The development of tagged Uyghur corpus

Yusup Aibaidula
Department of Computer Science of Physi-Math, Information Xinjiang Normal University, Urumqi, Xinjiang, China
E-mail: yusup2002@sohu.com

Kim-Teng Lua*
School of Computing, National University of Singapore, 3 Science Drive 2, Singapore 119260
luakt@comp.nus.edu.sg

Abstract: The history and development of Uyghur language is introduced. After a brief introduction to the development of Uyghur words, morphology and syntax, we explain our developing of a computer-aided contemporary Uyghur language tagging system. The coverage of this corpus, the resources building, the rules for syncopating and tagging etyma and termination, and the tagging of a corpus using a small tagset are explained. Some practical methods solving problems in Uyghur language tagging are also proposed.

Key word: history and developmnet of Uyghur language, Uyghur tagged corpus, Uyghur language tagging system.

1 Introduction

Processing raw corpuses is fundamental to the research of the corpus linguistics. Since the first corpus was built in USA in 1970's, many other corpuses had been developed world wide. In the early days, a corpus usually contains only one million words and only text corpuses are collected. From 1980's, research in linguistic corpuses becomes very active. Some major projects at that time are: the Lancaster Treebank, the tagged LOB Corpus and the Birmingham Corpus. In the 1980's the corpuses are large and there are more varieties. Before 1990s, almost all are English corpuses.

From 1986 to 1994, several different part-of-speech (POS) automatic tagging tools such as CLAWS1, CLAWS2, CLAWS3 and CLAWS4 are developed. At the same time, many progresses had been made in the corpus syntax analysis and tagging. The PennTreebank, developed by the Pennsylvania University of USA combined and integrated many corpus processing tools such as: the Church POS tagging tools, the Fidditch syntax analysis kit of the Hindle system and etc. These tools greatly enhanced the efficiency of the corpus tagging.

After 1990s, many non-English corpuses are developed. These include the Japanese ER corpus and the news manuscript corpus of NHK.

In the past ten years, much progress has been made in the development of Chinese corpuses. These include: the automatic syncopating ([LNY87], [XHS91]), POS tagging ([BXH92], [BXH92]), dependency tagging ([ZH94], [LZH93]) and etc. From 1992, researchers in the Institute of Computational Linguistics in Peking University have been working in the integrated processing of the Chinese corpuses. They have developed many new technologies and tools for the Chinese corpus processing.

From the viewpoint of the sources of a language, a corpus can be divided into text or speech, spoken language or written language and single language or multi languages.

So far, many tagsets have been developed for POS tagging, phrase tagging, dependency relation tagging, case relationship, syntax tree, semantic relationship and so on. As much as we know, the beginning of building Uyghur corpus is in 2001. So there are far more works to be done for this language.

2 Tagging of the contemporary Uyghur corpus

The development of Uyghur alphabets has a long history. Uyghur means "combination" or "solidarity". Before the 9th century, Uyghur people used Turkic alphabets for writing the language. Another kind of letters used by the Uyghur was the ancient Uyghur letters. These ancient Uyghur letters had been widely used in the Xinjiang and Central Asian areas from 8th century to 13th century. Later, the Mongolian and Manchu letters were actually developed based on it. In the middle of the 19th century, after the Uyghur's conversion to the Islam, words based on the Arabia letters were used. After many years of improvements and modifications, the Arabic letters based written system becomes accepted by the Uyghurs. This system now consists of 32 letters, and follows a right to left writing rules.

Uyghur is the main ethnic group in the Xinjiang Uyghur Autonomous Region. It has a population of 8.34 millions (according to the Xinjiang Uyghur Autonomous Region 2000 census, pp46, Xinjiang People Publishing House 2002/10, Urumqi). Uyghur lives in almost every parts of Xinjiang, with the vast majority stays in the South Xinjiang, especially in Kashqar, Aqsu, Khotane. Uyghur language plays an important role in the development of the Uyghur culture. The heritages of Uyghur culture are recorded in Uyghur. Uyghur belongs to Altai phylum of the Turki Austronesian. In history, the development of the Uyghur had experienced three stages: ancient Turki (7th-13th), Chaghatai (14th-18th) and contemporary Uyghur(19th-todate).

The Uyghur has 8 vowels and 24 consonants. Morphologically, Uyghur belongs to the coherence language category. Uyghur is a very rich language and it has a large number of words. Besides words deriving from the expanding of the Turki paronym, Uyghur absorbes many words from other language including Chinese, Iranic, Arabic, Russian, Mongolic, Tibeta and Sanskrit, among which Chinese words were the earliest ones borrowed, and Iranic and Arabic words take up the largest parts of the borrowed words. There are almost no differences between words and syntax among different Uyghur dialects, except for their different pronunciations. So the Uyghur people from different places can communicate to each other without much difficulties.

