Part of Speech Tagging for Mongolian Corpus

Purev Jaimai and **Odbayar Chimeddorj**
Center for Research on Language Processing
National University of Mongolia
{purev, odbayar}@num.edu.mn

Abstract

This paper introduces the current result of a research work which aims to build a 5 million tagged word corpus for Mongolian. Currently, around 1 million words have been automatically tagged by developing a POS tagset and a bigram POS tagger.

1 Introduction

In the information era, language technologies and language processing have become a crucial issue to our social development which should benefit from the information technology. However, there are many communities whose languages have been less studied and developed for such need.

Mongolian is one of the Altaic family languages. It has a great, long history. Nonetheless, till now, there are no corpora for the Mongolian language processing (Purev, 2008). Two years ago, a research project to build a tagged corpus for Mongolian began at the Center for Research on Language Processing, National University of Mongolia. In the last year of this project, we developed a POS tagset and a POS tagger, and tagged around 1 million words by using them.

Currently, we have manually checked 260 thousand automatically tagged words. The rest of the tagged words have not checked yet because checking the tagged corpus needs more time and effort without any automatic or semi-automatic tool and method.

The statistical method is used in our POS tagger. The rule based method requires the Mongolian language description which is appropriate to NLP techniques such as POS tagger. But, the current description of Mongolian is quite difficult to model for the computer processing. The tagger is based on a bigram method using HMM. The tagger is trained on around 250 thousand manually tagged words, and its accuracy is around 81 percent on tagging around 1 million words.

2 POS Tagset Design

We designed a POS tagset for Mongolian corpus by studying the main materials in Mongolia (PANLocalization, 2007). According to the agglutinative characteristics of Mongolian, the number of tags is not fixed, and it is possible to be created a lot of combinations of tags.

The POS tagset consists of two parts that are a high-level tagset and a low-level tagset. The high-level tagset is similar to English tags such as noun, verb, adword etc. It consists of 29 tags (see Table 1), while the low-level tagset consists of 22 sub tags (see Table 2). The annotation of our tagset mainly follows the tagsets of PennTreebank (Beatrice, 1990) and BNC (Geoffrey, 2000).

No.	Description	Tag
	Noun	
1.	Noun	N
2.	Pronoun	PN
3.	Proper noun	RN
4.	Adjective	JJ
5.	Pro-adjective	PJ
6.	Ad-adjective	JJA
7.	Superlative	JJS
8.	Number	CD
9.	Preposition	PR
10.	Postposition	PT
11.	Abbreviation	ABR
12.	Determiner	DT
13.	Morph for possessive	POS
	Verb	
14.	Verb	V

Proceedings of the 7th Workshop on Asian Language Resources, ACL-IJCNLP 2009, pages 103–106,
Suntec, Singapore, 6-7 August 2009.

15.	Proverb	PV
16.	Adverb	RB
17.	Ya pro-word	PY
18.	Ad-adverb	RBA
19.	Modal	MD
20.	Auxiliary	AUX
21.	Clausal adverb	SRB
22.	Ge-rooted verb	GV
23.	Co-conjunction	CC
24.	Sub-conjunction	CS
Others		
25.	Interjection	INTJ
26.	Question	QN
27.	Punctuation	PUN
28.	Foreign word	FW
29.	Negative	NEG

Table 1. High-Level Tagset for Mongolian

No.	**Description**	**Tag**
Noun		
1.	Genitive	G
2.	Locative	L
3.	Accusative	C
4.	Ablative	B
5.	Instrumental	I
6.	Commutative	M
7.	Plural	P
8.	Possessive	S
9.	Approximate	A
10.	Abbreviated possessive	H
11.	Direction	D
Verb		
12.	Past	D
13.	Present	P
14.	Serial verb	S
15.	Future	F
16.	Infinitive/Base	B
17.	Coordination	C
18.	Subordination	S
19.	1st person	1
20.	2nd person	2
21.	3rd person	3
22.	Negative	X

Table 2. Low-Level Tagset for Mongolian

The high-level tags are classified into noun, verb and others as shown in Table 1. In the noun column, parts of speech in the noun phrase such as adjective, number, abbreviation and so on are included. In the verb column, parts of speech in the verb phrase are included. In the other column, the parts of speech except those of the noun and verb phrases are included.

