Language Model for Mongolian Polyphone Proofreading

Min Lu, Feilong Bao✉ and Guanglai Gao

College of Computer Science, Inner Mongolia University, Hohhot 010021, China
csfeilong@imu.edu.cn

Abstract. Mongolian text proofreading is the particularly difficult task because of its unique polyphonic alphabet, morphological ambiguity and agglutinative feature, and coding errors are currently pervasive in the Mongolian corpus of electronic edition, which results in Mongolian statistic and retrieval research toughly difficult to carry out. Some conventional approaches have been proposed to solve this problem but with limitations by not considering proofreading of polyphone. In this paper, we address this problem by means of constructing the large-scale resource and conducting n-gram language model based approach. For ease of understanding, the entire proofreading system architecture is also introduced in this paper, since the polyphone proofreading is the important component of it. Experimental results show that our method performs pretty well. Polyphone correction accuracy is relatively improved by 62% and overall system accuracy is relatively promoted by 16.1%.

Keywords: Mongolian, Polyphone, Automatic Proofreading System, Morphological Ambiguity.

1 Introduction

Coding errors are more seriously and universally presented in Mongolian corpus of electronic edition than other languages, which directly affects the development of Mongolian information technology such as Mongolian Named Entity Recognition (NER) [1], machine translation [2], Mongolian speech recognition [3], etc. From the objective perspective, the main reason is that Mongolian is an easily mistaken language due to its unique polyphonic alphabet, i.e. Mongolian letters and presentations are not corresponded one by one. Commonly, words written by different spellings can present one surface form if the intended word replaced by the letters with same shape. By the naked eye, we cannot judge from their appearances whether they are written in correct spelling or malapropism. Taking the Mongolian word "ᠲᠠᠭᠠ" (meaning: hair) and "ᠠᠮᠤ" (meaning: ocean, waist belt) for example, the former one can be written in four different spellings. When presented in there national Latin transliteration (keyboard correspondence), they are spelled as "usu", "uso", "oso", "osu" respectively, among which only the first spelling "usu" (meaning: hair) is correct. The latter one "ᠠᠮᠤ" can be spelled in 8 different ways of "talai","dalai","telei","delei","talei","delai","dalei","delei". Unlike the former instance, however, there are two correct spellings of "dalai" (meaning: ocean) and "telei"

(meaning: waist belt) corresponded to the latter form "ᠲᠠᠢ". This kind of word is called polyphone. i.e. one form corresponds to multiple spellings, pronunciation and meanings as well. If the intended word "dalai" occurred in the phrase "ᠪᠠᠭᠠᠮᠤ ᠲᠠᠢ" (meaning: Pacific Ocean) is replaced by the other correct spelling "telei", it is considered making misusing polyphone mistake. To sum up, Mongolian typo errors are frequently committed mainly because that typist usually just care about the correct shapes instead of the Mongolian orthographic (correct writing) rules and syntactic or semantic rules.

Misusing polyphone is contained in real-word error which refers to that the intended word is replaced by other correct spelling word with syntax or semantic error. Real-word error generally go unnoticed by most spellcheckers as they deal with words in isolation, accepting them as correct if they are found in the dictionary, and flagging them as errors if they are not [4]. So, as one kind of real-word error, polyphone proofreading is considered the more difficult task. Several approaches have been proposed in the literature. Su [5] adopted language model to correct the coding errors. Rule-based approaches to deal with the correction problem were discussed in [6], [7]. However, these approaches can only correct part of the coding errors and the real-word error did not be discussed systematically. Only in [6], human computer interaction approach is put forward which can be called half automatic operation implemented by manually selecting the proper one from the candidates generated from tools.

To solve this problem, we present a method for correcting polyphone mistakes using statistical language model based approach by building our own resource library. In this paper, we also introduce the MAPS (Mongolian automatic proofreading system), which applied the integration of rule-based approach and statistical language model approach as the polyphone proofreading module is the component of this system. Our intention is to improve the accuracy of polyphones and thus to improve the overall performance of the system. The system is built under the assumption that shapes in the text are correct, that is to say, the system doesn't correct the word form but their internal code assuming all the word form correct. Its implementation steps are as follow: Firstly, we use the approach of intermediate code [8] to unify the words with same presentation to one Latin letter lists. Secondly, according to dictionary and rule-based approach, the correct sets of the input tokens are obtained. Finally, polyphone correction module applies n-gram language model to select the proper spelling of polyphone, in which unigram, bigram and trigram model are conducted respectively. Experimental result shows that our proposed method performs very well.

