Development of Traditional Mongolian Dependency Treebank

Xiangdong Su Guanglai Gao Xueliang Yan

College of Computer Science
Inner Mongolia University
Huhhot China 010021
csggl@imu.edu.cn

Abstract. This paper describes the development of Traditional Mongolian dependency treebank (TMDT) which aims to facilitate the dependency analysis on Traditional Mongolian. The annotation scheme of the dependency treebank is established according to Traditional Mongolian grammar and its usability in syntactic analysis. In the treebank, morphological and analytical information are annotated. At morphological level, a semi-automation strategy is adopted. Part-Of-Speech (POS) and stem of each word in the sentence are tagged and extracted respectively with automation tools, and then manually corrected. At analytical level, the dependencies in the sentence are only annotated manually according to constituent structure and the annotation scheme. This treebank formulates the foundation of dependency parsing on Traditional Mongolian and can be extended to a multi-dependency Treebank.

Keywords: Traditional Mongolian, Dependency Treebank, Morphological Information, Analytical Dependency

1 Introduction

Syntactic parsing is always the active research area in natural language processing. In this field, much progress has been made in its theories and several applicable systems have been developed. As a subfield of syntactic parsing, dependence parsing recently gains more and more attention among researchers since it provides useful information in many document-analysis related applications. In dependency parsing, dependency treebank is required to train the parser and evaluate its performance. So far, treebanks have been constructed for many languages, including English, French, German, Spanish, Turkish, Russian, and so on. To facilitate the dependency analysis in Traditional Mongolian, we develop a Traditional Mongolian dependency treebank (TMDT) on the basis of in-depth study of Traditional Mongolian grammars and its usability in syntactic analysis.

As one Asian language, Traditional Mongolian is derived from Uyghur and used in Inner Mongolia and neighboring regions. It has over 5 million speakers. Traditional Mongolian possesses agglutinative word structure with complex inflection and deriva-

tion. Its word is consisted of letters which are connected along a straight line, called spine. Each letter has as many as three different shapes depending on whether the letter appears in an initial, medial, or final position. In some cases, additional graphic variants are selected for visual harmony with the subsequent letter. From the syntactic viewpoint, Traditional Mongolian has SOV constituent order and the predicate is not necessary to be a verb or copula. There are totally eight kinds of case in Traditional Mongolian. Case markings on nominal constituents usually indicate their syntactic role. Some cases act as preposition or conjunction. We take the above mentioned characteristics into consideration in building the dependency treebank.

This paper mainly describes the annotation scheme of the treebank, and simply reports the annotation procedure as well as the final production. TMDT takes a two-level annotation structure including morphological level and analytical level. Morphological level annotates the POS and stem of each word in the sentence, considering their importance for Traditional Mongolian word. Analytical level annotates the binary dependency relationship holding between a syntactically subordinate word, called child, and the other word on which it depends, called parent. Morphological information contributes to the analytical annotation. The annotation process is divided into two phases according to the annotated information: morphological annotation phase and analytical annotation phase. In morphological phase, a semi-automation strategy is employed to speed up the annotation process. In the analytical phase, the dependency relationships are only annotated manually without any automaton's assistance. The whole Treebank is reviewed carefully to ensure its correctness.

The rest of this paper is structured in the following way: Section 2 mentions some related work about dependency treebank development. Section 3 describes the annotation scheme including the annotation principles, structure and tag set. Section 4 reports the annotation workflow, treebank format and final production. An example is provided to intuitively present the resulting treebank. Finally, section 5 concludes this paper and points out the future direction.

2 Related Work

Much past work is related with treebank building. M.P. Marcus et al. in [1] present the influential treebank, penn treebank, which leads the way in building annotated treebank and serves as an excellent model for treebank building of Traditional Mongolian. B. Rajesh et al. in [2] depict the creation of Hindi/Urdu multi-representational and multi-layered treebank. A. Böhmová et al. in [3] describe a three level annotation scheme in building the Prague dependency treebank, including Morphological level, analytical level and tectogrammatical level respectively. They manually annotate the dependency treebank, and automatically generate the phrase structure tree bank from the dependency treebank. C.-R. Huang et al. in [4] specify the design criteria and annotation guidelines of Sinica treebank. The three design criteria are: Maximal Resource Sharing, Minimal Structural Complexity, and Optimal Semantic Information. P. Pajas and J. Štěpánek in [5] propose an annotation framework that was designed to be extensible and independent of any particular annotation schema. M.-C.d. Marneffe

and C.D. Manning in [6] examine the Stanford typed dependencies representation, which was designed to provide a straightforward description of grammatical relationships.

