PAN LOCALIZATION PROJECT

ERROR ANALYSIS OF MONGOLIAN CORPUS

Phase 1.2

April 11, 2008

CENTER FOR RESEARCH ON LANGUAGE PROCESSING
NATIONAL UNIVERSITY OF MONGOLIA, ULAANBAATAR, MONGOLIA

Table of Contents

List of Tables

Abstract

This document provides an overview of the error analyzing of a Mongolian corpus, which consists of 1 million words, 100 thousand of which are tagged manually. The error analysis is discussed focusing on the used method and occurred errors.

1. Introduction

Building a corpus needs a lot of work and time. Millions of words are collected and tagged. This is implemented by the effort of many researchers working on the corpus for a long time, but the corpus building still requires working, that is, error analyzing [4, 5]. Studying and processing a language based on the corpus, it has to be standardized. Thus, the corpus has to be analyzed by hand or machine in both cases of the corpus has created manually or automatically. The error analyzing needs a lot of work like collecting, cleaning and tagging the corpus again.

For building a corpus for any language, it is necessary to design and create firstly a core corpus, and the corpus itself is built by extending the core corpus according to its principles. Any kind of errors is not allowed to be in the core corpus. We are creating a corpus for Mongolian. It will be 5 million words and tagged. This corpus will be the first core corpus in Mongolia.

In analyzing errors occurred in the corpus, we focused on verifying that the words are tagged correctly and the POS tagset is designed correct and appropriate. In the related works [2] on building a corpus for Mongolian in the past, there has been no POS tagset that is sufficient to use for the corpus tagging, and Mongolian linguists have not been using such technique for developing and describing Mongolian. Thus, we have spent much time on creating the tagged corpus because there have been many discussions and modifications after analyzing the tagged corpus.

We analyzed the errors of 100 thousand tagged words as three separate parts according to the POS tagsets [1]. The first analysis is in the first 15 thousand words tagged by the first POS tagset, the second is in the next 62 thousand words tagged by the draft version of the second POS tagset, and the third is in the last 27 thousand words tagged by the second POS tagset. Finally, we summarized these analyses.

2. Error Analysis

The main activity of error analyzing of the corpus is to verify that the developed POS tagset is appropriate and the words are tagged correctly. The error analyzing is a part of the corpus building, and how the tagged corpus is analyzed has to be described in this report.

The methodology to analyze the corpus is to concordance texts by POS tagset and words. We are using MTagger and a freeware concordancer program, AntConc [3]. Using these tools, we analyzed the tagged texts by searching, filtering, comparing, calculating the frequency and listing the words and the tags.

3. Error

In this section, the errors are presented in three separate parts according to three POS tagsets used for tagging 100 thousand words.

In the first tagging phase of the tagged corpus, there are around 15 thousand words tagged by the first POS tagset designed in the first phase (see Appendix A). The errors, occurred in this phase, are mainly caused by the inappropriate POS tagset design because its tags were too general, and a few are caused by the human mistakes. Most of the errors are occurred in adjective, modal, conjunction and compound words (see Table 1).