The rest of the paper is organized as follows. Section 2 describes the Mongolian feature. Section 3 presents the whole system architecture. Our proposed method is described in Section 4. Experimental settings and results are discussed in section 5. Section 6 draws the conclusion.

2 Mongolian Feature

Mongolian as language of great influence over the world, its main users are distributed in China, Mongolia and Russia. Currently, despite other approaches such as code transformation [9] (from Founder code, Menk code, etc. to National standard

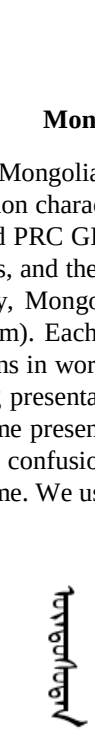
code), machine translation [2], speech recognition [3], etc., the Mongolian resource of electrical edition mainly comes from keyboard entry. The resource with serious typo errors is hardly utilized for the development of Mongolian information technology, which cause that, as one of the minority language in China, its informatization level is still lower than others like Tibetan and Uighur language. The following section presents detailed description of Mongolian character set and morphological ambiguity to illustrate the Mongolian unique feature. The intuitionistic interpretation of the reason why the typo errors are so seriously can be obtained in this section.

Table 1. Example of the same presentation forms with distinct codes.

No.	Presentation from	Position in word	Nominal form	Keyboard mapping	Code
1	ᠠ / ᠡ	medial, final	ᠠ	a	0x1820
			ᠡ	e	0x1821
			ᠨ	n	0x1828
2	ᠢ	final	ᠠ	a	0x1820
			ᠡ	e	0x1821
3	ᠤ / ᠥ	medial, final	ᠧ	q	0x1823
			ᠸ	v	0x1824
			ᠣ	o	0x1825
			ᠤ	u	0x1826
			ᠦ	w	0x1838
4	ᠶ / ᠧ	initial, medial	ᠧ	q	0x1823
			ᠸ	v	0x1824
5	ᠣ /ᠤ / ᠥ / ᠦ	initial, medial, final	ᠣ	o	0x1825
			ᠤ	u	0x1826
6	ᠷ / ᠽ	medial, final	ᠺ	i	0x1822
			ᠷ	y	0x1836
7	ᠲ / ᠳ	initial, medial	ᠲ	t	0x1832
			ᠳ	d	0x1833
8	ᠬ/ ᠭ/ ᠴ /ᠵ ᠨ/ ᠩ/ ᠬ	initial, medial	ᠬ	h	0x182c
			ᠭ	g	0x182d
9	ᠵ	medial	ᠵ	j	0x1835
			ᠶ	i	0x1832
			ᠷ	y	0x1836
10	ᠦ/ ᠤ	medial, final	ᠦ	w	0x1838
			ᠡ	E	0x1827

2.1 Mongolian Character Set

Mongolian characters contain two character types: nominal characters and presentation characters. According to Universal Coded Character Set (UCS) ISO/IEC 10646 and PRC GB 13000-2010, Mongolian character set only includes the nominal characters, and the units larger than one letter or less than one letter are not encoded. Generally, Mongolian letter set refers to the nominal characters (also known as nominal form). Each nominal character has several presentation forms according to its positions in words [10]. Table 1 shows Mongolian nominal characters and its corresponding presentation forms. Moreover, some characters have different nominal forms but same presentation forms. Mistakes are mainly committed by misusing those letters in the confusion set such as {a, e, n} (keyboard mapping) whose presentation forms are same. We use an example to illustrate this.



