Mongolian Prosodic Phrase Prediction using Suffix Segmentation

Rui Liu, Feilong Bao, Guanglai Gao, Weihua Wang
College of Computer Science, Inner Mongolia University
Hohhot, China, 010021
liurui_imu@163.com; {csfeilong, csggl}@imu.edu.cn; wangweihuacs@163.com

***Abstract*—Accurate prosodic phrase prediction can improve the naturalness of speech synthesis. Predicting the prosodic phrase can be regarded as a sequence labeling problem and the Conditional Random Field (CRF) is typically used to solve it. Mongolian is an agglutinative language, in which massive words can be formed by concatenating these stems and suffixes. This character makes it difficult to build a Mongolian prosodic phrase predictions system, based on CRF, that has high performance. We introduce a new method that segments Mongolian word into stem and suffix as individual token. The proposed method integrates multiple features according to the characteristics of Mongolian word formation. We conduct the contrast experiment by selecting the following features: word, multi-level Part-of-Speech (POS), multi-level lexical for suffix and the existence for suffix. The experimental results show that our method has significantly enhanced the performance of the Mongolian prosodic phrase prediction system through comparing with the conventional method that treats Mongolian word as token directly. The word feature, level one lexical for suffix feature and existence for suffix feature are effective. The best result is measured by F1-measure as 82.49%.**

Keywords- Mongolian, speech synthesis, word formation, Conditional Random Field, prosodic phrase prediction.

I. INTRODUCTION

The speech synthesis system, which can transform input text into speech, is an artificial intelligence system. The problem of how to automatically generate high-quality speech is attracting more and more attention from researchers. For the study of speech synthesis, the automatic prosodic prediction is an important part of improving the naturalness of the speech synthesis. The prediction is intended to estimate the right pause position of speech. Mongolian prosodic structure includes the following four units: syllable, prosodic word, prosodic phrase and intonational phrase. Among them, the prosodic phrase is an important prosodic unit in the process of speech synthesis. Thus, this paper is aimed at the prediction of prosodic phrase boundary.

Most early researches of the prosodic prediction focus on rules primarily. But the rules have limited ability to cover all of the language phenomena. This is mainly the disadvantage of the Rule-Based methods which is difficult to overcome. The study about the Mongolian prosodic phrase prediction is still in its primary stage and many issues are needed to be solved. In recent years, more and more researchers is using statistical learning for the prosodic boundary prediction [1,2,3,4], while the CRF is a common choice.

As far as we know, there are few studies reported on the topic of Mongolian prosodic phrase prediction. While for the Mongolian speech synthesis: Ochir et al. proposed a Mongolian speech synthesis system based on waveform concatenation [5]; Monghjaya conducted a research on the Mongolian speech synthesis based on stem and affixes [6]; Aomin carried out a study on Mongolian speech synthesis based on the prosodic [7]; Zhao used the HMM-Based methods in the Mongolian speech synthesis [8]. These studies have made contribution to the Mongolian speech synthesis, but the naturalness of Mongolian speech synthesis is less than satisfactory.

In this study, we build a CRF-based prosodic phrase prediction system for Mongolian. However, it is difficult to train a CRF model for the Mongolian prosodic phrase prediction directly. The main reason lies on the data sparseness, which is caused by the Mongolian's large vocabulary. Thus, we propose two new methods to segment the ending suffix followed by the Narrow No-Break Space (NNBSP) in Mongolian nouns according to the characteristics of Mongolian word formation. This process improves the performance of the Mongolian prosodic phrase prediction system and its results are better than the systems without this process. This method also can facilitate other agglutinative languages like Turkish and Korean and so on.

The remaining of the paper is organized as follows: Section 2 describes the characteristics of Mongolian word formation in details; Section 3 illustrates the two new processing methods and introduces all features in the CRF-based Mongolian prosodic phrase prediction system; experimental setups and results are shown in Section 4; finally, the conclusions are summarized in Section 5.

II. CHARACTERISTICS OF THE MONGOLIAN WORD FORMATION

Mongolian is an agglutinative language. A Mongolian word can be decomposed into a root and several suffixes. Mongolian does not have an infix and prefix, but a suffix. A suffix can be categorized as word-formation suffix, inflectional suffix and ending suffix.

