Survey of Natural Language Processing for Vertical Mongolian: Current Situation

Sargai Yunshaab
November 2019

Prepared for the project “Named-Entity Recognition in Tibetan and Mongolian Newspapers”
For the Mongolian and Inner Asian Studies Unit, Cambridge University
Funded by: Cambridge Language Sciences, Cambridge University
Project PI: Dr Hildegard Diemberger; Coordinator: Dr Robert Barnett

Contents

1. The state of Natural Language Processing for vertical Mongolian

Mongolian is a language spoken, written and read by about 5.5m people, of whom some 3m live in Mongolia, a democratic, independent state in eastern Inner Asia, while about 2.5m live in the “Inner Mongolian Autonomous Region” (IMAR), a province-level administrative area in the People’s Republic of China. Mongolian is agglutinative, with suffix chains in the verbal and nominal domains, and has subject–object–predicate as its basic word order. In Mongolia, the Cyrillic script is used for writing Mongolian, whereas in the IMAR an older script is used, which is written vertically. Also known as the *Hudum Mongol bichig*, this script is derived from the Old Uyghur script and is written in vertical lines top-down, right to left. For foreign-language or mixed-language texts such as this report, the script is often printed horizontally.

Vertical Mongolian natural language processing research started in the 1980’s, considerably later than many other languages. In 30 years of development, the basic corpus, grammatical information dictionary, editing and typesetting systems, and common office software have been developed and put into use.[1] Researchers at Inner Mongolia University in Hohhot, the capital of the IMAR, have completed phrase boundary recognition and phrase-structure relationship determination,[2] as well as shallow parsing, such as predicate automatic identification.[3] In terms of code conversion, S. Loglo proposed a code conversion method based on dictionaries and rules, which converts non-national standard code to international standard code, with a conversion accuracy of 98.19%.[4] In terms of spelling proofreading (spellchecks), Hua Shabao has implemented a rule-based spelling error discrimination system, which can distinguish root errors, additional component errors and vowel collocation errors without correction.[5] S. Loglo further proposed a Mongolian proofreading method based on uncertain finite

[1] S. Loglo, "Design and realization of a graphical editing system for Mongolian dependency treebank, " *2016 International Conference on Asian Language Processing (IALP)*, Tainan, pp. 176-179, 2016.
[2] Hua Shabao, Dabhurbayar, "A phrase-tagging research in Mongolian corpus", *Journal of the Central University for Nationalities (Philosophy and Social Sciences Edition)*, vol. 33, no. 5, pp. 64-67, 2006.
[3] Serguleng Wang, D. Sarana, Nasunurtu, "Design and realization of automatic annotation for modern Mongolian predicate segment", *Proceedings of 11th national symposium on minority languages*, pp. 420-427, 2007.
[4] S. Loglo, "Research on the General Algorithm of Mongolian Code Conversion [J]." *Journal of Inner Mongolia University (Philosophy and Social Sciences)* 2 (2009).
[5] Hua S. "Modern Mongolian automatic proofreading system–MHAHP. J. Inner Mongolia Univ." *Philos. Soc. Sci. Ed* 4 (1997): 49-53.

automata.[6] The method uses finite state automata to organize Mongolian dictionaries, improves the operation speed of the algorithm and corrects the Mongolian words with correct font but wrong pronunciation. However, manual intervention is needed to select the correct pronunciation. In terms of morphological processing of Mongolian, Nashunuritu combined the characteristics of Mongolian word-endings to form a dictionary and used rules to segment suffixes in Mongolian.[7] Hou Hongxu et al. put forward a Mongolian word segmentation method combining statistics and rules.[8] This method uses a Mongolian statistical language model as the basis for disambiguation of multiple candidate results, effectively improving the accuracy of segmentation. Bao et al. used the method of affix segmentation to improve the recognition rate of large-scale Mongolian speech recognition.[9] Zhao Jiandong et al. added a lookahead mechanism to the historical model, which improved the accuracy of part-of-speech tagging in Mongolian.[10]

The main work carried out on Named Entity Recognition (NER) for Mongolian has been as follows: Tonglaga analyzed personal names in Mongolian and their national, temporal and regional characteristics, along with the changing rules governing the internal model regarding people's names, which he used to build a corpus of personal names.[11] A personal name recognition system based on the maximum entropy model was realized, obtaining an F value=89.61 in 1,040 open name tests. Using the conditional random field model, Cai Jingjing constructed seven sets of different feature templates and ran a test of human name recognition, obtaining an F value of 92.64 in a closed test.[12] Wu Jinxing et al. developed a method for place-name recognition by combining the conditional random field method with use of a dictionary look-up for analyzing the composition of place names in Mongolian.[13]

However, the limited available resources and lack of a public annotated corpus have restricted progress in research and development. In addition, Mongolian's word formation mode is different from those of Chinese and English, and there are some problems, such as inconsistency in coding, which have held back the development of Mongolian NER and limited further research on text understanding, machine translation, public opinion analysis, knowledge map construction, intelligent Q & A, etc.[14]

2. Problems facing Mongolian NER

In order to investigate the current state of vertical Mongolian NER, interviews were carried out in Inner Mongolia, China between August and October in 2019. The interviewees were from Mongolian information-processing teams at the Inner Mongolia University, the Inner Mongolia Publishing House, the Inner Mongolia Daily, and leading Mongolian information technology companies. I collected literature before and after the interviews to support my findings, and drew, for the majority of this report, on Chinese-language papers which I have summarised and translated into English (these are cited in the footnotes). These sources indicated the following difficulties facing NER for vertical Mongolian:

[6] Loglo, S. "A Proofreading Algorithm of Mongolian Text Based on Nondeterministic Finite Automaton." *Journal of Chinese Information Processing* 23.6 (2009): 110-115.