The contemporary Uyghur spoken language has three dialects: the Central dialect, the Khotanese dialect and the Lop dialect. The Lop dialect has most distinct features. On the other hand, modern Uyghur written language bases mainly on the Central dialect. It is spoken by 80% of Uyghur people. Compared with other written language, Uyghur written language has its own distinct features:

- There are 32 letters in Uyghur. Uyghur words are delimited by blanks. There are many inflections in the Uyghur written language, especially the verbal inflections. So far we know that there are 10,000 verbal inflections and about 30 nunnations.
- Uyghur is a kind of *coherence* language. The meaning of etyma may be changed by adding a *termination* is added. Different nunnations represent different cases and numbers. For example, *Hesen* is a third person singular nominative noun. When it becomes *Hesenning* (adding the termination *ning*), it becomes a possessive. In some other cases, when a termination is added to a noun, it becomes a verb. For example, the noun *tash* (stone and soil) will be turned into a verb *taxla* (throw away) by adding termination *la*. This linguistic feature provides useful clues for POS tagging of the Uyghur written language.
- There are many *declensions* in Uyghur, for examples, the verb, adjective and noun. These declensions are helpful for part-of-speech tagging.
- Most Uyghur words have morphological changes, which represent different functions. For example, besides being using as a predicate or an adverbial modifier in most of cases, a verb can also be used as subject, object and attribute without any morphological changes. The

plural words also have no morphological changes, which makes the POS tagging more difficulty. Examples are:

U at minishni bilmeydu. (1) He cannot ride a horse.
U yengi tughulghan buwaqqa at qoydi. (2) His newborn baby was given a name.
Sen miltiq bilen dvshmenni at. (3) You use the gun to shoot the enemy.

These features make it difficulty, if not impossible to manually develop general rule sets for POS tagging. One of our objectives of the Uyghur POS tagging research is to extract syntactic rules from the corpus. And by combining the statistical information with the rule-based POS tagging system, we aim to develop an automatic processing software for the Uyghur language.

3 Sources of Corpuses

The sources we used in this research is an electronically text file of 8 millions words. It contains text published in the first six months "Xinjiang Daily" and additional text provided by the publisher of the Xinjiang Technology, Reference News, Xinjiang Education Publisher and Xinjiang People's Publisher. These materials cover all the words in the textbooks of our elementary schools, junior schools, high school, technical secondary school, university and the words in law, literature, health, technology (especially the agriculture technology), history, and economics.

4 Development of Uyghur corpus

4.1 Text collecting and text format transformation

The preliminary work of building a corpus is text collecting. We use the electronically compiled text as mentioned in Section 3. However, we have one problem here. These text resources are compiled by Beida Fengzhang Typeset System under DOS and can not be used directly under Windows system. We had to develop a text format transformation program to transfer the format to a suitable format. So far more than 8 millions words had been processed. Meantime, we have also compiled a Uyghur syntax information dictionary, which consists of the etyma and phrase dictionary and the termination dictionary. The etyma and phrase dictionary has of 60,000 words and termination dictionary has of 13,000 words. Based on these work, we developed a computer-aided contemporary Uyghur language corpus processing system.

4.2 Rules of the Uyghur language corpus

We begin our work by referencing to the Contemporary Uyghur Corpus Processing: Manual of Rules for Syncopating and Tagging Etyma and Termination", a document compiled by the Xinjiang Normal University. Details are:

1) Small tagset is used for POS tagging. These are:

Part of speech	Symbol	Part of speech in Uyghur
Noun	N	Isim
Verb	V	Pzh'zhl
Adjective	A	Svpet
Pronoun	R	Almash
Numeral	M	San
Adverb	D	Rewish
Quantifier	Q	Miqtar
onomatopoeia	Z	Teqlidi scz
exclamation	X	Vndesh
conjunction	C	Baghlighuchi
postposition	H	Tirkelme
tone	Y	Yvklime

The nouns are divided into the following sub categories,

Sub- Category of Noun(In English)	Symbol	Sub- Category of Noun(In Uyghur)
Time	Nw	Waqit Ismi
Location	No	Orun-terep Ismi
Abbreviation	Nq	Qisqartilma
Name of People	Nk	Kishi ismi
Name of place	Ny	Yer nami
Name of organization	Ng	Organ-teshkilatlar nami
Metal	N1	Mitallorgiye
Transportation	N2	Qatnash

Besides the 12 tags provided, we add in more than 50 new tags for different noun categories. For example: the Nk for the name of people, Ny for the name of place and Ng for the name of organization.