The low-level tagset is divided into two general types: noun phrase and verb phrase. It also consists of sub tags for inflectional suffixes such as cases, verb tenses etc. These tags are used mainly in combination with high-level tags.

Currently, around 198 combination tags have been created. Most of them are for noun and verb inflections. Tag marking length is 1 – 5 letters. Below we show some tagged sentences (see Figure 1).

Би	морь	ундаг
PN	**N**	**VP**
I	horse	ride
I ride a horse		

Би	мориноос	айдаг
PN	**NB**	**VP**
I	from horse	fear
I fear horses		

Би	мориноосоо	буулаа
PN	**NBS**	**VD**
I	from my horse	got off
I got off my horse		

Figure 1. Mongolian Tagged Sentences

Three example sentences are shown in Figure 1. Mongolian sentence is placed in the first line, and the following lines, second, third and fourth are POS tags, English parts of speech translation and English translation, respectively. A word '*морь*' (horse) is used with different morphological forms such as nominative case in the first sentence, ablative case in the second sentence and ablative case followed by possessive in the last sentence. The noun inflected with nominative case is tagged N, the noun inflected with ablative case is tagged NB, and the noun inflected with ablative case and possessive is tagged NBS according to the two level tagset.

3 Bigram-POS Tagger

Although the statistical method needs a tagged corpus which takes a lot of time and effort, it is more reliable for languages whose linguistic descriptions have difficulties in NLP and CL purposes. Thus, we are developing a statistical POS tagger for the project.

The statistical method has been used on POS taggers since 1960s (Christopher, 2000). Some of these kinds of methods use HMM (Hidden Markov Model). The main principle of HMM is to assign the most possible tag to an input word in a sentence by using the probabilities of training data (Brian, 2007 and Daniel, 2000).

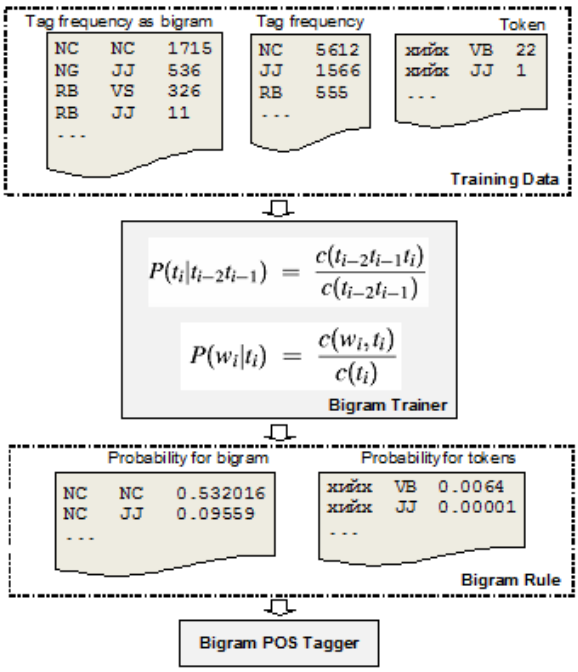

Figure 2. Overview of Mongolian Bigram tagger

The probabilities for the bigram tagger are calculated with the uni-tag frequency, the bi-tag frequency and the tokens from the training data (see Figure 2 for more detail).

4 Automatic POS Tagging

One million words of the Mongolian corpus have been tagged as the current result of the project. The tagging procedure is shown in Figure 3.

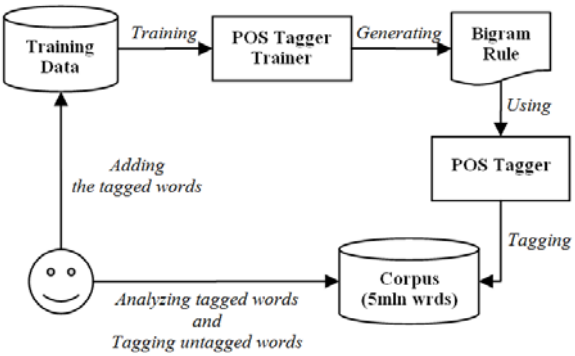

Figure 3. Automatic Tagging Procedure for Mongolian Corpus

When using the statistical POS tagger, the corpus tagging needs a training data. We have manually tagged around 110 thousand words. These 110 thousand words are used as the first training data. The statistical information on the first training data is shown in Table 3.