Generally, ending suffixes affect the form only, such as the tense, person and so on. Every root can be followed by many word-formation suffixes and inflectional suffixes to generate the stem. Most of the generated stem can be further followed by an ending suffix and a little of them are followed by two ending suffixes [9]. Moreover, the pronunciation of suffix will be different depending on the stem they follow. Fig. 1 illustrates the relationship among the root, stem and suffix of a Mongolian word. For example, the verb-stem "  " (Latin-transliteration: yabv, means: go) can be followed by more than 40 ending suffixes to generate different words. Another example, noun-stem "  " (Latin-transliteration : sandal, means:

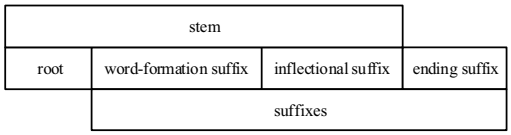

Figure 1. The relationship of root, stem and suffix of a word in mogolian

TABLE I. THE EXAMPLE OF NEW WORDS GENERATED BY MONGOLIAN STEM ADD ENDING SUFFIX.

Stem	ending suffix	Type	generated words
ᠶᠠᠪᠤ (yabv)	ᠬᠤ (hv)	present tense	ᠶᠠᠪᠬᠤ (yabvhu, means: go)
	ᠪᠠᠢ (jai)	perfect tense	ᠶᠠᠪᠪᠠᠢ (yabvjai, means: have/had gone)
	ᠯᠠ (l_a)	past and future tense	ᠶᠠᠪᠯᠠ (yabvl_a means: went/will go)
ᠰᠠᠨᠳᠠᠯᠢ (sandali)	ᠳᠤ (-dv)	dative-locative	ᠰᠠᠨᠳᠠᠯᠢ ᠳᠤ (sandali-dv means: to the chair)
	ᠶᠢ (-yi)	accusative	ᠰᠠᠨᠳᠠᠯᠢ ᠶᠢ (sandali-yi means: the chair)

chair) can be followed by 16 different ending suffixes and generate 16 different words. Table I partially lists the possible generated words for the two stems above.

We establish the stem and suffix library according to our statistics from Mongolian dictionary. The library has 60,294 stems and 214 ending suffixes that are frequently used in Mongolian. Massive words can be formed by concatenating these stems and ending suffixes. A main issue of Mongolian research is the data sparseness problem, even in a large labeled corpus, because it is difficult to include all Mongolian words. The problem has brought a serious difficulty and a big challenge for the Mongolian prosodic phrase prediction.

III. APPROACH

A. Segmentation process

For many Mongolian nouns: NNBSP, a special character, is used to concatenate root and suffixes. The NNBSP characters are written as "-" in Latin representations. The ending suffix which is concatenated with stem by NNBSP is called NNBSP-suffix by us. In this study, we segment the Mongolian nouns into stems and NNBSP-suffixes.

It is possible to segment all of the suffixes from the stem, but much harder work is needed. Because there may be some letters insertion, dropping or substitution [9] occurring when a suffix is being segmented from the stem. The NNBSP-suffix is the most common suffix that can be segmented correctly. In addition, this kind of ending suffixes comes after the verb-stem directly and we call it VSES, and it is unable to segment accurately. In order to ensure the accuracy of segmentation, we only segment the NNBSP-suffix in this study.

After the segmentation process, we propose two methods to alleviate the large-vocabulary problem.

- Treat the stem and the NNBSP-suffixes as individual tokens. (SEG)
- Remove every NNBSP-suffix, and only treat the stems as tokens. (SEG_RM)

TABLE II. THE EXAMPLE OF TWO SEGMENTATION PROCESSING METHODS

Mongolian sentence	Latin	SEG	SEG_RM
ᠴᠢ	qi	qi	qi
ᠬᠥᠬᠡᠳ᠎ᠡᠶᠢ	heuhed-i	heuhed	heuhed
		-i	
ᠰᠢᠶᠳᠡᠩ	siidhen	siidhen	siidhen
ᠲᠤᠭᠷᠭᠠᠰᠠᠨ᠎ᠤ	tvrgagsan-v	tvrgagsan	tvrgagsan
		-v	
ᠬᠤᠢᠨ᠂	hwin_a	hwin_a	hwin_a
ᠬᠥᠬᠡᠳ ᠲᠡᠭᠡᠨ	heuhed-tegen	heuhed	heuhed
		-tegen	
ᠵᠤᠷᠢᠯᠭ᠂ ᠪᠠᠨ	jwrilg_a -ban	jwrilg_a	jwrilg_a
		-ban	
ᠮᠡᠳᠡᠭᠡᠯᠦᠭᠰᠡᠨ	medegulugsen	medegulugsen	medegulugsen
ᠪᠠᠢᠬᠤ	baihv	baihv	baihv
ᠬᠡᠷᠡᠭᠲᠡᠢ	heregtei	heregtei	heregtei