correct spelling: **undusuten**

undusuden(36187)	**undusuten(24708)**	undvsvden(7902)	undvsvdan(5141)
ondosoden(2403)	undusudan(1989)	undusutan(1895)	undqsqden(1828)
undvsvten(1181)	ondvsvdan(976)	untusuten(915)	undusudee(869)
ondqsqden(860)	undvsvdaa(840)	ondvsvden(788)	untusutee(723)
uedvsvden(706)	undqsqdan(661)	untvsvtan(658)	ondosoten(650)
undvsuden(622)	undusvden(510)	uedvsvdee(474)	undusudaa(450)
uadvsvdan(406)	uadqsqden(363)	undvsudan(281)	undvsvtan(259)
uedqsqdan(256)	ondqsqdan(245)	uadusudee(240)	uedvsvdan(235)
uedusuden(230)	uadusudan(217)	oedvsvdan(199)	oetvsvten(193)
uadqsqdan(189)	untvsvten(187)	undusvdan(177)	ondosodan(160)
undqsuden(158)	undqsvden(136)	ondvsoden(132)	ondvsvten(128)
ondosotan(128)	undvsqden(128)	uetqsqtea(123)	undusoden(118)
oetvsvtan(115)	undvsuten(114)	undqsqtan(113)	undqsqten(106)
uadvsvden(106)	uadvsvdaa(100)		

Fig. 1. Different spelling and frequency about the same Mongolian word "ᠳᠤᠤᠰᠤᠨ"

For the Mongolian word "ᠳᠤᠤᠰᠤᠨ" (meaning: minority), its keyboard mapping is "undusuten". According to the analysis on a Mongolian corpus including 76 million Mongolian words, this word appears 102532 times, and only 24708 times of its codes are correctly. The other 78124 ones are typed as other words with the same presentation forms. Actually, there are 291 words that have the same presentation forms as the word "ᠳᠤᠤᠰᠤᠨ" (meaning: minority). Fig. 1 shows the Mongolian word "ᠳᠤᠤᠰᠤᠨ" (meaning: minority) and its typos whose frequency is greater than 100 in the corpus.

2.2 Morphological Ambiguity

Morphological ambiguity is the possibility that a word is understood in multiple ways out of the context of their discourse. Words whose presentations look the same but spellings, pronunciations and meanings distinct according to the text called Polyphone. In Mongolian, polyphones are one of the most problematic objects in morpho-

logical analysis because they prevail all around frequent lexical items. Table 2 arranges polyphonic words with their corresponding pronunciations, meanings and part of speeches.

Table 2. Example of some Polyphone words.

No	Word from	Latin transliteration	Meaning	Font size and style
1	ᠬᠳᠬ	qdq	right now, stars	adverb, noun
2	ᠬᠲᠬ	qtq	omen	verb
3	ᠪᠲᠪ	vtv	smoke	verb
4	ᠪᠳᠪ	vdv	estrus	noun
5	ᠵᠢᠨ	jin	jin(loan word)	noun
6	ᠶᠢᠨ	-yin	's, of	noun

As for the upward 4 tokens, word form "ᠬᠳᠬ" has four kinds of pronunciation and obviously four kinds of coding. It can be represented by its corresponding Latin-transliteration (keyboard mapping) "qdq", "qtq", "vtv" and "vdv". The first word "qdq" is homonym which has multiple meanings of "now, right now" with part of speech adverb and "stars" with part of speech noun. The correct pronunciation or coding of them depends on the context of their discourse. The situation is same to the downward two tokens whose word form " ᠵᠢᠨ" maps to two distinct words of "jin" and "-yin". "jin" is the loanword whose pronunciation is about /dʒɪn/ , commonly used in person name and geographic name. Phrase "ᠵᠢᠨ ᠬᠣᠣᠨ" (meaning: state) means *jin dynasty*. "-yin", being a genitive suffix (meaning: 's, of), is concatenated to stem by Mongolian space (0x202f) which is 2/3 length of common space to form one word. The phrase "ᠪᠠᠰᠠ (meaning: teacher) ᠵᠢᠨ" means the teacher's. Although there is space between two tokens, "ᠪᠠᠰᠠ ᠵᠢᠨ" is one word comprised of stem "ᠪᠠᠰᠠ" and suffix"ᠵᠢᠨ". If "-yin" was mistakenly replaced by "jin", that is thought taking the misusing polyphone error by both changing its original meaning and even token quantity (from one to two).

The amount of polyphone in Mongolian corpus is comparatively larger than other languages which enjoyed the monophonic alphabet, and polyphone errors are badly serious in Mongolian corpus. Typists are always puzzled by the selection of the correct pronunciation of it or input the non-word which is out of correct spelling sets by just caring about correct shape instead of the correct coding for their laziness. Polyphone error detection and correction is one of the important tasks in Mongolian proofreading technology.