Table 1. Errors in the first 15K words tagged with the first POS Tagset

#	Actual Error	Tags (Incorrect → Correct)	Error Path in Corpus
1.	***бүх***	ADJ→PADJ	17219-2
2.	***эн тэргүүний*** зорилт	ADJ,NN→ADJ,ADJ	17221-13
3.	***орчин үеийн***	ADJ→ADJ,NC	17221-20
4.	***ураг зураг***	ADJ→ADJ,NC	17238-8-21
5.	***эрин үед***	ADJ→ADJ,NC	17238-9-7
6.	үйл ажиллагааг нь ***хөнгөн шуурхай, илүү үр нөлөөтэй*** болгох	ADJ→ADV	17219-24
7.	***илүү ажил хэрэгч*** болохыг	ADJ→ADV	17219-34
8.	***амин чухал*** болж байна	ADJ→ADV	17219-36
9.	***харьцангүй өндөр хурдацтай*** байсан	ADJ→ADV	17220-7
10.	***удаашралтай...зогсонги*** болсны	ADJ→ADV	17220-8
11.	***хязгаарлагдмал*** байгаатай	ADJ→ADV	17221-3
12.	***дутагдалтай*** байгааг	ADJ→ADV	17221-9
13.	***янз бүр*** байж болох	ADJ→ADV	17221-14
14.	***сонгогч*** түүний идэвхтэй шургуу ажиллагаагаар	ADJ→NC	17219-15
15.	***бусад*** салбарын	ADJ→PADJ	17219-29
16.	зарим	ADJ→PADJ	17220-9
17.	***ямарваа***	ADJ→PADJ	17221-11
18.	***нийт***	ADJ→PADJ	17221-13,18,21,25,27
19.	***одоо*** үед	ADV→ADJ	17219-13
20.	***шууд*** хамаатай	ADV→ADJ	17219-31
21.	***бол***	MM→CJ	17219-5
22.	эхлэл нь ***нэгэнт*** тавигдсан	MW→	17221-21
23.	***ач холбогдолтой*** юм	MW→ADAJ	17251-11,41
24.	***зүйн хэрэг***	MW→ADJ,NC	17221-20
25.	***улмаар***	MW→CJ	17219-13
26.	***ялангуяа***	MW→CJ	17219-15
27.	***харин***	MW→CJ	17219-25
28.	юуны өмнө	MW→CJ	17221-3
29.	***гол төлөв***	MW→CJ	17221-10
30.	***үнэн хэрэгтээ***	MW→CJ	17221-11
31.	***үнэндээ***	MW→CJ	17221-15
32.	***зөвхөн***	MW→CJ	17221-17
33.	***ер нь***	MW→CJ	17221-24
34.	***бие даасан***	NC,Vt→ADJ	17219-25
35.	***бүгд хурал***	NC→ ADJ,NC	17219-3

36.	***хүчигтэй***	NC→ADJ	17221-20
37.	***аварга*** малчдын эгнээ	NC→ADJ	17249-6,12
38.	***дээрээс*** *нь* урьдчилан харж	NC→ADV	17219-15
39.	***Намын Төв Хорооны***	NC→NC,ADJ,NC	17221-13
40.	***нэг нь*** нийтийн төлөө	NC→NN	17265-17,38
41.	нийт нь ***нэгийн*** төлөө	NC→NN	17265-17,43
42.	***Батмөнх***	NC→NP	17271-24,4
43.	***Сүхбаатар***	NC→NP	17245-13,56
44.	***ажиллахтай*** шууд холбоотой	NC→Vin	17219-32
45.	***хангах*** нь	NC→Vt	17267-17,28
46.	***хурлын*** хяналт	NP→NC	17219-16
47.	***юм бүхэн***	NPPN,CJ→NC,CJ	17221-11
48.	***тэдний***	NPPN→PPN	17221-24
49.	сайжруулбал ***зохино***	Vin→MW	17221-16
50.	***зонхилон*** хэрэглэж	Vin→VT	17219-36
51.	эргэж ***чадахгүйд*** хүрвэл	Vt→MW	17919-12
52.	***чадахуйц***	Vt→MW	17220-8
53.	салбарт ***хийх***	Vt→Vin	17219-29

In Table 1, the actual error forms are listed in the first column. The incorrectly tagged words are underlined, italic and bold. In the next column, two tags are aligned; one in the left side of the arrow is the incorrect tag and other in the right side of the arrow is the correct tag. In the last column, the numbers indicate the text ID and the sentence ID where the errors are located in the corpus. The groupings of these errors are shown in Table 2.