[7] Nashunwuritu, "An Automatic Words Segmentation system of the Mongolian root and stem", *Journal of Inner Mongolia University*, vol. 29, no. 2, pp. 53-67, 1997.

[8] Hou, Hongxu, et al. "Mongolian word segmentation based on statistical language model." *Pattern recognition and artificial intelligence* 22.1 (2009): 108-112.

[9] Bao F., Gao G., Yan X., et al. Segmentation-based Mongolian LVCSR approach[C]. In: Proceedings of Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013: 8136-8139.

[10] ZHAO, Jian-dong, Guang-lai GAO, and Fei-long BAO. "Research on History-based Mongolian Automatic POS Tagging." *Journal of Chinese Information Processing* 27.5 (2013).

[11] 通拉嘎. 基于蒙古文语料库的人名自动识别[D]. 北京: 中央民族大学, 2013.

[12] 才晶晶. 基于 CRF 的蒙古文人名自动识别[D].呼和浩特: 内蒙古大学, 2016.

[13] 吴金星, 丽丽, 杨振新.CRF 和词典相结合的蒙古文地名识别研究[J].计算机工程与科学, 2016, 38(05):1046-1051.

[14] Wang W., “Research on Mongolian name entity recognition” (蒙古文命名实体识别研究), Ph.D. thesis, 2018, Inner Mongolia University (in Chinese).

- There isn't a large enough vertical Mongolian corpus of named entities, coding is not uniform or standardized, and there are many wrong words in the existing corpora, such as those with an incorrect font (glyph) or the wrong pronunciation. Compared with English, Chinese and other developed languages, the information technology foundation for vertical Mongolian is quite basic, and depends upon a named entity corpus that is too small, leading to difficulties for research into NER. For a long time, there have been different coding methods used in Mongolian, and this has made it more difficult to collect and organize a corpus. The typography used in the corpus also brings difficulties for compiling statistics and analysing Mongolian corpora[15].
- Vertical Mongolian has no capital letters, unlike the Latin alphabet. This makes it more difficult to judge the boundary of Mongolian named entities, unlike many languages where the initial capital is a clear boundary marker for a named entity. For example, the English or German name "Einstein" is written in traditional Mongolian as "ᠡᠶᠢᠨᠱᠲᠠᠶᠢᠨ", but this does not have an initial capital, making it difficult to judge its boundary.[16]
- The Mongolian lexicon is large because of the characteristics of Mongolian word formation. Unlike English and Chinese, Mongolian words are completed by affixing suffixes, and this process generates different words from the same stem. This method of word formation leads to a large vocabulary in Mongolian, which increases the number of unknown words in the corpus and makes it difficult to train and test the model.[17]
- The grammar order in Mongolian, subject+object+predicate, makes it difficult to recognise the boundaries of a named entity, as a named entity can act as subject or as an object in any sentence.
- The forms in which named entities appear in of Mongolian are changeable and varied. For example, there are clear differences between Mongolian names and Chinese names because of their different characteristics. Chinese names written in Mongolian usually have the same number of words as the number of syllables in the equivalent Chinese names. For example, the Chinese name 李纪恒 (Li Jiheng) is "ᠯᠢ ᠵᠢ ᠾᠧᠩ", which is composed of three words, whereas the Mongolian personal name agula "ᠠᠭᠤᠯᠠ" is one word. Another difference is that some Mongolian names use the patronym or father's name/surname, such as "ᠳ · ᠪᠠᠲᠤᠮᠥᠩᠬᠡᠴᠢᠯᠠᠭᠤ".[18]
- In Mongolian, the phenomenon of compound words is very common. As in other languages, some common nouns, adjectives and even verbs in Mongolian may act as personal names, and the occurrence of compound words is more frequent. For example, the adjective "ᠤᠬᠠᠭᠠᠲᠠᠢ" (meaning "smart") is a common Mongolian name. It is impossible to determine its category by depending only on the word itself, so it must be judged according to context.[19]
- Another problem is that new entity names keep popping up. Like other languages, Mongolian is a dynamic language. It creates new words constantly and absorbs words from other languages to realize its own development. These new words bring more named entities into the lexicon. For example, "ᠬᠣᠷᠢᠨᠭᠡᠷ ᠰᠢᠨ᠎ᠡ ᠲᠣᠭᠣᠷᠢᠭ" (Helingeer New District) is a new place name, which cannot be found in the current corpus. This can bring difficulties for classification.[20]

[15] Wang W., "Research on Mongolian name entity recognition" (蒙古文命名实体识别研究), Ph.D. thesis,2018,Inner Mongolia University (in Chinese).