We compare the different approaches for NNBSP-suffix with the method that treats the word as token directly to complete Mongolian prosodic phrase prediction task. In Table II, we give an example to illustrate these methods, in which the Mongolian sentence: "ᠴᠢ ᠬᠥᠬᠡᠳ᠎ᠡᠶᠢ ᠰᠢᠶᠳᠡᠩ ᠲᠤᠭᠷᠭᠠᠰᠠᠨ᠎ᠤ ᠪᠤᠶᠤ᠂ ᠬᠤᠢᠨ᠂ ᠬᠥᠬᠡᠳ ᠲᠡᠭᠡᠨ᠂ ᠵᠤᠷᠢᠯᠭ᠂ ᠪᠠᠨ ᠮᠡᠳᠡᠭᠡᠯᠦᠭᠰᠡᠨ ᠪᠠᠢᠬᠤ ᠬᠡᠷᠡᠭᠲᠡᠢ" (means: you should tell your children the reason after you teach them a lesson.) has been segmented into stem and NNBSP-suffixes. The SEG method takes the stem and NNBSP-suffixes as new tokens, while the SEG_RM method removes the NNBSP-suffixes and leaves the stem only. Our experiment is carried out based on CRF++ toolkit[1].

B. Features

According to the characteristics of the Mongolian word formation, nine features are used in this paper. Here, these features are shown as follows:

Word (*w*): the Mongolian word itself.

Level one POS (*p1*): level one POS tag of the word.

Level three POS (*p3*): level three POS tag of the word.

Level one lexical tag for VSES (*v1*): If a Mongolian word has a VESE, we mark the level one lexical tag for VSES.

Level three lexical tag for VSES (*v3*): If a Mongolian word has a VESE, we label the level three lexical tag for VSES.

Level one lexical tag for NNBSP-suffix (*n1*): If a Mongolian word has a NNBSP-suffix, we label the level one lexical tag for NNBSP-suffix.

Level three lexical tag for NNBSP-suffix (*n3*): If a Mongolian word has a NNBSP-suffix, we label the level three lexical tag for NNBSP-suffix.

NNBSP-suffix (*s*): NNBSP-suffix itself.

Existence of NNBSP-suffix (*t*): The tag is showing whether the NNBSP-suffix exist or not in a Mongolian word.

We label the level one and level three tag for word according to the national standard (GB/T 26235-2010) [11], which contains 15 level one POS tags and 72 level three POS tags. For NNBSP-suffix, they have 1 level one lexical tag and 22 level three lexical tags; VSES lexical tag includes 1 level one lexical tag and 58 level three lexical tags.

[1] http://taku910.github.io/crfpp/

C. Establishment of the Labeling Data

We select the 11,345 voice data and their corresponding text from the Mongolian speech synthesis Database in Inner Mongolia University's College of Computer Science. The Database is recorded by one standard Mongolian announcer. The text data set includes 449,000 Mongolian words, the number of the words which from the data is processed by deleting duplicate words is about 22,050. According to the voice data, we label the prosodic phrase boundary for the corresponding text [12]. The idea of word segmentation as tagging assigns a tag to each word, namely B (Begin and Middle), E (End and Single) and S (Pause), according to its position in the prosodic phrase in a certain context.

We label the POS or lexical tag by the Mongolian lexical analysis tool [13] designed by Inner Mongolia University's School of Mongolia Studies.

In our experiments, we adopt the Latin-transliteration representation for the labeled data and conducted three group experiments under the different processing method. The reference group that uses the data without processing is called Group-0, the experimental group that uses the data processed by SEG method is called Group-1, the experimental group that uses the data processed by SEG_RM method is called Group-2. For each group, we keep the data set in random order five times to generate five data sets, and use the five data sets to evaluate the experimental results with 90% for training and 10% for testing. The Word and Out-of-vocabulary (OOV) number of three groups are shown in Table III and Table IV. For the SEG and SEG_RM methods, they segment and remove the NNBSP-suffixes respectively, which does not change the stem number. Hence the stem number of the Group-1 and Group-2 are identical. We can see from the statistical results. The average percentage of OOV in Group-0 is 16.68 on average and in Group-1 and Group-2 is 12.94. It demonstrates that segmenting NNBSP-suffix helps to decrease the amount of OOV.

D. Evaluation standard

The commonly evaluation metrics, prosodic phrase boundary precision (P), Recall (R) and F1-measure (F), are used as performance evaluation for prosodic phrase boundary prediction. As the punctuation which means suspend or pause can be regarded as boundary characteristics obviously, we use two F1-measure values as the evaluation standard.