3 System Architecture

The polyphone proofreading module is one of the important tasks of the MAPS (Mongolian automatic proofreading system). It cannot run independently separated

from the whole system. In this section, we give whole framework of the system as illustrated in Fig. 2. Polyphone proofreading is framed by rough line.

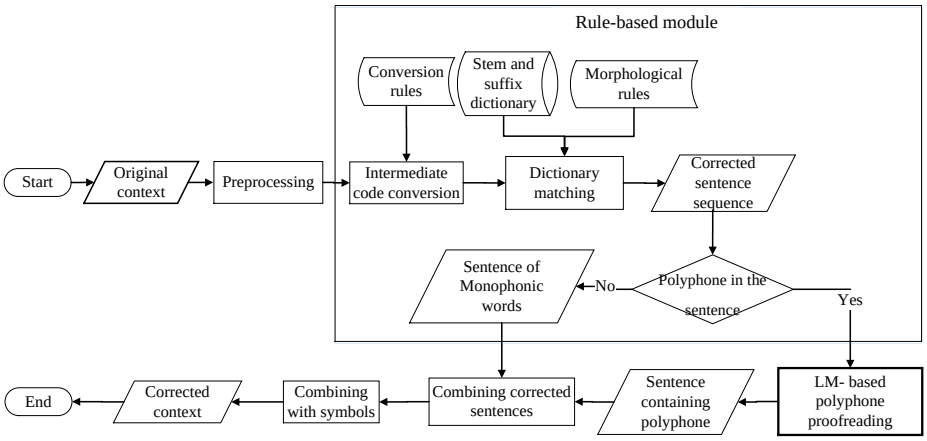


Fig. 2. System architecture

For considering that words with correct form but incorrect coding are far more than ones with incorrect form, the system dedicates to correct the words with correct form excluding those with formal error. The system takes the input text written by National Standard code in its original format, which undergoes preprocessing, rule-based module and LM-based process sequentially. In addition, it is worth noting that because of Mongolian agglutinative nature, the dictionary resource which is collated according to [11] applied in the system is comprised of the stem and suffix tokens instead of the whole word tokens.

Errors to be dealt with can be summed up as following three categories: (1) misspelled monophonic words, whose shapes and correct spellings are corresponded one by one, (2) misspelled plural case suffixes which are punctuated from the stem by Mongolian space and (3) misspelled polyphones, whose shape maps to multiple correct coding. The rule-based process, which is framed by dotted block, tackles the monophonic words and suffixes. The polyphones are processed in the LM-based component. The implementation steps are as follows:

Preprocessing: Sentence segmentation and special symbol processing are executed in this step.

Intermediate code transaction: This step is dedicated to convert each Mongolian input one by one into intermediate codes form utilizing the intermediate code transaction rules [8]. As the Fig. 3 is shown below, despite the variety of writings, the conversion approach makes one shape uniquely mapped to one intermediate code list.

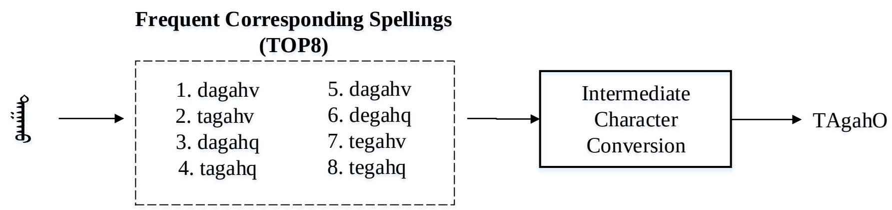


Fig. 3. Intermediate character conversion example

Dictionary matching: Taking intermediate codes received from the previous step as finding entry, correct spelling sets of words are acquired based on morphological rules and dictionaries. As shown in Fig. 4, firstly, the intermediate code is segmented to stem and suffix according to morphological rules [12]. Then, Latin-transliteration form of the correct stem sets and suffix sets are obtained respectively by matching from the dictionaries. Finally, suffixes are concatenated to the stems based on morphological rule. The output of this process is list of sentences, each of which is presented as chain of nodes (correct spelling sets of the word) as in the Fig. 5.