[16] *Ibid.*

[17] *Ibid.*

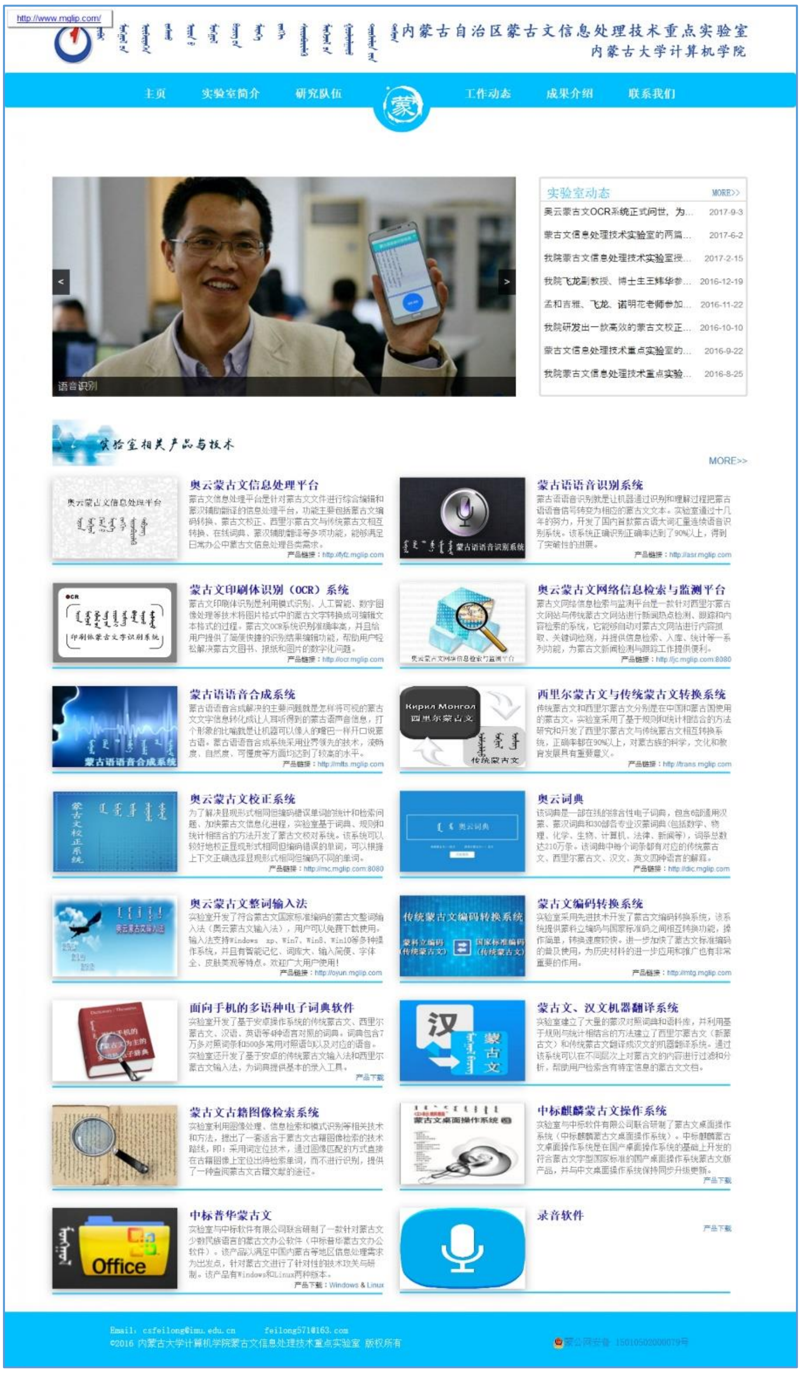
[18] *Ibid.*

[19] *Ibid.*

[20] *Ibid.*

3. Currently available products developed by the Inner Mongolia University Computer Science Lab

The Computer Science Lab at the Inner Mongolia University in Hohhot offers a number of directly related products and tools. This information below reproduces the claims made by the Computer Science Lab on its websites about its products (see Figures 1 and 2).



http://www.mglip.com/

内蒙古自治区蒙古文信息处理技术重点实验室
内蒙古大学计算机学院

主页 实验室简介 研究队伍 蒙 工作动态 成果介绍 联系我们

语音识别

实验室动态 MORE>>

奥云蒙古文OCR系统正式问世，为...	2017-9-3
蒙古文信息处理技术实验室的两篇...	2017-6-2
我院蒙古文信息处理技术实验室授...	2017-2-15
我院飞龙副教授、博士生王炜华参...	2016-12-19
孟和吉雅、飞龙、诺明花老师参加...	2016-11-22
我院研发出一款高效的蒙古文校正...	2016-10-10
蒙古文信息处理技术重点实验室的...	2016-9-22
我院蒙古文信息处理技术重点实验...	2016-8-25

实验室相关产品与技术

MORE>>

奥云蒙古文信息处理平台
蒙古文信息处理平台是针对蒙古文文件进行综合编辑和蒙汉辅助翻译的信息处理平台，功能主要包括蒙古文编码转换、蒙古文校正、西里尔蒙古文与传统蒙古文相互转换、在线词典、蒙汉辅助翻译等多项功能，能够满足日常办公中蒙古文信息处理各类需求。
产品链接：http://fyfz.mglip.com