1. **F1-a**: the F1-measure includes prosodic boundary expressed by punctuation which means suspend or pause.
2. **F1-b**: F1-measure fails to include prosodic boundary expressed by the punctuation which means suspend or pause.

E. Experiment

Experiment 1: Determination of best Window Size

We analyze the impacts of each single feature and template under the CRF framework. The Group-0 contains *w, p1, p3, v1, v3, n1, n3, t* features, the Group-1 has *w, p1, p3, v1, v3, t* features and Group-2 contains *w, p1, p3, v1, v3, s, t* features. Because the SEG method segment NNBSP-suffixes as new tokens, so the features of Group-1 take no account of *n1, n3, s*. For Group-2, we remove all NNBSP-suffixes, so its features take no account of *s*. For Group-0, we don't choose *s* because this method does not do anything about NNBSP-suffix alone. Firstly, we adjust the window size for each feature in the range 1 to 4 for the three groups experiments. By comparing each feature's F1-measure under every window size, we get the best window size for each feature which is illustrated in Table V. In the experiment we find that the effect of *p3, v3, n3* on the prosodic phrase prediction is better than that of *p1, v1, n1*. So we continue the next experiments by combining these six features: *w, p3, v3, n3, s* and *t*.

Experiment 2: Determination of best Feature Combinations.

We employ different feature combinations for our experiment, which includes *w, p3, v3, n3, s* and *t*. The results are defined as shown in Table VI.

Baseline: The reference group uses the data without any NNBSP-suffix process. The window size of *w* feature was fixed at 3. F1-a achieved 81.35 and F1-b achieved 60.12.

As illustrated in Table Ⅵ

1. The F1-a or F1-b of Group-1 shows higher than the other two groups obviously.

2. The best F1-measure has appeared when the feature combination form for Group-0 is "*w+v3+t+n3*" and for Group-2 is "*w+v3+t+s*". It is also proves that the performance of the system can be promoted by regarding NNBSP-suffix itself as a feature. The *v3, n3* features had effects on the prosodic phrase prediction in some degree.

3. The *p3, t* features had no obvious promotional effects on the prosodic phrase prediction. The most useful feature is *s*. It can increase F1-measure by 1.13 on average.

4. When we use the features that combine *w, v3* and *t*, the F1-a and F1-b reached their highest values respectively

TABLE III. THE DETAILS OF WORD AND OOV IN GROUP-0.

Group-0	Test1	Test2	Test3	Test4	Test5
Word number	6,604	6,543	6,595	6,770	6,591
OOV number	1,089	1,082	1,089	1,154	1,116
OOV percentage	16.5	16.5	16.5	17.0	16.9

TABLE IV. THE DETAILS OF WORD AND OOV IN GROUP-1/2.

Group-1/2	Test1	Test2	Test3	Test4	Test5
Word number	4,715	4,650	4,636	4,813	4,688
OOV number	603	588	600	639	616
OOV percentage	12.8	12.6	12.9	13.3	13.1

TABLE V. THE BEST WINDOW SIZE OF EACH FEATURE.

Data	*w*	*p1*	*p3*	*v1*	*v3*	*n1*	*n3*	*s*	*t*
Group-0	3	4	4	1	4	1	1	-	1
Group-1	3	4	3	1	4	-	-	-	1
Group-2	3	4	4	1	4	-	-	1	1

A short line means the Data don't have this feature.

TABLE VI. THE RESULTS OF THE COMBINATION FEATURES EXPERIMENT.

ID	Features	Group-0		Group-1		Group-2	
		F1-a	*F1-b*	*F1-a*	*F1-b*	*F1-a*	*F1-b*
1	*w*	**81.35**	**60.12**	82.36	63.00	80.43	59.19
2	*w+n3*	81.74	62.37	-	-	-	-
3	*w+s*	-	-	-	-	82.04	63.08
4	*w+t*	80.93	60.29	82.08	62.92	80.53	59.63
5	*w+t+n3*	81.77	62.45	-	-	-	-
6	*w+t+s*	-	-	-	-	82.04	63.09
7	*w+p3*	80.87	60.15	82.03	62.96	80.41	59.34
8	*w+p3+n3*	81.72	62.48	-	-	-	-
9	*w+p3+s*	-	-	-	-	81.74	62.51
10	*w+p3+n3+t*	81.69	62.37	-	-	-	-
11	*w+p3+s+t*	-	-	-	-	81.79	62.64
12	*w+p3+t*	80.99	60.57	82.13	63.17	80.56	59.76
13	*w+p3+v3*	80.78	60.07	82.22	63.44	80.26	59.10
14	*w+p3+v3+n3*	81.86	62.87	-	-	-	-
15	*w+p3+v3+s*	-	-	-	-	81.88	62.90
16	*w+p3+v3+t*	81.05	60.88	82.26	63.56	80.56	59.87
17	*w+p3+v3+t+n3*	81.86	62.85	-	-	-	-
18	*w+p3+v3+t+s*	-	-	-	-	81.78	62.68
19	*w+v3*	81.00	60.54	82.47	63.90	80.53	59.63
20	*w+v3+n3*	81.97	63.08	-	-	-	-
21	*w+v3+s*	-	-	-	-	82.03	63.26
22	*w+v3+t*	81.11	60.95	**82.49**	**63.97**	80.63	59.99
23	*w+v3+t+n3*	81.99	63.13	-	-	-	-
24	*w+v3+t+s*	-	-	-	-	**82.12**	**63.43**