Table 2. Head errors in the first 15K words tagged with the first POS Tagset

#	POS Tag Name (Incorrect → Correct)	POS Tag (Incorrect → Correct)
1.	Adjective → Pro-adjective	ADJ→PADJ
2.	Adjective, Numeral → Adjective , Adjective	ADJ,NN→ADJ,ADJ
3.	Adjective → Adjective, Common Noun	ADJ→ADJ,NC
4.	Adjective → Adverb	ADJ→ADV
5.	Adjective → Common Noun	ADJ→NC
6.	Adjective → Pro-adjective	ADJ→PADJ
7.	Adverb → Adjective	ADV→ADJ
8.	Modal Morph → Conjunction	MM→CJ
9.	Modal Word → Adverb	MW→ADV
10.	Modal Word → Adjective	MW→ADJ
11.	Modal Word → Adjective, Common Noun	MW→ADJ,NC
12.	Modal Word → Conjunction	MW→CJ
13.	Common Noun, Transitive Verb → Adjective	NC, Vt →ADJ
14.	Common Noun → Adjective, Common Noun	NC→ ADJ,NC
15.	Common Noun → Adjective	NC→ADJ
16.	Common Noun → Adverb	NC→ADV
17.	Common Noun → Common Noun, Adjective, Common Noun	NC→NC,ADJ,NC
18.	Common Noun → Numeral	NC→NN
19.	Common Noun → Pronoun	NC→NP
20.	Common Noun → Intransitive Verb	NC→Vin
21.	Common Noun → Transitive Verb	NC→Vt
22.	Pronoun → Common Noun	NP→NC
23.	Non-possessive Pronoun, Conjunction → Common	NPPN,CJ→NC,CJ

	Noun, Conjunction	
24.	Non-possessive Pronoun → Possessive Pronoun	NPPN→PPN
25.	Intransitive Verb → Modal Word	Vin→MW
26.	Intransitive Verb → Transitive Verb	Vin→VT
27.	Transitive Verb → Modal Word	Vt→MW
28.	Transitive Verb → Intransitive Verb	Vt→Vin

As shown in Table 2, the errors are occurred in adwords, modals, nouns, and verbs. Most of these errors are considered confusion tags in the part of speech tagging for Mongolian.

In the second tagging phase of the corpus tagging, the next 62 thousand words were manually tagged by the draft version of the second POS tagset. During this time, we were developing the second POS tagset practically. The occurred head errors are shown in Table 3.

Table 3. Errors in 62K words tagged with the draft version of the second POS tagset

#	POS Tag Name (Incorrect → Correct)	POS Tag (Incorrect → Correct)
1.	Adjective,adjective,common noun plural nominative →Common noun nominative,common noun nominative,numeral cardinal	AJ,AJ,NPN→NN,NN,NUC
2.	Adjective,adjective→Adverb,adverb	AJ,AJ→AV,AV
3.	Adjective→Adjective,adjective	AJ→AJ,AJ
4.	Adjective→Adjective,numeral cardinal	AJ→AJ,NUC
5.	Adjective→Adjective superlative	AJ→AJS
6.	Adjective→Adverb	AJ→AV
7.	Adjective→Proper noun	AJ→RN
8.	Adjective comparative→Adjective	AJC→AJ
9.	Adverb→Adverb comparative	AV→AVC
10.	Adverb→Verb active connecting	AV→VAC
11.	Adverb active connecting→Verb active connecting	AVAC→VAC
12.	Adverb active connecting→Verb passive connecting	AVAC→VPC
13.	Auxiliary verb future negative→Modal future negative	AXFX→MDFX
14.	Abbreviated noun possessive→Abbreviated noun possessive nominative	HAN→HANN
15.	Modal→Modal future	M→MODF
16.	Modal negative→Modal	MODX→MOD
17.	Common noun→Common noun locative	N→NL
18.	Common noun→Common noun nominative	N→NN
19.	Common noun ablative→?	NB→?
20.	Common noun accusative→Verb active connecting	NC →VAC
21.	Common noun accusative →Adjective,common noun accusative	NC→AJ,NC
22.	Negative,negative→Verb active future, verb active future	NEG,NEG→VAF,VAF
23.	Negative→Punctuation	NEG→PUN
24.	Common noun genitive→Adjective,common noun genitive	NG→AJ,NG
25.	Common noun genitive→Adjective	NG→AJ
26.	Common noun genitive→ Adjective,common noun genitive	NG→AJ,NG
27.	Common noun nominative→Adjective,common noun nominative	NN→AJ,NN
28.	Common noun nominative→Common noun plural nominative	NN→NPN
29.	Common noun plural accusative→Common noun accusative	NPC→NC
30.	Common noun plural accusative→Common noun accusative,common noun accusative	NPC→NC,NC
31.	Common noun plural genitive→Common noun genitive	NPG→NG
32.	Common noun plural nominative→Common noun nominative	NPN→NN
33.	Common noun plural nominative→Pronoun plural nominative	NPN→PPN