蒙古语语音识别系统
蒙古语语音识别就是让机器通过识别和理解过程把蒙古语语音信号转变为相应的蒙古文文本。实验室通过十几年的努力，开发了国内首款蒙古语大词汇量连续语音识别系统。该系统正确识别正确率达到了90%以上，得到了突破性的进展。
产品链接：http://asr.mglip.com

蒙古文印刷体识别（OCR）系统
蒙古文印刷体识别是利用模式识别、人工智能、数字图像处理等技术将图片格式中的蒙古文字转换成可编辑文本格式的过程。蒙古文OCR系统识别准确率高，并且给用户提供了简便快捷的识别结果编辑功能，帮助用户轻松解决蒙古文图书、报纸和图片的数字化问题。
产品链接：http://ocr.mglip.com

奥云蒙古文网络信息检索与监测平台
蒙古文网络信息检索与监测平台是一款针对西里尔蒙古文网站与传统蒙古文网站进行新闻热点检测、跟踪和内容检索的系统，它能够自动对蒙古文网站进行内容抓取、关键词检测，并提供信息检索、入库、统计等一系列功能，为蒙古文新闻检测与跟踪工作提供便利。
产品链接：http://jc.mglip.com:8080

蒙古语语音合成系统
蒙古语语音合成解决的主要问题就是怎样将可视的蒙古文文字信息转化成让人耳听得到的蒙古语声音信息，打个形象的比喻就是让机器可以像人的嘴巴一样开口说蒙古语。蒙古语语音合成系统采用业界领先的技术，流畅度、自然度、可懂度等方面均达到了较高的水平。
产品链接：http://mtts.mglip.com

西里尔蒙古文与传统蒙古文转换系统
传统蒙古文和西里尔蒙古文分别是在中国和蒙古国使用的蒙古文。实验室采用了基于规则和统计相结合的方法研究和开发了西里尔蒙古文与传统蒙古文相互转换系统，正确率都在90%以上，对蒙古族的科学，文化和教育发展具有重要意义。
产品链接：http://trans.mglip.com

奥云蒙古文校正系统
为了解决显现形式相同但编码错误单词的统计和检索问题、加快蒙古文信息化进程，实验室基于词典、规则和统计相结合的方法开发了蒙古文校对系统。该系统可以较好地校正显现形式相同但编码错误的单词，可以根据上下文正确选择显现形式相同但编码不同的单词。
产品链接：http://mc.mglip.com:8080

奥云词典
该词典是一部在线的综合性电子词典，包含6部通用汉蒙、蒙汉词典和30部各专业汉蒙词典(包括数学、物理、化学、生物、计算机、法律、新闻等)，词条总数达210万条。该词典中每个词条都有对应的传统蒙古文、西里尔蒙古文、汉文、英文四种语言的解释。
产品链接：http://dic.mglip.com

奥云蒙古文整词输入法
实验室开发了符合蒙古文国家标准编码的蒙古文整词输入法（奥云蒙古文输入法），用户可以免费下载使用。输入法支持Windows xp、Win7、Win8、Win10等多种操作系统，并且有智能记忆、词库大、输入简便、字体全、皮肤美观等特点。欢迎广大用户使用！
产品链接：http://oyun.mglip.com

蒙古文编码转换系统
实验室采用先进技术开发了蒙古文编码转换系统，该系统提供蒙科立编码与国家标准码之间相互转换功能，操作简单，转换速度较快。进一步加快了蒙古文标准编码的普及使用，为历史材料的进一步应用和推广也有非常重要的作用。
产品链接：http://mtg.mglip.com

面向手机的多语种电子词典软件
实验室开发了基于安卓操作系统的传统蒙古文、西里尔蒙古文、汉语、英语等4种语言对照的词典。词典包含7万多对照词条和500多常用对照语句以及对应的语音。实验室还开发了基于安卓的传统蒙古文输入法和西里尔蒙古文输入法，为词典提供基本的录入工具。
产品下载

蒙古文、汉文机器翻译系统
实验室建立了大量的蒙汉对照词典和语料库，并利用基于规则与统计相结合的方法建立了西里尔蒙古文（新蒙古文）和传统蒙古文翻译成汉文的机器翻译系统。通过该系统可以在不同层次上对蒙古文的内容进行过滤和分析，帮助用户检索含有特定信息的蒙古文文档。

蒙古文古籍图像检索系统
实验室利用图像处理、信息检索和模式识别等相关技术和方法，提出了一套适合于蒙古文古籍图像检索的技术路线，即：采用词定位技术，通过图像匹配的方式直接在古籍图像上定位出待检索单词，而不进行识别，提供了一种查阅蒙古文古籍文献的途径。