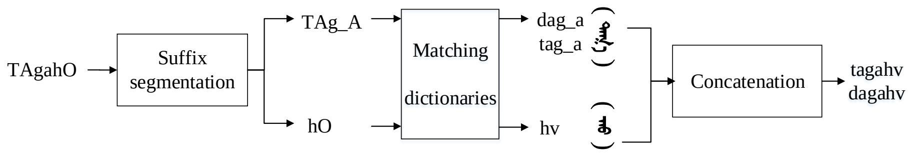


Fig. 4. Process of the matching from dictionary

Polyphone proofreading: Each sentence produced from the *Dictionary matching step* can be thought as node chains. Nothing will be done to an atom-chain, i.e. the quantity of each node of the chain is only one. In the other word, the sentence is composed of monophonic words. ML-based process is carried out on those sentences which contain polyphones.

4 Language Model Establishment

This study aims at improving the performance of the polyphone proofreading. With the observation that words a writer intends are semantically related to their surrounding words, the polyphone proofreading can be dealt by performing a word-level N-gram model analysis which specifies a priori probability of a particular word sequence. In this section, we introduce our statistical language model methodology for processing polyphone. Polyphone has high occurrence frequency in Mongolian documents. In statistic, more than 46,000 sentences contained the polyphonic words in the corpus of 50,000 sentences. Take the polyphone contained sentence " (minu) (bqdqgsan) (ni) (qdq) (ehileged) (bi) (harigvcahv) (bqlvn_a)" for example, in the sentence, the word "" ("bqdqgsan", "bvdvgsan") and "" ("vdv", "vtv", "qdq", "qtq") are polyphones. As illustrated in Fig. 5, Latin-

transliteration form was annotated below each Mongolian word. The word "" (meaning: think, paint) corresponds to two kinds of Latin form, the word "" (meaning: omen, smoke, now, estrus) corresponds to four correct spellings; The correct sentence is denoted by the path with the line in bolder, i.e. "minu bqdqgsan ni qdq ehileged bi harigvcahv bqlvn_a. "

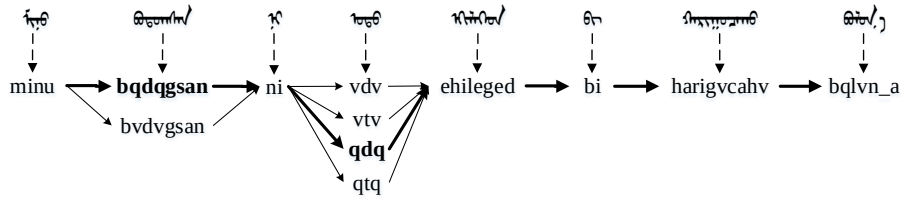


Fig. 5. Lexical chains of sentence

N-gram language model [13] has been widely used in statistical language model. The probability of a Mongolian word sequence $w = w_1 w_2 \dots \dots w_m$ can be written in the form of conditional probability:

$$p(w) = p(w_1 w_2 \dots \dots w_m) = \prod_{i=1}^{m} p\left(w_i \vee w_1^{i-1}\right) \approx \prod_{i=1}^{m} p\left(w_i \middle| w_{i-n+1}^{i-1}\right) \quad (1)$$

The probability of the m-th words w_m depends on all the words $w_1 w_2 \dots \dots w_{m-1}$. We can now use this model to estimate the probability of seeing sentences in the corpus by providing a simple independence assumption based on the Markov assumption [14]. Corresponding to the language model, the current word is only related to the previous n-1 words. From the equation (1), we can see that the target of language model is how to estimate the conditional probability of the next word in the list using $p\left(w_i \vee w_{i-n+1}^{i-1}\right)$. The most commonly probability estimation method we used is the maximum likelihood estimation (MLE).

$$p\left(w_i \middle| w_{i-n+1}^{i-1}\right) = \frac{c\left(w_{i-n+1}^{i}\right)}{c\left(w_{i-n+1}^{i-1}\right)} \quad (2)$$

$c\left(w_{i-n+1}^{i-1}\right)$ means the total count of the N-gram in the corpus. However, a drawback of the MLE is that the N-tuple corpus which does not appear in the training set will be given zero-Probability. Smoothing algorithm can be used to solve this kind of zero-Probabilities problem. In this paper, we use the Kneser-Ney smoothing algorithm [15].