34.	Common noun plural nominative→Verb active connecting,common noun plural nominative	NPN→VAC,NPN
35.	Numeral approximate→Numeral approximate,numeral cardinal	NUA→NUA,NUC
36.	Numeral ordinal→Adjective	NUO→AJ
37.	Pronoun→Pronoun genitive	P→PG
38.	Pronoun plural→Pronoun plural nominative	PP→PPN
39.	Pronoun plural genitive→Adjective	PPG→AJ
40.	Pronoun plural genitive→Common noun genitive	PPG→NG
41.	Postposition→Conjunction coordinate,adjective, common noun nominative,possessive	PTP→JC,AJ,NN,POS
42.	Proper noun genitive→Adjective,common noun genitive	RG→ADJ, NG
43.	Proper noun genitive→Common noun genitive	RG→NG
44.	Verb active connecting→Common noun accusative, verb active connecting	VAC→NC,VAC
45.	Verb active connecting→Verb active continues	VAC→VAT
46.	Verb active connecting→Verb passive connecting	VAC→VPC
47.	Verb active future→?	VAF→?
48.	Verb active future→Common noun accusative,verb active future	VAF→NC,VAF
49.	Verb active future→Verb active connecting	VAF→VAC
50.	Verb active future negative→Verb active past negative	VAFX→VAPX
51.	Verb active past→Verb active present	VAP→VAPr
52.	Verb active person→?	VA***PS***→?
53.	Verb passive future→Adverb,auxiliary verb passive future	VPF→AV,***AXPF***
54.	Verb passive future→Verb active connecting,verb passive future	VPF→VAC,VPF
55.	Adverb→Adverb comparative	AV→AVC

In Table 3, 55 kinds of errors occurred in the 62 thousand tagged words are listed. During tagging these 62 thousand words, we were still developing the draft version of the second POS tagset. Thus, a lot of errors were occurred in this phase, but most of them had been corrected after detected. Consequently, when the errors began occurring few, we considered that the draft version of the new POS tagset developed, and released the second POS tagset.

In the third phase of the tagging, there have been still errors eventhough the words were tagged with the second POS tagset. The errors were made by human mistakes, incorrectly analyzing the words and marked the incorrect tags. The errors occured in the last 27 thousand words tagging are shown in Table 4.

Table 4. Errors in the last 27K words tagged with the second POS tagset

#	POS Tag Name (Incorrect → Correct)	POS Tag (Incorrect → Correct)
1.	Common noun→Common noun nominative	N→NN
2.	Common noun locative→Verb passive connecting , common noun locative	NL→VPC,NL
3.	Common noun genitive→Adjective	NG→JJ
4.	Common noun plural nominative→Common noun plural	NPN→NP
5.	Common noun plural ablative→Adjective,common noun ablative	NPB→JJ,NB
6.	Pronoun genitive→Adjective,common noun genitive	PNG→PJ,NG
7.	Ad-Adjective→Adjective	JJA→JJ
8.	Adjective→Adjective,adjective	JJ→JJ,JJ
9.	Adjective→Adverb	JJ→RB
10.	Verb active connecting→Postpostion	VAC→PT
11.	Verb active connecting→Verb active continues	VAC→VAG
12.	Verb active connecting→Verb passive continues	VAG→VPG
13.	Verb active continues→Verb active connecting	VAG→VAC

14.	Verb active present→Verb active past	VAP→VAD
15.	Verb active past→Common noun accusative,verb active past	VAD→NC,VAD
16.	Verb passive future→Verb passive infinitive/base	VPF→VPB
17.	Verb active future nominative→Verb active infinitive/base nominative	VAFN→VABN
18.	Verb active infinitive/base negative→Verb active present negative	VABX→VAPX
19.	Proper noun genitive→Common noun genitive	RNG→NG
20.	Adverb→Common noun instrumental	RB→NI
21.	Conjunction subordinate→Conjunction coordinate	CS→CC
22.	Conjunction coordinate→Conjunction subordinate	CC→CS
23.	Abbreviation→Abbreviation nominative	ABR→ABRN
24.	Abbreviation nominative→Abbreviation	ABRN→ABR
25.	Abbreviation genitive,abbreviation genitive→ Abbreviation	ABRG,ABRG→ABR
26.	Modal active future negative→Modal active infinitive/base negative	MDAFX→MDABX

In Table 4, most of the errors are caused by ambiqusties of Mongolian. The most confusing parts of speech are occured in verbs, and then modal and conjungation.