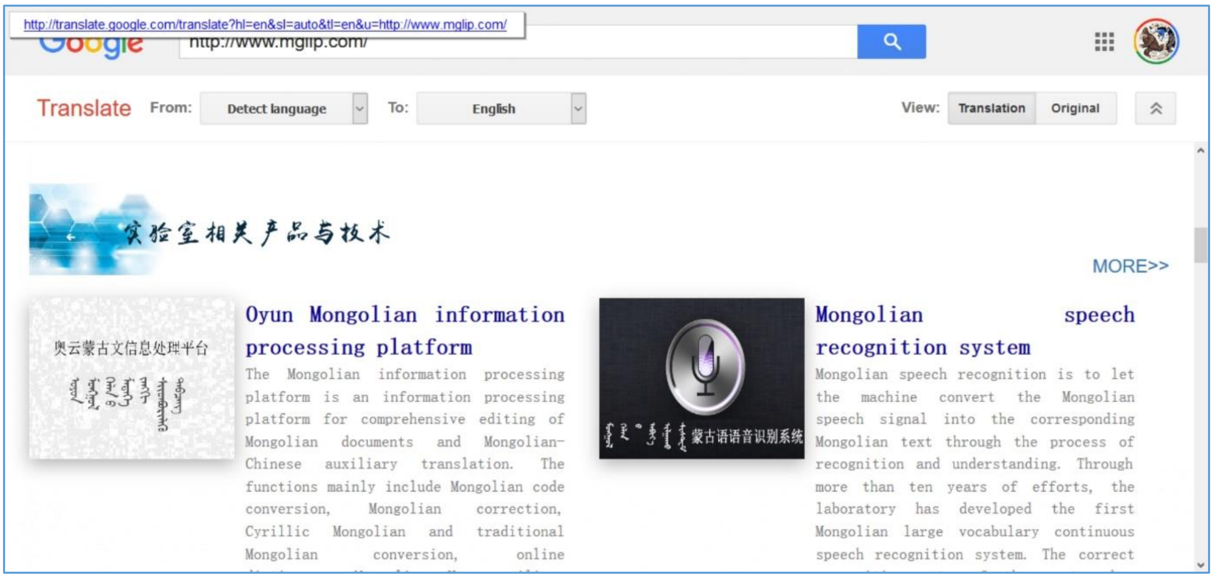
中标麒麟蒙古文操作系统
实验室与中标软件有限公司联合研制了蒙古文桌面操作系统（中标麒麟蒙古文桌面操作系统）。中标麒麟蒙古文桌面操作系统是在国产桌面操作系统的基础上开发的符合蒙古文字型国际标准的国产桌面操作系统蒙古文版产品，并与中文桌面操作系统保持同步升级更新。
产品下载

中标普华蒙古文
实验室与中标软件有限公司联合研制了一款针对蒙古文少数民族语言的蒙古文办公软件（中标普华蒙古文办公软件）。该产品以满足中国内蒙古等地区信息处理需求为出发点，针对蒙古文进行了针对性的技术攻关与研制。该产品有Windows和Linux两种版本。
产品下载：Windows & Linux

录音软件
产品下载

Email：csfeilong@imu.edu.cn feilong571@163.com
©2016 内蒙古大学计算机学院蒙古文信息处理技术重点实验室 版权所有
蒙公网安备 15010502000079号

Figure 1: Product page of the Computer Science Lab at Inner Mongolia University (October 2019)



Oyun Mongolian information processing platform

The Mongolian information processing platform is an information processing platform for comprehensive editing of Mongolian documents and Mongolian-Chinese auxiliary translation. The functions mainly include Mongolian code conversion, Mongolian correction, Cyrillic Mongolian and traditional Mongolian conversion, online

Mongolian speech recognition system

Mongolian speech recognition is to let the machine convert the Mongolian speech signal into the corresponding Mongolian text through the process of recognition and understanding. Through more than ten years of efforts, the laboratory has developed the first Mongolian large vocabulary continuous speech recognition system. The correct

Figure 2: Google Translation of two items from the Computer Science Lab's product page

Almost all of these tools are inaccessible to outsiders – they can only be accessed by members of the university – and the site, apart from the home page, appears to be no longer accessible, so I was unable to assess the claims by the Science Lab about most of their products

- The Mongolian Information Processing Platform. This is an information processing platform for comprehensive editing and for auxiliary translation of documents from Mongolian to Chinese. Its functions mainly include Mongolian code conversion, Mongolian correction, conversion between Cyrillic Mongolian and traditional (vertical) Mongolian, an online dictionary, Mongolian-Chinese auxiliary translation, and so on. According to the site, this can meet most of the document information processing needs in Inner Mongolia.
- The Mongolian Print Recognition (OCR) System.[21] According to the site, Mongolian print recognition is a process of transforming Mongolian characters in graphic formats into editable text format by using pattern recognition, artificial intelligence, digital image processing and other technologies. The site claims that recognition accuracy of the Mongolian OCR system is high, and it provides users with a simple and fast editing function in terms of recognition results. This helps users easily solve the digital problems of Mongolian books, newspapers and pictures.
- The Mongolian Speech Synthesis System. The main problem of Mongolian speech synthesis is how to transform visual Mongolian text information into spoken Mongolian, i.e., converting it into data or information that can be heard by the human ear. The Mongolian Speech Synthesis System adopts industry-leading technology, with what is described as a high level of fluency, naturalness and intelligibility.
- The Aoyun Correction System for Mongolian. In order to solve the problem of statistics and retrieval of words with the same appearance but wrong coding, and to speed up the process of developing Mongolian information technology, the laboratory developed a Mongolian proofreading system based on a combination of dictionary look-up, rules and statistics. The system can better correct the words with the same visual form but the wrong coding, and correctly select words with the same visual form but different coding, based on analysis of context.
- The Mongolian Speech Recognition System. Speech recognition in Mongolian aims to let machines transform Mongolian speech signals into the corresponding Mongolian text

[21] H. Zhang, H. Wei, F. Bao and G. Gao, "Segmentation-Free Printed Traditional Mongolian OCR Using Sequence to Sequence with Attention Model," *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Kyoto, 2017, pp. 585-590.

through the process of recognition and understanding. As a result of over ten years' efforts, the laboratory has developed the first large-vocabulary continuous speech recognition system for Mongolian in China. The correct recognition rate of the system is said to be more than 90%.