In dependency treebank annotation, the core task is discerning the dependency relationships which vary with languages and dependency grammars. I.A. Melčuk in [7] discusses morphological, syntactic and semantic dependencies in Meaning-Text theory. R. Hudson in [8] details the English dependency theory. J. Nivre in [9] reviews the dependency detection criteria.

Manually and semi-automation annotations are the mainstream of annotation strategy. T. Brants et al. in [10] explore (1) the automation of Treebank annotation, (2) the comparison of conflicting annotations and (3) the inconsistencies detection in automatic annotation. L.v.d. Beek et al. in [11] use Alpino parser and parse selection tool to facilitate the annotation process of Alpino dependency treebank.

The following works are also related to our work. J. Lafferty et al. in [12] view POS tagging as a sequence labeling problem and obtain a better performance with CRFs tagger. M.-Y. Ma in [13] brings forward a mixed model for Traditional Mongolian Stemming. W.-B. Jiang et al. in [14] employ Lexical Analyzer based on directed graph to segment the stem and affix of Mongolian word. More information about Traditional Mongolian Lexical and syntactic grammar can be found in the book [15].

3 Annotation Scheme

So far, most of the available dependency treebanks take into account both morphological and analytical information. Yet Prague dependency treebank encodes the dependencies at semantic level. In TMDT, we just annotate the sentences at two levels: morphological level and analytical level. Firstly, the morphological captures the basic attributes of the syntactic units (words) and helps to the annotation at analytical level. The analytical level specifies the dependency information of the sentences. Secondly, semantic annotation is very complex process and relates to the deep structure of the sentence. We take no account of the semantic dependency in TMDT since there are still some disagreements among researches about the semantic relationship in Traditional Mongolian. Furthermore, the criterion of minimal structural complexity is adopted to ensure that the assigned structural information can be used without any assumption about the user's background. We will deal with the annotation levels in turn, starting with morphological level in section 3.1 and continuing with the analytical level in section 3.2.

3.1 Morphological Level

Morphological information expresses the attributes of the syntactic units and plays an important role in syntactic analysis. In TMDT, The morphological annotation principles are as follows.

1. Annotation unit

As mentioned above, many words in Traditional Mongolian are produced through inflection and derivation. The change in the form of a word (typically the ending) usually expresses a grammatical function or attribute such as tense, mood, number, case, and gender. We treat the derivative as annotation unit and take no consideration of the inflection and derivation phenomenon, except the case inflection phenomenon. For instance, adding a tense suffix "ᠵᠠ", "ᠪᠤ" (whose meaning is "go" in English) becomes "ᠵᠠᠪᠤᠰᠤ" (whose meaning is "have gone" in English). "ᠵᠠᠪᠤᠰᠤ" is treated as a basic annotated unit. Although our strategy increases the amount of lemmas in dictionary, it simplifies the annotation process and treebank representation.

2. Case inflection

Case inflection is the phenomenon that adding a case to a noun, adjective, or pronoun that express the semantic relation of the word to other words in the sentence. Some cases are separated with the previous word they attach to by a common blank (Unicode 0X0020). The other cases are separated with the previous word they attach to by Mongolian blank (Unicode 0X202F). In Traditional Mongolian grammar, the attached word and the case in the latter situation are considered as a single word. We take a different perspective and treat them as two annotation units considering the case function in sentences. This is more suitable to practical application.

3. Annotating Part-Of-Speech (POS) tag

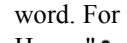
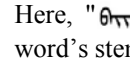
POS are known as word classes or lexical categories and greatly related to the constituent role of syntactic unit. So POS is annotated at morphological level. POS tagging is the process of classifying words into their POS. The collection of POS tags used in TMDT is listed in Table 1.

4. Annotating word's stem

In computational linguistics, each word in Traditional Mongolian is made up of a

Table 1. POS Tags of Traditional Mongolian in TMDT

Category	Description	Category	Description
NN	noun	NUM	number
QUAN	quantity	PRON	pronoun
ADJ	adjective	ADV	adverb
VB	verb	POSS	Possecive pronoun
CONJ	conjunction	MOD	modal verb
MOOD	mood word	WH	interrogative word
LEXAUX	combining form	AUXI	auxiliary word
PUNC	punctuations	FORW	foreign words
NOMCA	nominative case	ACCA	accusative case
REFCA	reflexive case	INSCA	instrumental case
GENCA	genitive case	DLCA	dative-locative case
ABLCA	ablative case	COMCA	comitative case
TEPO	temporal and positional words	REFVB	special words link thinking and speech

solo stem or a stem with one or more suffixes. The stem of a word is the part which is common to all its inflected variants. This means that the stem is an essential attribute of Traditional Mongolian word and represents the original meaning of the word. For example, "" is a new word produced by adding "  " to "". Here, "" is the stem, and "  " is the suffix. Therefore, we integrate the word's stem into the target dependency treebank.