Words	Word type	Texts	Tags
112,754	21,867	200	185

Table 3. First Training Data

As shown in Table 3, the training data consists of 112,754 words. These words are divided into 21,867 types. This training data can be a good representative of the corpus because the texts in which distinct to total word ratio is higher are chosen (see Table 4).

No.	**Texts** (Files)	**Distinct Words**	**Total Words**	**Percent**
1.	MNCPR00320	113	125	0.9
2.	MNCPR00312	157	179	0.87
3.	MNCPR00118	118	136	0.86
4.	MNCPR00384	162	187	0.86
5.	MNCPR00122	238	279	0.85
6.	MNCPR00085	190	224	0.84
7.	MNCPR01190	320	379	0.84
8.	MNCPR00300	159	189	0.84
9.	MNCPR00497	241	288	0.83
10.	MNCPR00362	251	300	0.83

Table 4. Some Texts Chosen for Training Data

In Table 4, some of the texts that are chosen for the training data are shown. The most appropriate text that should be tagged at first is MNCPR00320 because its total words are 125 and distinct words are 113. Consequently, its equality of words types and total word is almost the same, 0.9. The first 200 texts from the corpus are manually tagged for the training data.

After training the bigram POS tagger, the corpus is tagged with it by 100 texts by 100 texts. After that, we manually checked the automatically tagged texts, and corrected the incorrectly tagged words and tagged the untagged words, in fact, new words to the training data. After manually checking and tagging, the automatically tagged texts are added to the training data for improving the tagger accuracy. Then, this whole process is done again and again. After each cycle, the training data is increased, and the accuracy of the tagger is also

improved. The statistics of automatic tagging the first 100 texts is shown in Table 5.

Words	Word type	Texts	Tags	Untagged word
73,552	9,854	100	108	16,322

Untagged word type	Mistagged words	Accuracy
3,195	310	76.5

Table 5. First 100 Texts Automatically Tagged

As shown in Table 5, the untagged words are 22 percent of the total words, and 0.5 percent is tagged incorrectly. Incorrectly tagged words are manually checked. The mistagged words are caused from the insufficient training data. In the result of the first automatic tagging, the tagger that is trained on around 110 thousand words can tag 76.5 percent of around 73 thousand words correctly.

In tagging the second 100 texts, the accuracy is almost the same to the previous one because the training data is collected from texts containing more word types. The correctly tagged words are 78 percent. After checking and tagging the automatically tagged 400 texts, the training data is around 260 thousand words as shown in Table 6.

Words	Word types	Texts	Tags
260,312	27,212	400	198

Table 6. Current Training Data

We tagged another 900 texts based on the training data in Table 6. They consist of around 860 thousand words, and 81 percent is tagged. The statistics is shown in Table 7.

Words	Word type	Texts
868,258	41,939	900

Untagged words	Untagged word types	Accuracy
168,090	19,643	81

Table 7. Automatically tagged words

As shown in Table 7, the bigram POS tagger trained on 260 thousand words has tagged around 700 thousand of 868 thousand words. The accuracy is nearly the same to the previous tagging accuracy. That means the first training data is well selected, and includes main usage words. Therefore the accuracy of the first tagged 200 texts is very close to that of 900 texts tagged later.

5 Conclusion

A research project building a 5 million word corpus is in its last phase. We have automatically tagged 1 million words of the corpus by developing a POS tagset and a bigram-POS tagger for Mongolian. The tagging accuracy is around 81 percent depending on the training data. Currently, the training data is around 260 thousand words. As increasing the training data, the accuracy of the tagger can be up to 90 percent. However, the increasing training data takes a lot of time and effort. The tagset currently consists of 198 tags. It may increase in the further tagging. In this year, we are planning to tag and check the 5 million word corpus.

Acknowledgments
Here described work was carried out by support of PAN Localization Project (PANL10n).

References

Brian Roark and Richard Sproat. 2007. *Computational Approaches to Morphology and Syntax*. Oxford University Press.

Christopher D. Manning and Hinrich Schutze. 1999. *Foundations of Statistical NLP*. MIT Press.

Daniel Jurafsky, James H. Martin. 2000. *Speech and Language Processing*. Singapore.

PANLocalization Project. 2007. *Research Report on Tagset for Mongolian.* Center for Research on Language Processing, National University of Mongolia.

Purev Jaimai and Odbayar Chimeddorj. 2008. *Corpus Building for Mongolian*. The Third International Joint Conference on Natural Language Processing, Hyderabad, India.