5 Experiment

The principle contribution in this paper is twofold: (1) we built our own resource library including dictionaries containing all polyphones, and dataset used in training corpus and test corpus; (2) We conduct the language model based method to deal with polyphone errors. In this section, we describe how the resource is created and show the experimental evaluation and analysis.

5.1 Data Resource

In general, there is a limitation in the number of Mongolian linguistic resources that are publicly available free for the research purpose. Therefore, we have to spend tangible efforts to acquire/annotate and verify our own linguistic resources in order to properly develop the proofreading system.

The proposed statistical approach rely on pre-defined confusion sets, which are comprised of commonly confounded words , such as polyphone sets of {"qdq", "qtq", "vtv", "vdv"} illustrated in Table 2 and the good-quality dataset used as training and testing dataset. After a period of collecting and collating, finally we finished creating the confusion sets by 252 verbal stems all put into the verbal stem dictionary and 998 whole words injected in nominal stem dictionary. Concatenated by verbal suffixes and case suffixes, the verbal stems can derive about 22,971 tokens and 998 whole words can derive about 19,407 tokens when concatenated by case suffixes. Since the textual resource in the Internet is full of coding errors, dataset used for creating training set and test data is constructed by following three steps: (1) Original Mongolian texts of about 50,000 sentences written in national standard code are obtained from the Mongolian news web. (2) The texts are corrected preliminarily by automatic proofreading system without polyphone correction module. For the polyphone, randomly select one candidate. Then, sentences which contain polyphone are picked out. (3) The manual annotation task carries out on those selected sentences under the open source platform BRAT [16]. The annotation takes about one and a half months with four Mongolian native persons. The collated Mongolian corpus, each of which contained the polyphones, consists of 41,416 sentences and 2,822,337 words. That was split into training data of 38,416 sentences and test data of 3,000 sentences.

5.2 N-gram Language Model Based Approach

We take the Correction Accurate Rate (CAR) as the evaluate metric, which is defined as

$$CAR = \frac{N_{correct}}{N_{total}} \tag{3}$$

$N_{correct}$ denotes the number of all polyphone that are correctly proofread. And N_{total} is the total number of all the polyphone needed to be corrected. We conduct the n-gram language model by SRILM toolkit [15] with Kneser-Ney discounting.

The calibration progress can be divided into two steps: Firstly, correct all Mongolian words one by one according to the rule based approach; Then, we check whether polyphone is contained or not in those sentences. If polyphone is contained, taking sentence as the basic unit, we further determine the best one according to the Language Model. To improve the performance of CAR , we respectively conduct unigram, bigram and trigram model to evaluate the experiment. As the result shown in Fig. 6, trigram model performs best by accuracy rate 95.36%, which is 62% higher than that of polyphones in original text without correction. Both bigram and trigram model outperformed the unigram model. The result shows that polyphone proofreading performance is effectively improved when contextual information is utilized in the process. Because of data sparseness, performance of trigram model did not show sig-

nificant improvement with slight promotion of 0.06% compared to bigram model. Experiment will lead to better results if the experimental dataset become more adequate.

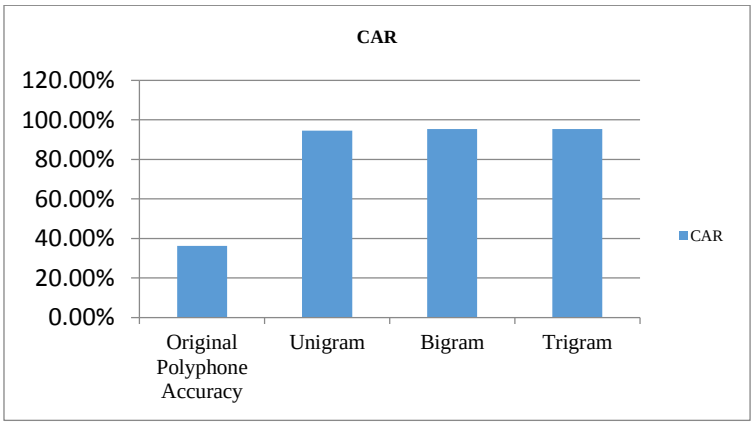


Fig. 6. Performance comparison between the Rule-based and LM based approaches

We also test the overall performance as the result illustrated in the Fig. 7. We can see that the overall system performance, when applied to the trigram model in polyphone proofreading, has the improvement by 16.1%.