3.2 Analytical level

In the framework of dependency grammar, we hold an assumption that the syntactic structure essentially consists of syntactic units (words) linked by binary, asymmetrical relationship, called dependency relationship. When building the analytical level of TMTD, our target is to annotate the dependency relationships of the sentences. To implement the treebank annotation, we first work out the annotation principles at analytical level as follows.

1. Annotation unit

The annotation unit is same as the annotation unit in morphological level.

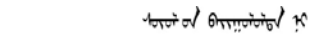
2. Annotation structure

The dependency can be represented in the form of an acyclic directed graph which is similar to the syntax graph described in Tiger Language [16]. In the graph, the words in the sentence are employed as the node labels, and each pair of nodes linked by an arc means there is a dependency relationship. No circles are allowed in the graph. The resulting graph is a tree, called dependency tree. The treebank is non-projective. That is to say, cross edges are allowed.

3. Nonverbal word as predicate

The nonverbal word is treated the same as the verb when it acts as a predicate.

4. Case processing

In Traditional Mongolian, the case utility is a notable feature. These cases perform different syntactic functions. Take the sentence  for example. Here, the word "" in the sentence is called accusative case (a constituent which appears after "" and indicates that "" is an object); the word "  " in is called nominative case (a constituent which appears after "" and indicates that "" is a subject). There are also other case types, including genitive case, instrumental case, dative-locative case, ablative case, and comitative case. We classify the dependency relationship between the case word and its related constituent according to its syntactic function.

The emphasis of syntactic annotation is to discern the dependency relationships including its presence, directions and category in the sentences. This is a linguistics-related task. We use heuristic strategies to detect dependency relations and distinguish the head and the dependent in such relations. The strategies are: (1) whether two syntactic units formulate a structure C, (2) whether the head H determines the category of C and (3) whether H selects the dependent D, determines the presence of D or restricts the position of D.

Table 2. Categories of Analytical Dependency in TMDT

Category	Description
clau	dependency between main clause and subordinate clause
indcla	dependency between main clause and independent constituent
nsubj	dependency between predict and subject
nobj	dependency between predict and object
modi	dependency between the modifier and the object been modified
aux	dependency between the auxiliary and the object it act on
advmod	dependency between the adverbial modifier and predict
nomca	dependency between the true subject and nominative case
acca	dependency between the true object and accusative case
genca	dependency between the modifier and genitive case
ablca	dependency between the object of adverbial phrase and ablative case
dlca	dependency between the object of adverbial phrase and dative-locative case
insca	dependency between the object of the adverbial phrase and instrumental case
comca	dependency between the object of adverbial phrase and comitative case
refl	dependency between the referenced content and REFVB
conj	dependency between the conjunction and the first object it linked
coord	dependency between the coordination or appositional components
lex	lexical relation between two words
punct	dependency between the punctuation and the part it act on
dep	dependency that is unable to determine a more precise relation
root	root

In TMDT, 21 kinds of analytical dependency are defined according to the annotation principles and the syntactic characters of Traditional Mongolian. Table 2 lists all the dependencies together with their descriptions.

4 Treebank Building

This section describes the annotation workflow, treebank format and final production in TMDT building.

4.1 Annotation workflow

Fig. 1 shows the annotation workflow in treebank development. To speed up the annotation process, we use a CRF tagger proposed in [12] to label POS tags and implement the algorithm (StemExtractor) proposed in [13] to extract the stem for each word. However, purely automatic annotation without supervision is not reliable. Therefore, we manually correct POS tag and stem resulting from the automation tools

subsequently. The dependency relationships at the analytical level are just annotated manually without any automaton’s assistant. In this process, two annotators independently choose analytical tag from the dependency list defined in the annotation scheme and use an annotator-arbiter strategy to resolve the conflicts between them. This ensures the correctness of the target dependency treebank.

