human introspection and supervision. Thus, a number of automatic methods have been proposed to extract loanwords and their translations from corpora, targeting various languages.

In this paper, we focus on extracting loanwords in Mongolian. The Mongolian language is divided into Traditional Mongolian, written using the Mongolian alphabet, and Modern Mongolian, written using the Cyrillic alphabet. We focused solely on Modern Mongolian, and use the word "Mongolian" to refer to Modern Mongolian in this paper.

There are two major problems in extracting loanwords from Mongolian corpora.

The first problem is that Mongolian uses the Cyrillic alphabet to represent both conventional words and loanwords, and so the automatic extraction of loanwords is difficult. This feature provides a salient contrast to Japanese, where the Katakana alphabet is mainly used for loanwords and proper nouns, but not used for conventional words.

The second problem is that content words, such as nouns and verbs, are inflected in sentences in Mongolian. Each sentence in Mongolian is segmented on a phrase-by-phase basis. A phrase consists of a content word and one or more suffixes, such as postpositional particles. Because loanwords are content words, then to extract loanwords correctly, we have to identify the original form using stemming.

In this paper, we propose methods for extracting loanwords from Cyrillic Mongolian and producing a Japanese–Mongolian bilingual dictionary. We also propose a stemming method to identify the original forms of content words in Mongolian phrases.

Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 657–664,
Sydney, July 2006.

2 Related work

To the best of our knowledge, no attempt has been made to extract loanwords and their translations targeting Mongolian. Thus, we will discuss existing methods targeting other languages.

In Korean, both loanwords and conventional words are spelled out using the Korean alphabet, called *Hangul*. Thus, the automatic extraction of loanwords in Korean is difficult, as it is in Mongolian. Existing methods that are used to extract loanwords from Korean corpora (Myaeng and Jeong, 1999; Oh and Choi, 2001) use the phonetic differences between conventional Korean words and loanwords. However, these methods require manually tagged training corpora, and are expensive.

A number of corpus-based methods are used to extract bilingual lexicons (Fung and McKeown, 1996; Smadja, 1996). These methods use statistics obtained from a parallel or comparable bilingual corpus, and extract word or phrase pairs that are strongly associated with each other. However, these methods cannot be applied to a language pair where a large parallel or comparable corpus is not available, such as Mongolian and Japanese.

Fujii et al. (2004) proposed a method that does not require tagged corpora or parallel corpora to extract loanwords and their translations. They used a monolingual corpus in Korean and a dictionary consisting of Japanese Katakana words. They assumed that loanwords in multiple countries corresponding to the same source word are phonetically similar. For example, the English word "system" has been imported into Korean, Mongolian, and Japanese. In these languages, the romanized words are "siseutem", "sistem", and "shisutemu", respectively.

It is often the case that new terms have been imported into multiple languages simultaneously, because the source words are usually influential across cultures. It is feasible that a large number of loanwords in Korean can also be loanwords in Japanese. Additionally, Katakana words can be extracted from Japanese corpora with a high accuracy. Thus, Fujii et al. (2004) extracted the loanwords in Korean corpora that were phonetically similar to Japanese Katakana words. Because each of the extracted loanwords also corresponded to a Japanese word during the extraction process, a Japanese–Korean bilingual dictionary was produced in a single framework.

However, a number of open questions remain from Fujii et al.'s research. First, their stemming method can only be used for Korean. Second, their accuracy in extracting loanwords was low, and thus, an additional extraction method was required. Third, they did not report on the accuracy of extracting translations, and finally, because they used Dynamic Programming (DP) matching for computing the phonetic similarities between Korean and Japanese words, the computational cost was prohibitive.

In an attempt to extract Chinese–English translations from corpora, Lam et al. (2004) proposed a similar method to Fujii et al. (2004). However, they searched the Web for Chinese–English bilingual comparable corpora, and matched named entities in each language corpus if they were similar to each other. Thus, Lam et al.'s method cannot be used for a language pair where comparable corpora do not exist. In contrast, using Fujii et al.'s (2004) method, the Katakana dictionary and a Korean corpus can be independent.

In addition, Lam et al.'s method requires Chinese–English named entity pairs to train the similarity computation. Because the accuracy of extracting named entities was not reported, it is not clear to what extent this method is effective in extracting loanwords from corpora.