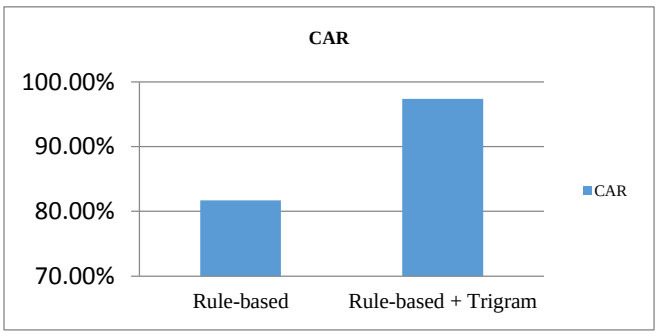


Fig. 7. Overall system performance comparison

6 Conclusion

In this paper, we present the statistical language model based approach after the description of the MAPS framework, and introduce in detail the construction of the resource library. Our purpose is the development of a high-quality correction module for polyphonic words which is one of the real-word correction problems. From the experiment result, N-gram language model was proved to be an effective approach to polyphone correction with the overall performance of the automatic proofreading

system improved by 16.1%. In future work, we plan to expand our training sets and try to use other methods to detect and correct polyphones. Moreover, we will extend our method to allow for other kinds of real-word errors such as semantic errors, malapropisms structural errors and pragmatic errors.

Acknowledgements. This paper is supported by The National Natural Science Foundation of China (No.61563040), Inner Mongolia Natural Science Foundation of major projects (No.2016ZD06) and Inner Mongolia Natural Science Fund Project (No.2017BS0601).

References

1. Wang, W., Bao, F., Gao, G.: Mongolian Named Entity Recognition System with Rich Features. COLING, 2016. 505-512 (2016).
2. Bao, F., Gao, G., Wang, H., et al.: Cyril Mongolian to traditional Mongolian conversion based on rules and statistics method. Journal of Chinese Information Processing 31(3), 156-162 (2013).
3. Bao, F., Gao, G., Yan, X., et al.: Segmentation-based Mongolian LVCSR approach. Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013. 8136-8139 (2013).
4. Islam, A., Inkpen, D.: Real-Word Spelling Correction using Google Web 1T n-gram Data Set. International Conference on Natural Language Processing and Knowledge Engineering, 2009. Nlp-Ke. IEEE, 1689-1692(2009).
5. Su, C., Hou, H., Yang, P., Yuan, H.: Based on the statistical translation framework of the Mongolian automatic spelling correction method. J. Chin. Inf. Proces. 175–179 (2013).
6. Si, L.: Mongolian proofreading algorithm based on nondeterministic finite automata. Chinese Journal of information 23 (6), 110-115 (2009).
7. Jiang, B.: Research on Rule-Based the Method of Mongolian Automatic Correction. Inner Mongolia University, Hohhot (2014).
8. Yan, X., Bao, F., Wei, H., et al.: A Novel Approach to Improve the Mongolian Language Model Using Intermediate Characters. China National Conference on Chinese Computational Linguistics. Springer International Publishing, 103-113 (2016).
9. Gong, Z.: Research on Mongolian code conversion. Inner Mongolia University (2008).
10. GB 25914-2010: Information technology of traditional Mongolian nominal characters, presentation characters and control characters using the rules (2011).
11. Surgereltu.: Mongolia orthography dictionary. 5th edn. Inner Mongolia People's Publisher , Hohhot (2011).
12. Inner Mongolia University.: Modern Mongolian. 2nd edn. Inner Mongolia People's Publisher, Hohhot (2005).
13. Zong, C.: Statistical Natural Language Processing. 2nd edn. Tsinghua University Press, Beijing (2008).
14. Jurafsky, D., Martin, J.: Speech and Language Processing. 2nd edn. Prentice Hall, Upper Saddle River (2009).
15. Stolcke, A.: SRILM - An Extensible Language Modeling Toolkit. In: Proc. Intl. Conf. Spoken Language Processing, Denver, Colorado (2002).
16. Pontus, S., Sampo, P., Goran T.: Brat: a web-based tool for nlp-assisted text annotation. In Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, 102–107.