- The Aoyun Mongolian Network Information Retrieval and Monitoring Platform. The platform is a system for news "hot spot" detection, tracking, and content retrieval for Cyrillic Mongolian websites and traditional Mongolian websites. Reportedly, it can automatically grab content, carry out keyword detection on Mongolian websites, and provide a series of functions such as information retrieval, warehousing, statistics, etc., which provides what is described as a convenient service for Mongolian news detection and tracking.
- The System for conversion between Cyrillic Mongolian and traditional Mongolian. The laboratory adopts the method of combining rules and statistics to research and develop a mutual conversion system between Cyrillic Mongolian and traditional Mongolian, with a claimed accuracy of more than 90%. It is claimed to be of great significance for the development of Mongolian science, culture and education.
- The Mongolian Code Conversion System. The laboratory has developed a Mongolian code conversion system using advanced technology. The system provides a mutual conversion capability between Mongolian code and national standard code, which is simple in operation and fast in conversion. It has further accelerated the popularization and use of Mongolian standard coding. This is said to play a very important role in the further application and promotion of historical materials.
- The Mongolian and Chinese machine translation system. The laboratory has established a large number of Mongolian-Chinese dictionaries and corpora. It has also established a machine translation system for Cyrillic Mongolian and traditional Mongolian into Chinese using a combination of rules and statistics. Through this system, content in Mongolian can be filtered and analysed at different levels to help users retrieve Mongolian documents with specific information.

4. Findings from testing the Cyrillic to Vertical Mongolian converter

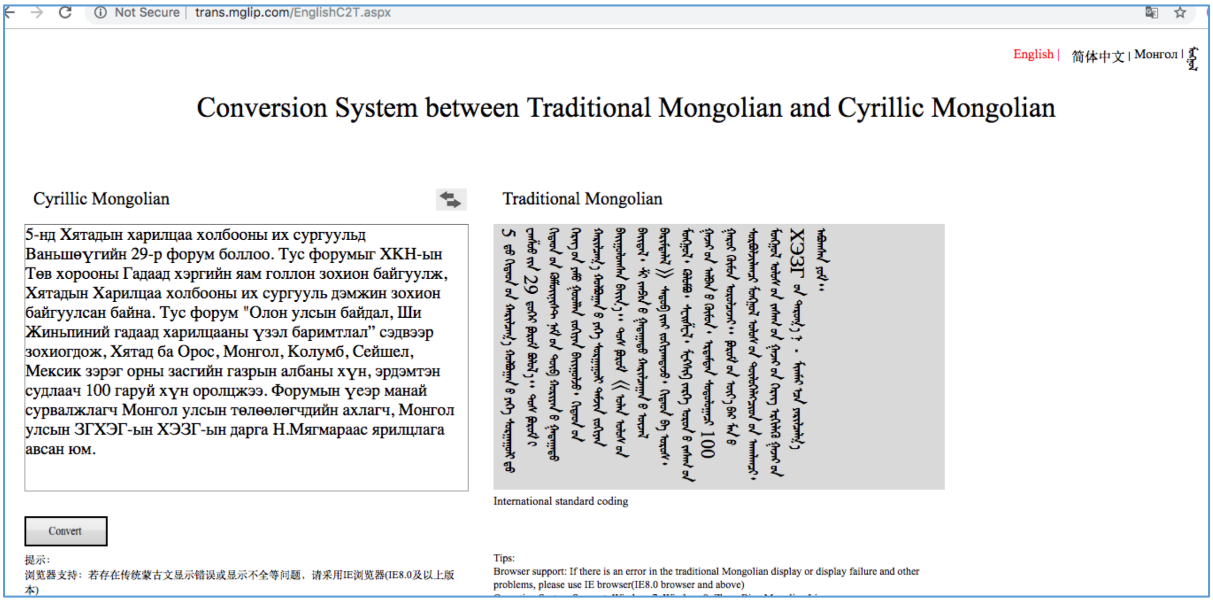
I was able to access and obtain one tool, the System for conversion between Cyrillic Mongolian and traditional Mongolian. As noted above, vertical (traditional) Mongolian and Cyrillic Mongolian are Mongolian scripts used respectively in China and Mongolia. However, a large portion of Cyrillic Mongolian words can be written in more than one way in vertical Mongolian, which makes conversion from Cyrillic Mongolian to vertical Mongolian difficult. To overcome this difficulty, Bao F et al.'s paper proposed a language model-based approach, which takes advantage of contextual information. Experimental results show that, for Cyrillic Mongolian words that have multiple correspondences in traditional Mongolian, the accuracy rate for this approach reaches 87.66%, thereby greatly improving overall system performance.[22] In this pilot project, our team tested the online converter extensively, as shown in Figure 3.