3 Methodology

3.1 Overview

In view of the discussion outlined in Section 2, we enhanced the method proposed by Fujii et al. (2004) for our purpose. Figure 1 shows the method that we used to extract loanwords from a Mongolian corpus and to produce a Japanese–Mongolian bilingual dictionary. Although the basis of our method is similar to that used by Fujii et al. (2004), "Stemming", "Extracting loanwords based on rules", and "N-gram retrieval" are introduced in this paper.

First, we perform stemming on a Mongolian corpus to segment phrases into a content word and one or more suffixes.

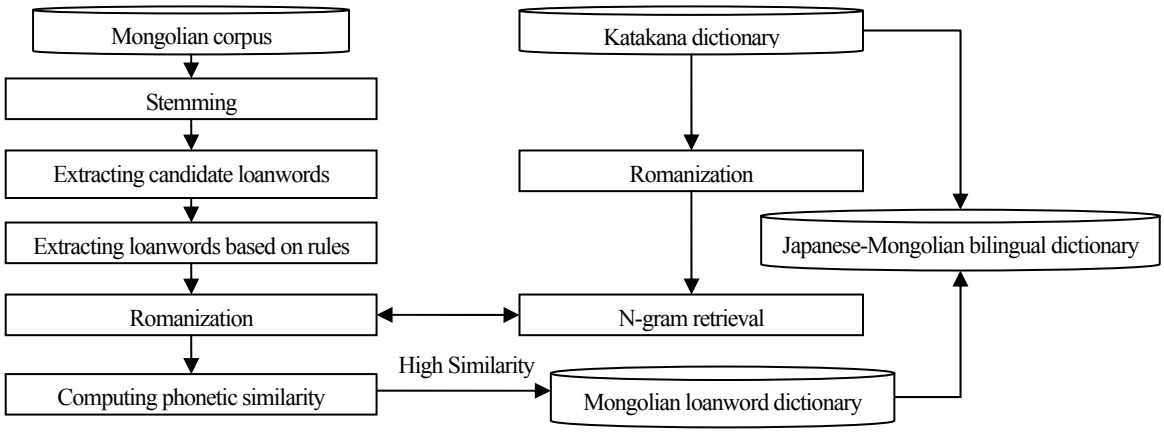

Figure 1: Overview of our extraction method.

Second, we discard segmented content words if they are in an existing dictionary, and extract the remaining words as candidate loanwords.

Third, we use our own handcrafted rules to extract loanwords from the candidate loanwords. While the rule-based method can extract loanwords with a high accuracy, a number of loanwords cannot be extracted using predefined rules.

Fourth, as performed by Fujii et al. (2004), we use a Japanese Katakana dictionary and extract a candidate loanword that is phonetically similar to a Katakana word as a loanword. We romanize the candidate loanwords that were not extracted using the rules. We also romanize all words in the Katakana dictionary.

However, unlike Fujii et al. (2004), we use N-gram retrieval to limit the number of Katakana words that are similar to the candidate loanwords. Then, we compute the phonetic similarities between each candidate loanword and each retrieved Katakana word using DP matching, and select a pair whose score is above a predefined threshold. As a result, we can extract loanwords in Mongolian and their translations in Japanese simultaneously.

Finally, to identify Japanese translations for the loanwords extracted using the rules defined in the third step above, we perform N-gram retrieval and DP matching.

We will elaborate further on each step in Sections 3.2–3.7.

3.2 Stemming

A phrase in Mongolian consists of a content word and one or more suffixes. A content word can potentially be inflected in a phrase. Figure 2 shows

Type	Example
(a) No inflection.	**ном + ын → номын** Book + Genitive Case
(b) Vowel elimination.	**ажил +аас+ аа→ ажлаасаа** Work + Ablative Case +Reflexive
(c) Vowel insertion.	**ах + д → ахад** Brother + Dative Case
(d) Consonant insertion.	**байшин + ийн→ байшингийн** Building + Genitive Case
(e) The letter "ь" is converted to "и", and the vowel is eliminated.	**сургууль+ аас→ сургуулиас** School + Ablative Case

Figure 2: Inflection types of nouns in Mongolian.

the inflection types of content words in phrases. In phrase (a), there is *no inflection* in the content word "ном (book)" concatenated with the suffix "ын (genitive case)".

However, in phrases (b)–(e) in Figure 2, the content words are inflected. Loanwords are also inflected in all of these types, except for phrase (b). Thus, we have to identify the original form of a content word using stemming. While most loanwords are nouns, a number of loanwords can also be verbs. In this paper, we propose a stemming method for nouns. Figure 3 shows our stemming method. We will explain our stemming method further, based on Figure 3.