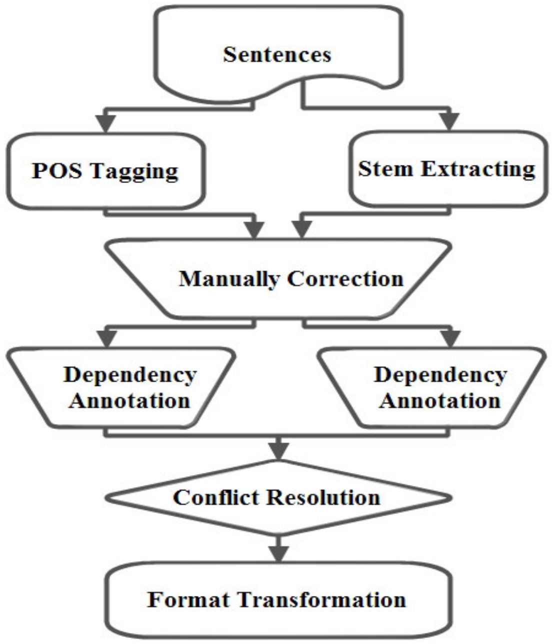


Fig. 1. Workflow of the Annotation Procedure

4.2 Output format and final production

Two types of treebank formats were employed in TMDT. The first format, which is similar to CoNLL-08, is demonstrated in Table 3. The first column represents the orders of syntactic units in sentences, and the second column represents the syntactic units. The remnants are stem, POS tag, dependency category and head respectively. The fifth column represents the dependency between the heads whose sequence is the corresponding number and the dependent which is the word in the corresponding row. TMDT can be extended to multi-dependency treebank. The output format is just adding a “rel” column and a “head” column in the first format. Table 3 is the first format of the sentence "[illegible]". The second format is XML because of its popularity, its ease of understanding and its wide use in description of linguistic information. It is automatically transformed from the first format.

The annotated corpus comes from Inner Mongolian daily. The content involves all aspects of the social life. These sentences conform to the grammar, and are few of mistakes. This is convenient for treebank annotation and exploiting the linguistic phe-

nomenon in Traditional Mongolian. The annotation corpus is in Mongolian Universal Coding.

TMDT is composed of 400 sentences (13028 annotated words) from Inner Mongolian daily. Among these sentences, the shortest one contains 9 words and the longest one contains 54 words. The average length is 32.57 words. Simple sentences account for 27%, and clauses account for 73%. There are 4460 distinct words and 1548 distinct stems. The non-projective trees account for 2.25% in the whole treebank. Fig. 2 shows the annotation result of "[illegible]" at morphological level and analytical level in TMDT.

Table 3. The first format of "[illegible]"

seq.	word	stem	POS	rel	head
1	[illegible]	[illegible]	NN	poss	2
2	[illegible]	[illegible]	POSSCA	modi	3
3	[illegible]	[illegible]	NN	subca	4
4	[illegible]	[illegible]	SUBCA	nsubj	11
5	[illegible]	[illegible]	NN	poss	7
6	[illegible]	[illegible]	NN	lex	5
7	[illegible]	[illegible]	POSSCA	modi	8
8	[illegible]	[illegible]	NN	objca	10
9	[illegible]	[illegible]	NN	lex	8
10	[illegible]	[illegible]	OBJCA	nobj	11
11	[illegible]	[illegible]	VB	root	0
12	[illegible]	[illegible]	AUXI	aux	11
13	[illegible]	[illegible]	PUNC	punct	11