We found the following problems in the conversion results:

- When Cyrillic is written in all capital letters, the converter doesn't recognize the words.
- The converter doesn't recognize text abbreviations well.
- The converter sometimes does not recognize English letters that are used in Cyrillic Mongolian (e.g. km, kg).
- The converter makes many mistakes when it comes to prepositions (connecting words).

[22] Bao F., Gao G., Yan X., Wang H. (2013) "Language Model for Cyrillic Mongolian to Traditional Mongolian Conversion." In: Zhou G., Li J., Zhao D., Feng Y. (eds), *Natural Language Processing and Chinese Computing. Communications in Computer and Information Science*, vol 400. Springer, Berlin, Heidelberg.

- The converter makes mistakes when there are foreign names since the foreign names don't exist in Mongolian vocabulary lists.
- The converter sometimes transliterates Cyrillic words inaccurately, mistaking it for a word with a similar pronunciation but a different meaning (and sometimes such words are obsolete or do not make sense).


Not Secure | trans.mglip.com/EnglishC2T.aspx
English | 简体中文 | Монгол |

Conversion System between Traditional Mongolian and Cyrillic Mongolian

Cyrillic Mongolian

5-нд Хятадын харилцаа холбооны их сургуульд Ваньшөүгийн 29-р форум боллоо. Тус форумыг ХКН-ын Төв хорооны Гадаад хэргийн яам голлон зохион байгуулж, Хятадын Харилцаа холбооны их сургууль дэмжин зохион байгуулсан байна. Тус форум "Олон улсын байдал, Ши Жиньпиний гадаад харилцааны үзэл баримтлал" сэдвээр зохиогдож, Хятад ба Орос, Монгол, Колумб, Сейшел, Мексик зэрэг орны засгийн газрын албаны хүн, эрдэмтэн судлаач 100 гаруй хүн оролцжээ. Форумын үеэр манай сурвалжлагч Монгол улсын төлөөлөгчдийн ахлагч, Монгол улсын ЗГХЭГ-ын ХЭЗГ-ын дарга Н.Мягмараас ярилцлага авсан юм.

提示：
浏览器支持：若存在传统蒙古文显示错误或显示不全等问题，请采用IE浏览器(IE8.0及以上版本)

Traditional Mongolian

International standard coding

Tips:
Browser support: If there is an error in the traditional Mongolian display or display failure and other problems, please use IE browser(IE8.0 browser and above)

Figure 3: An example of a test result of conversion from Cyrillic to vertical Mongolian script

5. Findings from testing the Mongolian Print Recognition (OCR) system

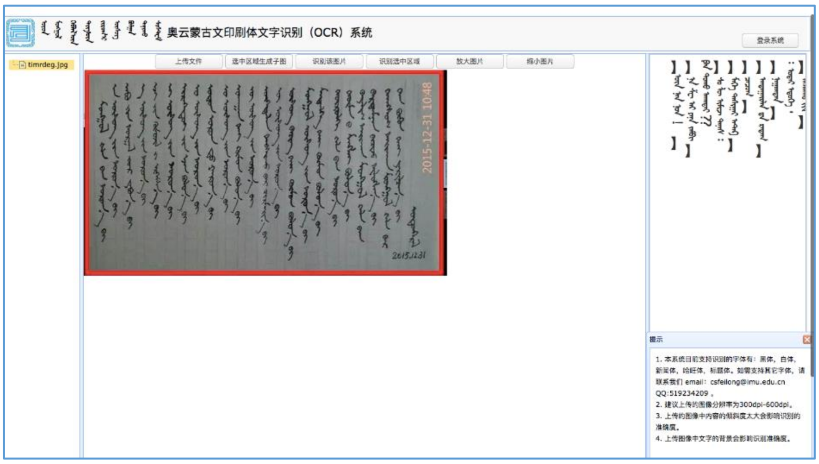
I was also able to run a test of the Mongolian Print Recognition (OCR) System. The tests showed some difficulties for the system in recognizing images of vertical Mongolian script:

- The OCR system only recognizes certain types of fonts (Mongolian Black, Mongolian White, Mongolian Title, Mongolian HaWang, Mongolian Sonin).
- The degree of tilt of the content in the uploaded image affects the accuracy of recognition.
- The background of the image affects the accuracy of recognition.
- The OCR system doesn't recognize hand-written documents (see Figure), including calligraphed texts (Figure).
- The OCR makes mistakes on the same word many times even when the source image is very clear (Figure 6, Figure).
- The OCR system makes mistakes when the contents are written in different fonts in the same image, even for the Mongolian White font and Mongolian Title font (Figure).

Overall, from the tests I ran, the accuracy of the OCR system appears to be relatively low, at around 40% or lower.

