Mongolian Inflection Suffix Processing in NLP: A Case Study

Xiangdong Su, Guanglai Gao(✉), Yupeng Jiang, Jing Wu, and Feilong Bao

College of Computer Science, Inner Mongolia University,
Hohhot 010021, People's Republic of China
csggl@imu.edu.cn

Abstract. Inflection suffix is an important morphological characteristic of Mongolian words, since the suffixes express abundant syntactic and semantic meanings. In order to provide an informative introduction of it, this paper implements a case study on it. Through three Mongolian NLP tasks, we disclose the following information: (1) views of inflection suffix in NLP tasks, (2) Inflection suffix processing ways, (3) Inflection suffix effects on system performance and (4) some suffix related conclusion.

Keywords: Mongolian · Inflection suffix · NLP · Case study

1 Introduction

Mongolian is an agglutinative language which normally ranks as a member of the Altaic language family, a family whose principal members are Turkish, Mongolian and Manchu (with Korean and Japanese listed as possible relations). There are about seven million Mongol speakers in the world, including two million in Mongolia, more than twice that number in Inner Mongolia and other parts of China, and another half million or so in the Buryat and Kalmyk Republic and elsewhere in Russia. The Mongols have written their language in several different scripts, the oldest and most durable of which, called the classical Mongol script, was introduced almost 800 years ago under Genghis Khan. It originated with the Sogdo-Uighur alphabet and has been revised several times. It derives the traditional Mongolian, which is used in Inner Mongolian Autonomous Region of China now. Their differences mainly concentrate on the following aspects [1]. Firstly, the glyphs of the same character between them are completely different even in the same position of one word. Secondly, the glyphs of the traditional Mongolian characters have diacritics, such as dots, while the corresponding glyphs of the classical Mongolian characters do not. The Mongolian used in Mongolia is called Cyrillic Mongolian, which was introduced in the 1940's [2]. It has the same pronunciation as traditional Mongolian. But its written form differs from traditional Mongolian. It uses the letters from the Russian alphabet. This paper introduces Mongolian according tradition Mongolian used in Inner Mongolian Autonomous Region of China. Without otherwise specified, Mongolian means traditional Mongolian in this paper.

J. Li et al. (Eds.): NLPCC 2015, LNAI 9362, pp. 347–352, 2015.
DOI: 10.1007/978-3-319-25207-0_29

The main features of Mongolian are a system of vowel harmony, agglutination and the SOV word order. Mongolian does not distinguish gender, has no definite article, and has only a very limited plural system. As other Altaic languages, there is hardly any difference between nouns and adjectives. There are 30 Mongolian characters. Each character has as many as three different glyphs (visual forms) depending on whether the character appears in an initial, medial, or final position of a word [3]. In some cases, some characters have the same glyph. In appearance, some glyphs are a part of other glyphs. The characters in words are connected along the baselines. Mongolian documents conform to a top-down writing style. The column order is from left to right.

Mongolian words are formed by attaching suffixes to roots. The suffix falls into two groups: derivational suffix and inflection suffix. Derivational suffix is also called the word-building suffix. They are added to the root and give the original words new meanings. The root adding one or more derivation suffixes is called a stem. Inflection suffix is also called word-changing suffix. They are added to the stems and give the original words grammatical meanings. These suffixes serve to integrate a word into sentence. Since the role difference between derivation suffix and inflection suffix, many NLP tasks pay more attention to the inflection suffix processing. The motivation of this paper elaborates how inflection suffix works in Mongolian NLP through a case study. The following aspects will be discussed: (1) views of inflection suffix in NLP tasks, (2) inflection suffix processing ways, (3) case effects on system performance and (4) some suffix related conclusion. The case study is taken out on Mongolian translation, Mongolian syntactic analysis, and Mongolian information retrieval. The experiments reveal some important points. We also refer some related works.

The remainder of this paper is organized as follows. Section 2 describes inflection suffix. Section 3 describes inflection suffix in Mongolian Syntactic Analysis. Section 4 describes inflection suffix in Mongolian information retrieval. Section 5 describes inflection suffix in Mongolian machine translation. Section 6 draws the conclusion.

2 Inflection Suffix

Inflectional suffix can be divided into plural suffixes, case suffixes, reflexive suffixes, voice suffixes, aspect suffixes and mood suffixes. Among this 6 group suffixes above, plural suffixes, case suffixes and reflexive suffixes only for nominal class; Voice suffixes, aspect suffixes and mood suffixes only for verb class. The nominal class includes noun, adjective, numeral, time-place word and pronoun. Figure 1 shows the structure of nominal word. The case suffixes express the relationship between two words and phrases, and the reflexive suffixes express the possessive relationship between two parts of sentence.