First, we consult a "Suffix dictionary" and perform backward partial matching to determine whether or not one or more suffixes are concatenated at the end of a target phrase.

Second, if a suffix is detected, we use a "Suffix segmentation rule" to segment the suffix and extract

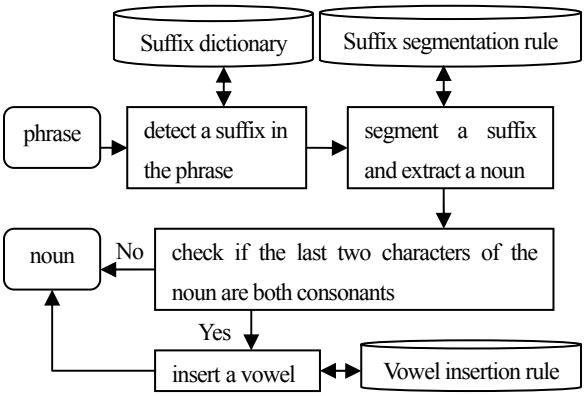

Figure 3: Overview of our noun stemming method.

the noun. The inflection type in phrases (c)–(e) in Figure 2 is also determined.

Third, we investigate whether or not the *vowel elimination* in phrase (b) in Figure 2 occurred in the extracted noun. Because the vowel elimination occurs only in the last vowel of a noun, we check the last two characters of the extracted noun. If both of the characters are consonants, the eliminated vowel is inserted using a "Vowel insertion rule" and the noun is converted into its original form.

Existing Mongolian stemming methods (Ehara et al., 2004; Sanduijav et al., 2005) use noun dictionaries. Because we intend to extract loanwords that are not in existing dictionaries, the above methods cannot be used. Noun dictionaries have to be updated as new words are created.

Our stemming method does not require a noun dictionary. Instead, we manually produced a suffix dictionary, suffix segmentation rule, and vowel insertion rule. However, once these resources are produced, almost no further compilation is required.

The suffix dictionary consists of 37 suffixes that can concatenate with nouns. These suffixes are postpositional particles. Table 1 shows the dictionary entries, in which the inflection forms of the postpositional particles are shown in parentheses.

The suffix segmentation rule consists of 173 rules. We show examples of these rules in Figure 4. Even if suffixes are identical in their phrases, the segmentation rules can be different, depending on the counterpart noun.

In Figure 4, the suffix "ийн" matches both the noun phrases (a) and (b) by backward partial matching. However, each phrase is segmented by a

Table 1: Entries of the suffix dictionary.

Case	Suffix
Genitive	**н, ы, ын, ны, ий, ийн, ний**
Accusative	**ыг, ийг, г**
Dative	**д, т**
Ablative	**аас (иас), оос (иос), ээс, өөс**
Instrumental	**аар (иар), оор (иор), ээр, өөр**
Cooperative	**тай, той, тэй**
Reflexive	**аа (иа), оо (ио), ээ, өө**
Plural	**уд (иуд), үүд (иүд)**

Suffix	Noun phrase	Noun
ийн Genitive	(a) **Ээж<u>ийн</u>** mother's	**ээж** mother
	(b) **Хараа<u>гийн</u>** Haraa'(river name)s	**Хараа** Haraa

Figure 4: Examples of the suffix segmentation rule.

deferent rule independently. The underlined suffixes are segmented in each phrase, respectively. In phrase (a), there is *no inflection*, and the suffix is easily segmented. However, in phrase (b), a *consonant insertion* has occurred. Thus, both the inserted consonant, "г", and the suffix have to be removed.

The vowel insertion rule consists of 12 rules. To insert an eliminated vowel and extract the original form of the noun, we check the last two characters of a target noun. If both of these are consonants, we determine that a vowel was eliminated.

However, a number of nouns end with two consonants inherently, and therefore, we referred to a textbook on Mongolian grammar (Bayarmaa, 2002) to produce 12 rules to determine when to insert a vowel between two consecutive consonants.

For example, if any of "м", "г", "л", "б", "в", or "р" are at the end of a noun, a vowel is inserted. However, if any of "ц", "ж", "з", "с", "д", "т", "ш", "ч", or "х" are the second to last consonant in a noun, a vowel is not inserted.

The Mongolian vowel harmony rule is a phonological rule in which female vowels and male vowels are prohibited from occurring in a single word together (with the exception of proper nouns). We used this rule to determine which vowel should be inserted. The appropriate vowel is determined by the first vowel of the first syllable in the target noun.