4. Conclusion

In this research report, we have provided the error analyzing of Mongolian corpus. The corpus consists of 1 million words, and 100 thousand of them have been manually tagged. The implementation of the corpus tagging was divided into three parts because of the POS tagset development phases. In other words, we have developed the POS tagset for Mongolian by designing it three times during 100 thousand words tagging. Therefore, the error analyzing of these tagged words are also divided into three parts: analyses on the first 15 thousand words, the next 62 thousand words, and the last 27 thousand words.

The tagged corpus has been manually analyzed by using a concordance program, AntConc, and MTagger. For analyzing and detecting errors in the tagged texts, we concordance the texts with POS tags and words, and create various kinds of tables and lists showing statistical information about the tagged corpus. Analyzing these kinds of information is our main methodology for the corpus analyzing.

We have spent a lot of time on tagging 100 thousand words because the POS tagset has been developed during at that time. Therefore, the error analyzing of the corpus is still continuing now. Analyzing and maintaining the corpus requires a lot of work like building it. Accordingly, it is necessary to determine the principles of the corpus analyzing and maintaining after the corpus are tagged. It takes more time, and needs more experience for us because it is the first corpus building for Mongolian. Thus, we are developing a guidelines based on the experience of tagging and analyzing 100 thousand words for how to analyze and maintain the tagged corpus.

Most of the errors occurred in the corpus tagging is caused by the POS tagset design. Mongolian is one of the highly agglutinative languages such as Finnish and Turkish [6]. In Mongolian syntax, the inflectional functions are appeared as both of a suffix or an isolated word (separated with white spaces in the orthographic form, but the meaning is just an inflectional function such as plural case). Moreover, there are many ambiguities between morphological structure and syntax position for one form. For example, "-тай" (*-tai* in Roman transcription) is a suffix featured as both of inflectional and derivational, and when it is occurred as a derivational suffix, there are still ambiguities except the ambiguity of being inflectional or derivational. It could be an adjective or a modal in both of morphological and syntax features. Currently, such challenges make it difficult to design the POS tagset completely for Mongolian.

Appendix A: First POS Tagset

Table 5. First POS Tagset

Category	POS Tag ID No.	POS Name	POS Tag
Noun	1	Common Noun	NC
	2	Proper Noun	NP
	3	Numeral	NN
Verb	4	Transitive Verb	Vt
	5	Intransitive Verb	Vin
	6	Auxiliary Verb	AUX
Ad-word	7	Adjective	ADJ
	8	Adverb	ADV
Modal	9	Modal Word	MM
	10	Modal Morph	MW
Pro-word	11	Non-possessive pronoun	NPPN
	12	Possessive pronoun	PPN
	13	Pro-numeral	PNN
	14	Pro-adjective	PADJ
	15	Pro-adverb	PADV
	16	Pro-verb	PV
Conjunction	17	Conjunction	CJ
Interjection	18	Interjection	INT
Abbreviation	19	Abbreviation	ABR
Punctuation	20	Punctuation	PUN

Reference

[1] Manual Corpus Tagging for Mongolian
Research Report, PANLocalization Project II

[2] Building a Corpus for Mongolian Language
Purev Jaimai and Odbayar Chimeddorj, The Third International Joint Conference on Natural Language Processing, January 2008, Hyderabad, India

[3] AntConc: *A freeware concordance program, Laurence Anthony, Waseda University, Japan*
www.antlab.sci.waseda.ac.jp

[4] Detecting Errors within a Corpus using Anomaly Dectection
Eleazar Eskin, Department of Computer Science, Columbia University

[5] Correction of Errors in a Modality Corpus Used for Machine Translation Using Machine-learning
Masaki Murata, Masao Utiyama, Kiyotaka Uchimoto, Qing Ma, and Hitoshi Isahara
Communications Research Laboratory

[6] Computational Approaches to Morphology and Syntax
Brian Roark and Richard Sproat, Oxford University Press, 2007