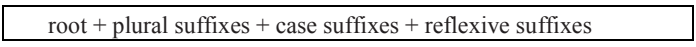

Fig. 1. The structure of nominal word

Case suffixes, reflexive suffixes and plural suffixes connected to the stem through a Narrow Non-Break Space (U+202F, Latin.:"-") [4]. For instance, in a word "ᠤᠯᠠᠭᠠᠨᠪᠠᠭᠠᠲᠤᠷ᠎ᠤᠨ" (transliteration: "vlaganbagatvr-vn", means: "Ulaanbaatar's"), the "ᠤᠨ" (-vn) is case suffixes. Case suffixes in Mongolian divided into eight types. These are Nominative Case, Accusative Case, Genitive Case, Dative-Locative Case, Ablative Case, Comitative Case, Instrumental Case and Directive Case. For case grammar, one stem is allowed to append one or two case(s) to from a new word. Table 1 demonstrates some examples of case suffixes.

Table 1. Examples of case suffixes

Stem	Case	Type	English Translation
ᠤᠯᠠᠭᠠᠨᠪᠠᠭᠠᠲᠤᠷ (vlaganbagatvr)	ᠤᠨ(-vn)	Genitive	Ulaanbaatar's
	ᠳᠤ(-tv)	Dative-Locative	in Ulaanbaatar
	ᠠᠴᠠ(-aqa)	Ablative	from Ulaanbatar
ᠤᠪᠠᠮᠠ ᠳᠡᠭᠡᠭᠦ (wbama deguu)	ᠲᠠᠢ ᠪᠠᠨ(-tai-ban)	Comitative	Obama with his brother
ᠰᠤᠷᠪᠤᠨ_ᠠ (svrvn_a)	ᠢ(-yi)	Accusative	let Surna
	ᠪᠡᠷ(-ber)	Instrumental	through Surna

The reflexive suffix helps to integrate nominal word with nominal syntagma, nominal word with verb syntagma or verbal word with verbal syntagma in a sentence coupling them to the subject.

3 Inflection Suffix in Mongolian Syntactic Analysis

Case is one kind of inflection suffix. Case inflection is the phenomenon that adding a case to a noun, adjective, or pronoun that expresses the syntactic relation of the word to other words in the sentence. Take the sentence "ᠪᠢ ᠮᠣᠩᠭᠣᠯ ᠬᠡᠯᠡ᠄ᠵᠢ ᠰᠤᠷᠴᠤᠯᠵᠤ ᠪᠠᠢᠨ᠎ᠠ ᠃" for example. Here, the word "ᠵᠢ" in the sentence is called accusative case (a constituent which appears after "ᠰᠤᠷᠴᠤᠯᠵᠤ" and indicates that "ᠰᠤᠷᠴᠤᠯᠵᠤ" is an object); the word "ᠪᠢ" in is called nominative case (a constituent which appears after "ᠪᠢᠴᠢᠭᠴᠢ" and indicates that "ᠪᠢᠴᠢᠭᠴᠢ" is a subject). In Mongolian grammar, the attached stem and the case are considered as a whole word. However, treating them individually bring significant benefit to syntactic analysis.

In order to utilize case information, X. Su et al. in [5] treated the stem and the case suffix in a word as two parts in syntactic analysis and annotated them individually in dependency treebank development. The annotation types of the cases take the case function into consideration. The treebank includes eight kinds of analytical dependencies which relate to the case units.

Meanwhile, X. Su et al. in [6] carried out an experiment on the Mongolian dependency treebank, which contains 400 sentences (13028 annotated words) from Inner Mongolian daily. There are 4460 distinct words and 1548 distinct stems. Labeled attachment score and dependency accuracy achieve 85.0% and 82.6% individually.

4 Inflection Suffix in Mongolian Information Retrieval

There are a few works focusing on the Mongolian information. Part of them investigated the suffix effect in this NLP task. G. Gao in [7] removed all the inflectional suffixes (includes "case") in the words before indexing since they assumed that stemming the Mongolian can not only effectively improve search efficiency, but also reduce index storage space. They constructed a stem and suffix dictionary and used it to remove the inflection suffixes. The dictionary includes 299 suffixes. Since Mongolian words include multiple suffixes, Y. Jin in [8] took three different strategies to processing the corpus before indexing, including (1) no suffix processing, (2) removing the inflection suffixes (including "case"), and (3) removing all the suffixes (derivation suffix and inflection suffix). They compared the retrieval performances when different suffix processing strategies were used. J. Yue in [9] calculated the occurrence of inflection suffix in Mongolian corpus, and used this information to construct the suffix set.

For the two words which have the same stem and different cases, they should be treated as relevance in information retrieval. Suppose that we do not remove the cases before indexing, it is happened that, when we use one of them as the query word, the retrieval result will not include the other one. This is unexpected. Therefore, removing the case in preprocessing section can solve this problem.

In the study, we carry out an experiment to test the effect of inflection suffix in Mongolian information. 27345 Mongolian documents are collected from the Inner Mongolian Newswire. We process the corpus with three methods: (1) no suffix processing, (2) removing inflection suffixes and (3) removing all suffixes. Then, we construct three indexes using the resultant corpus after preprocessing. The index number is 125796 before suffix processing. And the number of index terms decreased to 74001 when inflection suffix processing is used. It is suggested that Mongolian suffix processing can effectively reduce the number of term. And the retrieval efficiency improved in some extent.