2) Proper nouns such as name of people, name of place and name of organization are tagged. If a proper noun belongs to the phrase category, then it is marked by a pair of squarė brackets.

4.3 The Rules

It contains the following 3 parts: cutting, tagging and combination.

4.3.1 The cutting rules

(1) The cutting sub-units

The so-called *participle* unit is defined in the Contemporary Uyghur Literary Language Standard Dictionary and the Contemporary Uyghur Detailed Solution Dictionary which are established by the Chinese Xinjiang Uyghur Autonomous Region Language Committee. It is the fundamental unit used in the information processing and has clearly defined semantics and grammar functions. In this research, the concept of the *participle* unit is still applied but we call it as the **cutting sub-unit** now.

Independent or *non-element* sentences can also appear during the cutting under some special situations. Some words form of stem and suffix are showed as below:

(Put out) in the verb in "[[eli/v+p(suffix)]/d chiq/v]/vp", elip (take) /d chiq (come out) /v (predicate verb); [kal/v+di]/vp (come) kal is the verb stem, di is the verb suffix; /v;
Leave-merge form of noun

The noun *Azherning* (personal name) can be divided into *Azgher/n* (stem) and *ning* (suffix). *Azher* is a nominative noun, singular, and third person. However, the word *Azherning* has become the possessive. Other attributes remain unchanged.

(2) The relationship between the cutting sub-units and their dictionary entries

Usually, the cutting units are obtained from the participle dictionary. By referencing to the Contemporary Uyghur Literary Language Standard Dictionary and Contemporary Uyghur Detailed Solution Dictionary, we compile a Participle Word Table. Up to date, more than 60,000 entries have been included. These are listed in our Contemporary Uyghur Grammar Information Dictionary(CUGID).

There are differences between the cutting sub-units and the dictionary entries. For example, the practices terminology, abbreviation, geographic, or foreigner's name are cutting sub-units. But these are not collected. The numeral and the time, such as " 2858 ", " beshinchi (fifth) ", " 1988-yili(1988 years) ", " 4-ay (in April) ", " 20-küni (on 20th) " are also cutting sub-units, but these are too many to be included. Though some *front ingredients* and *after ingredients* are cutting sub-units, we have also seen a number of out-of-dictionary words in real texts.

4.3.2 Tagging rules

(1) CUGID provides us with the basis of our lexical category tagset. It classifies 60,000 words and expressions. If a word or an expression is a cutting sub-unit in the CUGID and it has only one POS, tagging only copies the POS. If a word has many POS, the tagging will choose the most suitable one.

(2) The directing function of the phrase grammar system

In the lexical category tagging, a difficult grammar decision that we will have to make is the relationship between the POS and its syntactical function. In Uyghur language there is not simple one-to-one correspondence between these two. In general, we try not to decide POS based on the syntactical function. For example, if a word has already been defined as a verb in CUGID, we would not tag them to be noun only because they are used as a subject or object in a sentence.

(3) Proper noun tagging

We have tagged a large number of proper nouns, such as personal names, geographic names that appear in news articles. We also added square brackets and type marks to the phrase appropriation names based on the word cutting and lexical category tagging (mainly for Nw, Ng, but a few for Ny).

5. Combination of Rules and the Statistics

To make sure that words are correctly tagged, we combine rules with statistical information. The advantage of using rules is that it allows the use of existing linguistic knowledge. On the other hand, statistical frequency data can be obtained from corpuses easily. These data are far more uniform. We would also have better coverage of real text.

The initial tagging is done by rules to remove the semantic ambiguity. Statistical disambiguation is then applied to deduce the lexical category of those out-of-dictionary words. Manual proofreading is then applied to obtain the final tags.

This result has two uses. On one hand we may obtain the different parameters which will be useful in the statistical disambiguation. On the other hand, we can compare the result of machine tagging (rules and statistical disambiguation) with the result of manual tagging. Mistakes in machine tagging can then be identified. From here, we may obtain a lot of useful information to supplement and adjust the content of rule base.

Our problem is that we do not have a big enough tagged corpus to begin with. At the beginning, we tag the corpuses by rules. Then we apply manual tagging to discover and correct the mistakes from rules tagging. We then modify the rules. With the revised rules, the corpuses are tagged again. We repeat this process recursively until no further improvement can be made.