For example, if there are “а” and “у” in the first syllable, the vowel “а” is inserted between the last two consonants.

3.3 Extracting candidate loanwords

After collecting nouns using our stemming method, we discard the conventional Mongolian nouns. We discard nouns defined in a noun dictionary (Sanduijav et al., 2005), which includes 1,926 nouns. We also discard proper nouns and abbreviations. The first characters of proper nouns, such as “Эрдэнэбат (Erdenebat)”, and all the characters of abbreviations, such as “ЦШШНИ (Nuclear research centre)”, are written using capital letters in Mongolian. Thus, we discard words that are written using capital characters, except those occurring at the beginning of sentences. In addition, because “ө” and “ү” are not used to spell out Western languages, words including those characters are also discarded.

3.4 Extracting loanwords based on rules

We manually produced seven rules to identify loanwords in Mongolian. Words that match with one of the following rules are extracted as loanwords.

(a) A word including the consonants “к”, “п”, “ф”, or “щ”.

These consonants are usually used to spell out foreign words.

(b) A word that violated the Mongolian vowel harmony rule.

Because of the vowel harmony rule, a word that includes female and male vowels, which is not based on the Mongolian phonetic system, is probably a loanword.

(c) A word beginning with two consonants.

A conventional Mongolian word does not begin with two consonants.

(d) A word ending with two particular consonants.

A word whose penultimate character is any of: “п”, ”б”, “т”, ”ц”, “ч”, ”з”, or “ш” and whose last character is a consonant violates Mongolian grammar, and is probably a loanword.

(e) A word beginning with the consonant “в”.

In a modern Mongolian dictionary (Ozawa, 2000), there are 54 words beginning with “в”, of which 31 are loanwords. Therefore, a word beginning with “в” is probably a loanword.

(f) A word beginning with the consonant “р”.

In a modern Mongolian dictionary (Ozawa, 2000), there are 49 words beginning with “р”, of which only four words are conventional Mongolian words. Therefore, a word beginning with “р” is probably a loanword.

(g) A word ending with “<consonant> + и”.

We discovered this rule empirically.

3.5 Romanization

We manually aligned each Mongolian Cyrillic alphabet to its Roman representation[1].

In Japanese, the Hepburn and *Kunrei* systems are commonly used for romanization proposes. We used the Hepburn system, because its representation is similar to that used in Mongolian, compared to the *Kunrei* system.

However, we adapted 11 Mongolian romanization expressions to the Japanese Hepburn romanization. For example, the sound of the letter “L” does not exist in Japanese, and thus, we converted “L” to “R” in Mongolian.

3.6 N-gram retrieval

By using a document retrieval method, we efficiently identify Katakana words that are phonetically similar to a candidate loanword. In other words, we use a candidate loanword, and each Katakana word as a query and a document, respectively. We call this method “N-gram retrieval”.

Because the N-gram retrieval method does not consider the order of the characters in a target word, the accuracy of matching two words is low, but the computation time is fast. On the other hand, because DP matching considers the order of the characters in a target word, the accuracy of matching two words is high, but the computation time is slow. We combined these two methods to achieve a high matching accuracy with a reasonable computation time.

First, we extract Katakana words that are phonetically similar to a candidate loanword using N-gram retrieval. Second, we compute the similarity between the candidate loanword and each of the retrieved Katakana words using DP matching to improve the accuracy.

We romanize all the Katakana words in the dictionary and index them using consecutive N

[1] http://badaa.mngl.net/docs.php?p=trans_table (May, 2006)

characters. We also romanize each candidate loanword when use as a query. We experimentally set N = 2, and use the Okapi BM25 (Robertson et al., 1995) for the retrieval model.

3.7 Computing phonetic similarity

Given the romanized Katakana words and the romanized candidate loanwords, we compute the similarity between the two strings, and select the pairs associated with a score above a predefined threshold as translations. We use DP matching to identify the number of differences (i.e., insertion, deletion, and substitution) between two strings on an alphabet-by-alphabet basis.

While consonants in transliteration are usually the same across languages, vowels can vary depending on the language. The difference in consonants between two strings should be penalized more than the difference in vowels. We compute the similarity between two romanized words using Equation (1).

$$1 - \frac{2 \times (\alpha \times dc + dv)}{\alpha \times c + v} \quad (1)$$

Here, *dc* and *dv* denote the number of differences in consonants and vowels, respectively, and α is a parametric consonant used to control the importance of the consonants. We experimentally set $\alpha = 2$. Additionally, *c* and *v* denote the number of all the consonants and vowels in the two strings, respectively. The similarity ranges from 0 to 1.