The retrieval performance is listed in Table 2. It is clear that inflectional suffix process achieves the best performance at different recall level. This implies that the inflection suffixes should be removed in Mongolian information retrieval. However, derivation suffixes should not be removed, since they give the stems new meanings.

5 Inflection Suffix in Mongolian Translation

As noted, among the many kinds of suffixes, inflectional suffix changes the formation of the word but does not change the meaning of the word. In the process of machine translation, removing the control marks (U202f) is necessary. For an example: a noun will append a case suffix when it acts as a subject, but append another case suffix when it acts as an object. We have to remove the control marks to get the stem of the word to reveal the word semantics, and to reduce the extent of data sparseness. For the template-based machine translation, we need to remove the control marks and the suffixes follow by them. That is because we have to match the Mongolian words in

Mongolian-Chinese dictionary, only if we get off the suffixes of plural and subordinative relationship, the words can be matched correctly.

Table 2. The effectiveness of inflection processing

Recall	No Suffix Processing	Removing Inflection Suffix	Removing All Suffixes
0.0	1	1	1
0.1	0.8101	0.8155	0.8155
0.2	0.7791	0.7992	0.8012
0.3	0.7523	0.7712	0.7712
0.4	0.7114	0.7231	0.7210
0.5	0.6543	0.7012	0.6980
0.6	0.5367	0.5958	0.5731
0.7	0.4367	0.5120	0.4872
0.8	0.3214	0.3964	0.3823
0.9	0.1928	0.2428	0.2516
1.0	0.1002	0.1484	0.1329
Avg	0.5723	0.6096	0.6031

For statistical machine translation, there are two ways to process the U202f control mark. One is to remove the control marks and the suffixes following them directly; the other one is to replace the control mark with space, then the word and the suffix will be taken as two separate words. Both of the way can reveal the real meaning of the words, and alleviate data sparse to get better performance of alignment and translation. The second one will take more abundant information to interference the alignment and translation. We conduct a translation experiment with statistical machine translation mode to evaluate these two ways. The test set includes 1000 bilingual sentence pairs.

Table 3 lists the evaluation result. It is clear that treating the inflection suffix as individual unit makes the system perform best. It also proves that inflection suffix is quite useful in Mongolian translation. Comparing with no inflection suffix processing, removing the inflection suffix slight improves the system. It reduces the sparsity of data. But removing the derivation suffixes changes the meanings of the words.

Table 3. The performance comparison of inflection processing in Mongolian-Chinese translation

Method	BLEU
no inflection suffix processing	21.88
removing the inflection suffix	21.96
treating inflection suffix as individual unit	23.49

6 Conclusion

This paper investigates the Mongolian inflection suffix in NLP tasks. The key findings are as follows: Firstly, inflection suffixes express abundant syntactic meanings. Especially, case suffixes play important roles in syntactic analysis. Secondly, there are three common ways to process the inflection suffix in NLP tasks, including treating them as parts of words, treating them as individual syntactic units, and removing them in the preprocessing section of NLP task. Thirdly, removing the inflection suffix brings a slight improvement in Mongolian information retrieval. Using the inflection suffixes as individual units improves the performance of translation. Furthermore, the inflection suffix inflection can be integrating into more Mongolian NLP tasks. Inflection suffix deserves more investigation to make full of its syntactic functions.

Reference

1. Wei, H., Gao, G.: A Keyword Retrieval System for Historical Mongolian Document Images. International Journal on Document Analysis and Recognition (IJDAR) **17**, 33–45 (2014)
2. Tserenpil, D., Kullmann, R.: Mongolian Grammar. Admon Co. Ltd., Mongolia (2008)
3. Qinggeertai: Traditional Mongolian grammar. Inner Mongolian Press, Huhhot, China (1992)
4. Quejingzabu: Mongolian Coding. Inner Mongolia University Press (2000)
5. Su, X., Gao, G., Yan, X.: Development of traditional mongolian dependency treebank. In: Sun, M., Zhang, M., Lin, D., Wang, H. (eds.) CCL and NLP-NABD 2013. LNCS, vol. 8202, pp. 247–256. Springer, Heidelberg (2013)
6. Su, X., Gao, G., Yan, X.: Dependency parsing for traditional mongolian. In: Proceedings of the 2013 International Conference on Asian Language Processing, pp. 181–184. IEEE Computer Society, Washington, DC (2013)
7. Gao, G., Jin, W., Long, F., Hou, H.: A first investigation on mongolian information retrieval. In: Proceedings of the 2nd International Workshop on Evaluating Information Access (EVIA) (2008)
8. Jin, Y.: Research of Mongolian Information Retrieval Model Based on Markvo Random Field. Inner Mongolia University (2011)
9. Yue, J.: Study on the Methods in the Selection of Retrieval Unit in Mongolian Information Retrieval System. Inner Mongolia University (2011)