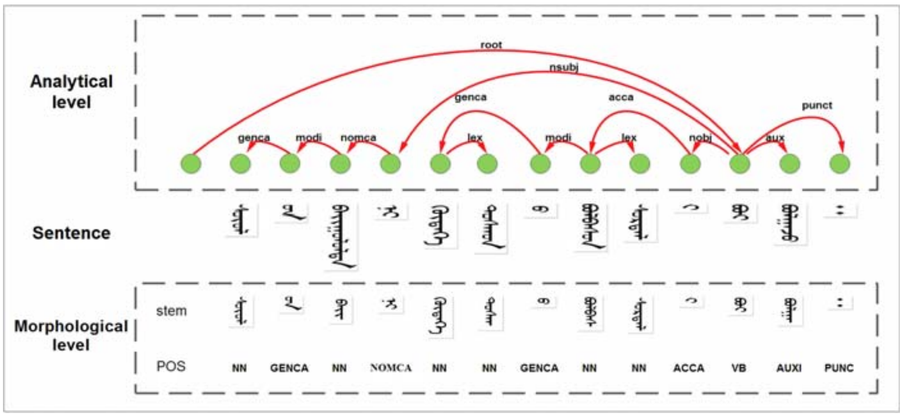


Fig. 2. Annotation Scheme of "[illegible]" in TMDT

5 Conclusion and Future Work

A dependency treebank is very important in syntactic analysis. However, there is no suitable dependency treebank for Traditional Mongolian. For this reason we start to develop Traditional Mongolian dependency treebank by annotating the corpus coming from Inner Mongolian daily with dependency structures. This paper describes the annotation scheme, annotation procedure and the final production. This treebank is annotated at morphological level and analytical level, which labels the POS tags, word's stem and the syntactic dependency relationships. Our work yields promising results, indicating the annotation scheme of TMDT treebank is essential to the success of building a multi-layered treebank. This treebank can be extended to a multi-dependency treebank, in which many types of dependency relationship co-exist between two syntactic units in sentences. This work formulates the foundation of dependency parsing on Traditional Mongolian. This is not an end, but rather a road-map to the syntactic analysis on Traditional Mongolian, with some progress along the way, since the theories of linguistics is still in development.

In the future, we will continue to expand the treebank's scale and attempt to convert it into a phrase structure treebank using statistical methods.

Acknowledgments

This work is supported by National Natural Science Foundation of China (Grant No. 61263037) and Major Program of Natural Science Foundation of Inner Mongolia of China (Grant No. 2011ZD11).

References

1. Marcus, M., Santorini, B., Marcinkiewicz, M.: Building a Large Annotated Corpus of English: The Penn Treebank. Computational Linguistics 19, 313-330 (1994)
2. Bhatt, R., Narasimhan, B., Palmer, M., Rambow, O., Sharma, D.M., Xia, F.: A Multi-representational and Multi-layered Treebank for Hindi/Urdu. Proceedings of the Third Linguistic Annotation Workshop, pp. 186-189. Association for Computational Linguistics, Suntec, Singapore (2009)
3. Böhmová, A., Hajič, J., Hajičová, E., Hladká, B.: The Prague Dependency Treebank: A Three-Level Annotation Scenario. In: Abeillé, A. (ed.) Treebanks: Building and Using Syntactically Annotated Corpora, pp. 103-127. Kluwer Academic Publishers (2001)
4. Huang, C.-R., Chen, F.-Y., Chen, K.-J., Gao, Z.-M., Chen, K.-Y.: Sinica Treebank: Design Criteria, Annotation Guidelines, and On-line Interface. Second Chinese Language Processing Workshop, pp. 29-37. Association for Computational Linguistics, Hong Kong, China (2000)
5. Pajas, P., Štěpánek, J.: Recent Advances in a Feature-Rich Framework for Treebank Annotation. Proceedings of the 22nd International Conference on Computa-

tional Linguistics, vol. 1, pp. 673-680. Association for Computational Linguistics, Manchester, United Kingdom (2008)
6. Marneffe, M.-C.d., Manning, C.D.: The Stanford Typed Dependencies Representation. Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation, pp. 1-8. Association for Computational Linguistics, Manchester, United Kingdom (2008)
7. Mel'čuk, I.A.: Dependency Syntax: Theory and Practice. State University of New York Press,, New York (1988)
8. Richard Hudson: An Introduction to Word Grammar. Cambridge University Press, Cambridge (2010)
9. Joakim Nivre: Dependency Grammar and Dependency Parsing. Technical Report, School of Mathematics and Systems Engineering,Växjö University (2005)
10. Brants, T., Skut, W.: Automation of Treebank Annotation. Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning, pp. 49-57. Association for Computational Linguistics, Sydney, Australia (1998)
11. Beek, L.v.d., Bouma, G., Malouf, R., Noord, G.v.: The Alpino Dependency Treebank. Computational Linguistics in the Netherlands (CLIN), (2002)
12. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Proceedings of the Eighteenth International Conference on Machine Learning, pp. 282-289. Morgan Kaufmann Publishers Inc. (2001)
13. Ma, M.-Y.: Researching of Mongolian Word Segmentation System Based on Dictionary, Rules and Language Model (in Chinese). Computer Science, Inner Mongolian University, vol. master, (2011)
14. Jiang, W.-B., Wu, J.-X., Wuriliga, Nashunwuritu, Liu, Q.: Discriminative Stem-Affix Segmentation for Directed-Graph-Based Mongolian Lexical Analyzer (in Chinese). Journal of Chinese Information Processing 25, 30-34 (2011)
15. Qinggeertai: Traditional Mongolian Grammar (in Chinese). Inner Mongolian press, Huhhot, China (1992)
16. König, E., Lezius, W.: The TIGER Language: A Description Language for Syntax Graphs, Formal Definition. (2003)