4 Experiments

4.1 Method

We collected 1,118 technical reports published in Mongolian from the "Mongolian IT Park"[2] and used them as a Mongolian corpus. The number of phrase types and phrase tokens in our corpus were 110,458 and 263,512, respectively.

We collected 111,116 Katakana words from multiple Japanese dictionaries, most of which were technical term dictionaries.

We evaluated our method from four perspectives: "stemming", "loanword extraction", "translation extraction", and "computational cost." We will discuss these further in Sections 4.2-4.5, respectively.

4.2 Evaluating stemming

We randomly selected 50 Mongolian technical reports from our corpus, and used them to evaluate the accuracy of our stemming method. These technical reports were related to: *medical science* (17), *geology* (10), *light industry* (14), *agriculture* (6), and *sociology* (3). In these 50 reports, the number of phrase types including conventional Mongolian nouns and loanword nouns was 961 and 206, respectively. We also found six phrases including loanword verbs, which were not used in the evaluation.

Table 2 shows the results of our stemming experiment, in which the accuracy for conventional Mongolian nouns was 98.7% and the accuracy for loanwords was 94.6%. Our stemming method is practical, and can also be used for morphological analysis of Mongolian corpora.

We analyzed the reasons for any failures, and found that for 12 conventional nouns and 11 loanwords, the suffixes were incorrectly segmented.

4.3 Evaluating loanword extraction

We used our stemming method on our corpus and selected the most frequently used 1,300 words. We used these words to evaluate the accuracy of our loanword extraction method. Of these 1,300 words, 165 were loanwords. We varied the threshold for the similarity, and investigated the relationship between precision and recall. Recall is the ratio of the number of correct loanwords extracted by our method to the total number of correct loanwords. Precision is the ratio of the number of correct loanwords extracted by our method to the total number of words extracted by our method. We extracted loanwords using rules (a)–(g) defined in Section 3.4. As a result, 139 words were extracted.

Table 3 shows the precision and recall of each rule. The precision and recall showed high values using "All rules", which combined the words extracted by rules (a)–(g) independently.

We also extracted loanwords using the phonetic similarity, as discussed in Sections 3.6 and 3.7.

Table 2: Results of our noun stemming method.

	No. of each phrase type	Accuracy (%)
Conventional nouns	961	98.7
Loanwords	206	94.6

[2] http://www.itpark.mn/ (May, 2006)

Table 3: Precision and recall for rule-based loanword extraction.

Rules	(a)	(b)	(c)	(d)	(e)	(f)	(g)	All rules
Words extracted automatically	102	63	21	6	4	5	24	150
Extracted correct loanwords	101	60	20	5	4	5	19	139
Precision (%)	99.0	95.2	95.2	83.3	100	100	79.2	92.7
Recall (%)	61.2	36.4	12.1	3.0	2.4	3.03	11.5	84.2

We used the N-gram retrieval method to obtain up to the top 500 Katakana words that were similar to each candidate loanword. Then, we selected up to the top five pairs of a loanword and a Katakana word whose similarity computed using Equation (1) was greater than 0.6. Table 4 shows the results of our similarity-based extraction.

Both the precision and the recall for the similarity-based loanword extraction were lower than those for the "All rules" data listed in Table 3.

Table 4: Precision and recall for our similarity-based loanword extraction.

Words extracted automatically	Extracted correct loanwords	Precision (%)	Recall (%)
3,479	109	3.1	66.1

We also evaluated the effectiveness of a combination of the N-gram and DP matching methods. We performed similarity-based extraction after rule-based extraction. Table 5 shows the results, in which the data of the "Rule" are identical to those of the "All rules" data listed in Table 3. However, the "Similarity" data are not identical to those listed in Table 4, because we performed similarity-based extraction using only the words that were not extracted by rule-based extraction.

When we combined the rule-based and similarity-based methods, the recall improved from 84.2% to 91.5%. The recall value should be high when a human expert modifies or verifies the resultant dictionary.

Figure 5 shows example of extracted loanwords in Mongolian and their English glosses.

4.4 Evaluating Translation extraction

In the row "Both" shown in Table 5, 151 loanwords were extracted, for each of which we selected up to the top five Katakana words whose similarity computed using Equation (1) was greater than 0.6 as

Table 5: Precision and recall of different loanword extraction methods.