6. Conclusion

This paper reports our works on the POS tagging of the contemporary Uyghur. We have explained our processing strategy that basically, it is a way of combining the rule-based and statistical-based disambiguation with the additional effort of manual checking and tagging.

In our future research, we will attempt to process more corpuses with this method recursively. We aim to obtain accurate POS to enhance our language database in the next 2 years. In this way, we are able to provide better statistics for researchers in Uyghur Information Processing

*the major part of work described in this paper is performed by the first author from Xinjiang University.

References

[1] 俞士汶、朱学锋、王惠、张芸芸，《现代汉语语法信息词典详解》，北京：清华大学出版社，第 1 版，1998 年 4 月

[2] K T Lua, An Efficient Inductive Unsupervised Semantic Tagger, Computer Processing of Oriental Languages, Vol 11, No 1, 1997, 35-47

[3] 玉素甫.艾白都拉 ，吾守尔.斯拉木，“维吾中心语驱动文法句法分析器中的上下文相关处理”，《计算机应用与软件》，1999/6。

[4] 玉素甫.艾白都拉，吾守尔.斯拉木“维吾尔语词法分析器成功”，<<中文信息>>，1997/4。

[5] Hearst Maarti.Noun Homograph Disambiguation Using Local Context in Large Corpora, in proceedings,ARPA Human Language Technology Workshop,1993

[6] Yarowsky D.Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In Proceedings of 33rd Annual Meeting of ACL, Cambridge, Massachusetts,USA,1995

[7] Lua, K T and K W Gan, Application of information theory binding in word segmentation. Computer Processing of Chinese & Oriental Languages, vol.8, no.1 (June 1994): 115-124

[8] 哈米提.铁木尔，“现代维语语法”（维吾尔文），民族出版社出版，1987 年 6 月

[9] 程适良等编，“现代维吾尔语语法”（维吾尔文），新疆人民出版社出版，1996 年 6 月

[10]俞士汶主编，《现代汉语语料库加工——词语切分与词性标注规范与手册》，北京大学计算语言学研究所，1999 年 4 月

[11] Lua, K T, Frequency-rank curves and entropy for Chinese characters and words. Computer Processing of Chinese & Oriental Languages, vol.8, no.1 (June 1994): 37-52. (United States).

[12]周强、段慧明，现代汉语语料库加工中的切词与词性标注处理，《中国计算机报》，1994 年 5 月 31 日，第 85 版

[13]周强、张伟、俞士汶，树库的构建，《中文信息学报》，1997 年第 4 期，42-51

[14]中国国家标准 GB13715《信息处理用现代汉语分词规范》，见刘源等著《信息处理用现代汉语分词规范及自动分词方法》，北京：清华大学出版社，第 1 版，1994 年

[15]白栓虎等，汉语语料库词性标注方法研究，见陈肇雄主编《机器翻译研究进展》，408-418，北京：电子工业出版社，1992 年

[16]刘开瑛等，语料库词类自动标注算法研究，见陈肇雄主编《机器翻译研究进展》，378-386，北京：电子工业出版社，1992 年

[17]孙茂松，《信息处理用现代汉语分词词表》的设计原则，见黄昌宁、董振东主编《计算语言学文集》，193-198，北京：清华大学出版社，1999

[18]朱学锋、俞士汶、王惠，现代汉语 5 万词语归类的实践，《语言文字应用》，1997 年第 4 期，88-94

[19]陆志韦等，《汉语的构词法》，科学出版社，1964 年

[20]朱德熙，《语法答问》，北京：商务印书馆，1985 年

[21]周强、俞士汶，一种切词和词性标注相融合的汉语语料库多级加工方法，见陈力为主编《计算语言学研究与应用》，126—131，北京：北京语言学院出版社，1993

[22] 哈米提.铁木尔，“现代维语语法”（维吾尔文），民族出版社出版，1987 年 6 月

[23] 程适良等编，“现代维吾尔语语法”（维吾尔文），新疆人民出版社出版，1996 年 6 月

[24] 吴蔚天，罗建林，“汉语计算语言学”，电子工业出版社，1994 年 7 月。

[25] 玉素甫.艾白都拉 “维语句法分析器中的词义排歧问题的研究”，《计算机应用与软件》，2002/4。

This article is sponsored by: 60163002, National Natural Science Foundation of China (NSFC), The Chinese & Oriental Languages Information Processing Society, Singapore (COLIPS), China Scholarship Council (CSC) and The Startup Fund of the Special Training Program for the Talent of Minority by the Science Department of Xinjiang Uyghur Autonomous Region.