in Group-1 experiment: 82.49 and 63.97, and higher than other feature combinations.

IV. CONCLUSIONS

In this paper, we explore two NNBSP-suffix process methods based on CRF for the Mongolian prosodic phrase prediction. The results of the experiment show that the processing of NNBSP-suffix can improve the performance of the Mongolian speech synthesis system. We find that the performance of the Mongolian prosodic prediction by the SEG method is better than the SEG_RM method and the above methods are better than the method without processing Mongolian word. It indicates that it is helpful to segment the Mongolian word into smaller token. In addition, in the process of feature selection in CRF training, we select nine features: word, level one POS, level three POS, level one lexical tag for VSES, level three lexical tag for VSES, level one lexical tag for NNBSP-suffix, level three lexical tag for NNBSP-suffix, NNBSP-suffix, existence of NNBSP-suffix. Word and NNBSP-suffix are the most effective features among the above features. Finally, by combining the three features: Word, Level three lexical tag for VSES and Existence of NNBSP-suffix, we obtain the best result (F1-a: 82.49, F1-b: 63.49).

Selecting the above features and segmenting NNBSP-suffix can also facilitate other agglutinative languages like Turkish and Korean and so on.

In future we will improve the accuracy of the POS tagging and explore to add new feature to the feature set, such as word vector. We would consider to use the Neural Network to continue the Mongolian prosodic phrase prediction task as well.

REFERENCES

[1] Y. Wang, " Key technologies for Text-to-Speech systems". Tsinghua University, 2013.

[2] A. Kaufmann, "Negation and prosody in British English: a study based on the London–Lund corpus. Journal of Pragmatics", 2002, vol. 34, pp. 10–11, 1473-1494.

[3] C. Ding, L. Xie, J. Yan and W. Zhang, "Automatic prosody prediction for Chinese speech synthesis using BLSTM-RNN and embedding features". In: Proceedings of ASRU , 2015.

[4] V. R. Reddy, S. Rao. K, "Prosody modeling for syllable based text-to-speech synthesis using feedforward neural networks. Neurocomputing", 2015, vol. 171, pp. 1323-1334.

[5] Ochir and Z. Gong, "A Test of The Speech Synthesis with the Waveform Concatenation". In: 3th NCMMSC, Chongqing, 1994, pp. 408-412 .

[6] Monghjaya, "A Research on Mongolian Speech Synthesis-system Based on Stems and Affixes". Journal of Inner Mongolia University, 2008, vol. 39, no. 6, pp. 693-697.

[7] Aomin, "Research on Mongolian speech synthesis based on prosody". Hohhot: Inner Mongolia University, 2012.

[8] J. Zhao, G. Gao and F. Bao, "Research on HMM-based Mongolian Speech Synthesis". Computer Science, 2014, vol. 41, no. 1, pp. 80-104 .

[9] F. Bao, G. Gao, X. Yan and W. Wang, "Segmentation-based Mongolian LVCSR approach". In: Proceeding of ICASSP, 2013, pp. 8136-8139.

[10] J. Lafferty, A. Mccallum and F. C. Pereira, "Conditional Random Fields: Probabilistic Models For Segmenting And Labeling Sequence Data", 2001, pp. 282-289.

[11] GB/T 26235-2010, The People's Republic of China national standards: Information technology-Mongolian word and expression marks for information processing.

[12] J. Zhao, G. Gao and F. Bao, "Designing a rule for annotation of corpus data in synthesis of Mongolian speech". Journal of Inner Mongolia University, 2013, vol. 44, no. 3, pp. 324-328 .

[13] S. Loglo, Sarula and S. Hua, "Research on Mongolian lexical analyzer based on NFA. In: IEEE International Conference on Intelligent Computing and Intelligent Systems", 2010, pp. 240 – 245.