	No. of words	No. that were correct	Precision (%)	Recall (%)
Rule	150	139	92.7	84.2
Similarity	60	12	20.0	46.2
Both	210	151	71.2	91.5

Mongolian	English gloss
альбумин	albumin
лаборатор	laboratory
механизм	mechanism
митохондр	mitochondria

Figure 5: Example of extracted loanwords.

translations. As a result, Japanese translations were extracted for 109 loanwords. Table 6 shows the results, in which the precision and recall of extracting Japanese–Mongolian translations were 56.2% and 72.2%, respectively.

We analyzed the data and identified the reasons for any failures. For five loanwords, the N-gram retrieval failed to search for the similar Katakana words. For three loanwords, the phonetic similarity computed using Equation (1) was not high enough for a correct translation. For 27 loanwords, the Japanese translations did not exist inherently. For seven loanwords, the Japanese translations existed, but were not included in our Katakana dictionary.

Figure 6 shows the Japanese translations extracted for the loanwords shown in Figure 5.

Table 6: Precision and recall for translation extraction.

No. of translations extracted automatically	No. of extracted correct translations	Precision (%)	Recall (%)
194	109	56.2	72.2

Japanese	Mongolian	English gloss
アルブミン	альбумин	albumin
ラボラトリー	лаборатор	laboratory
メカニズム	механизм	mechanism
ミトコンドリア	митохондр	mitochondria

Figure 6: Japanese translations extracted for the loanwords shown in Figure 5.

4.5 Evaluating computational cost

We randomly selected 100 loanwords from our corpus, and used them to evaluate the computational cost of the different extraction methods. We compared the computation time and the accuracy of "N-gram", "DP matching", and "N-gram + DP matching" methods. The experiments were performed using the same PC (CPU = Pentium III 1 GHz dual, Memory = 2 GB).

Table 7 shows the improvement in computation time by "N-gram + DP matching" on "DP matching", and the average rank of the correct translations for "N-gram". We improved the efficiency, while maintaining the sorting accuracy of the translations.

Table 7: Evaluation of the computational cost.

Method	N-gram	DP	N-gram + DP
Loanwords	100		
Computation time (sec.)	95	136,815	293
Extracted correct translations	66	66	66
Average rank of correct translations	44.8	2.7	2.7

5 Conclusion

We proposed methods for extracting loanwords from Cyrillic Mongolian corpora and producing a Japanese–Mongolian bilingual dictionary. Our research is the first serious effort in producing dictionaries of loanwords and their translations targeting Mongolian. We devised our own rules to extract loanwords from Mongolian corpora. We also extracted words in Mongolian corpora that are phonetically similar to Japanese Katakana words as loanwords. We also corresponded the extracted loanwords to Japanese words, and produced a Japanese–Mongolian bilingual dictionary. A noun stemming method that does not require noun dictionaries was also proposed. Finally, we evaluated the effectiveness of the components experimentally.

References

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).

Atsushi Fujii, Tetsuya Ishikawa, and Jong-Hyeok Lee. 2004. Term extraction from Korean corpora via Japanese. *Proceedings of the 3rd International Workshop on Computational Terminology*, pp. 71-74.

Pascal Fung and Kathleen McKeown. 1996. Finding terminology translations from non-parallel corpora. *Proceedings of the 5th Annual Workshop on Very Large Corpora,* pp. 53-87.

Wai Lam, Ruizhang Huang, and Pik-Shan Cheung. 2004. Learning phonetic similarity for matching named entity translations and mining new translations. *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* pp. 289-296.

Sung Hyun Myaeng and Kil-Soon Jeong. 1999. Back-Transliteration of foreign words for information retrieval. *Information Processing and Management,* Vol. 35, No. 4, pp. 523 -540.

Jong-Hooh Oh and Key-Sun Choi. 2001. Automatic extraction of transliterated foreign words using hidden markov model. *Proceedings of the International Conference on Computer Processing of Oriental Languages,* 2001, pp. 433-438.

Shigeo Ozawa. Modern Mongolian Dictionary. Daigakushorin. 2000.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1995. Okapi at TREC-3, *Proceedings of the Third Text REtriev*al *Conference (TREC-3), 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) .

Frank Smadja, Vasileios Hatzivassiloglou, Kathleen R. McKeown. 1996. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, Vol. 22, No. 1, pp. 1-38.

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