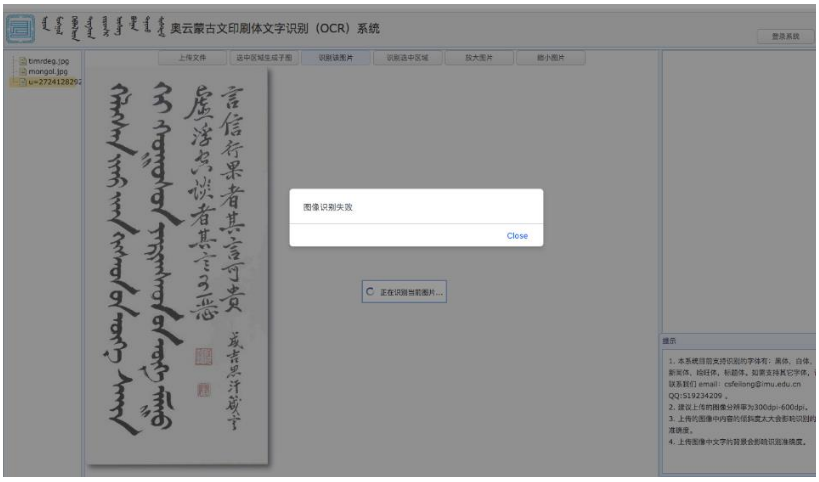
6. Conclusion

NLP is well developed for Cyrillic Mongolian script, but Vertical Mongolian script presents many difficulties for NLP tool developers. So far these difficulties cannot be fully resolved by using a Cyrillic-Vertical Script converter, as the conversion systems still need further development. A large number of tools for vertical Mongolian script have been developed at Inner Mongolia University, but for the most part these appear not to be accessible to those outside the university or outside the country. More research is needed to establish whether any of these tools could be made available for public use.



1. 本系统目前支持识别的字体有：黑体，白体，新闻体，哈旺体，标题体。如需支持其它字体，请联系我们 email：csfeilong@imu.edu.cn QQ:519234209 。
2. 建议上传的图像分辨率为300dpi-600dpi。
3. 上传的图像中内容的倾斜度太大会影响识别的准确度。
4. 上传图像中文字的背景会影响识别准确度。

Figure 4: An example of an OCR test result of a hand-written document



图像识别失败

正在识别当前图片...

1. 本系统目前支持识别的字体有：黑体，白体，新闻体，哈旺体，标题体。如需支持其它字体，请联系我们 email：csfeilong@imu.edu.cn QQ:519234209 。
2. 建议上传的图像分辨率为300dpi-600dpi。
3. 上传的图像中内容的倾斜度太大会影响识别的准确度。
4. 上传图像中文字的背景会影响识别准确度。

Figure 5: An example of an OCR test result of a calligraphy document

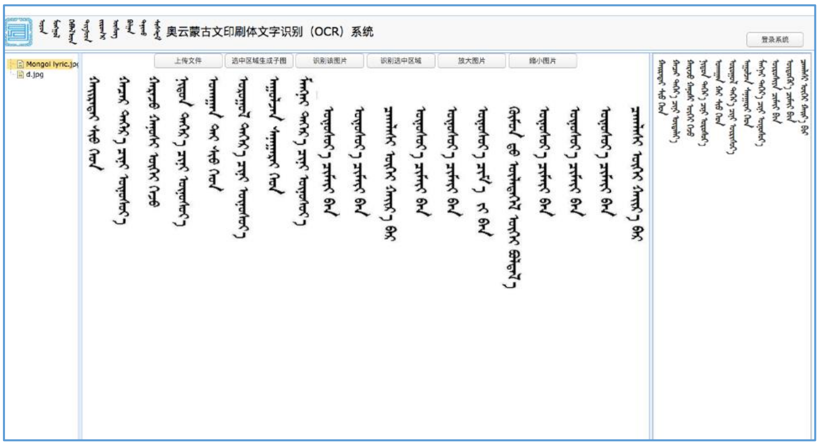
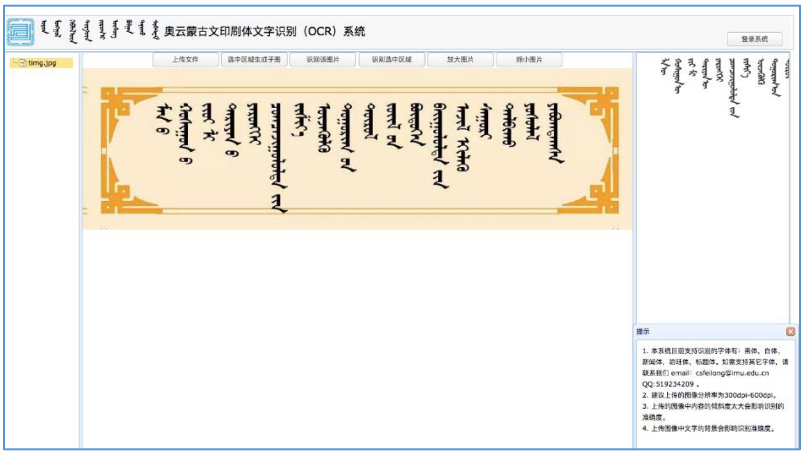


Figure 6: An example of an OCR test result of an image containing clear lyrics



1. 本系统目前支持识别的字体有：黑体，白体，新闻体，哈旺体，标题体。如需支持其它字体，请联系我们 email：csfeilong@imu.edu.cn QQ:519234209 。
2. 建议上传的图像分辨率为300dpi-600dpi。
3. 上传的图像中内容的倾斜度太大会影响识别的准确度。
4. 上传图像中文字的背景会影响识别准确度。

Figure 7: An example of an OCR test result of an image containing a slogan

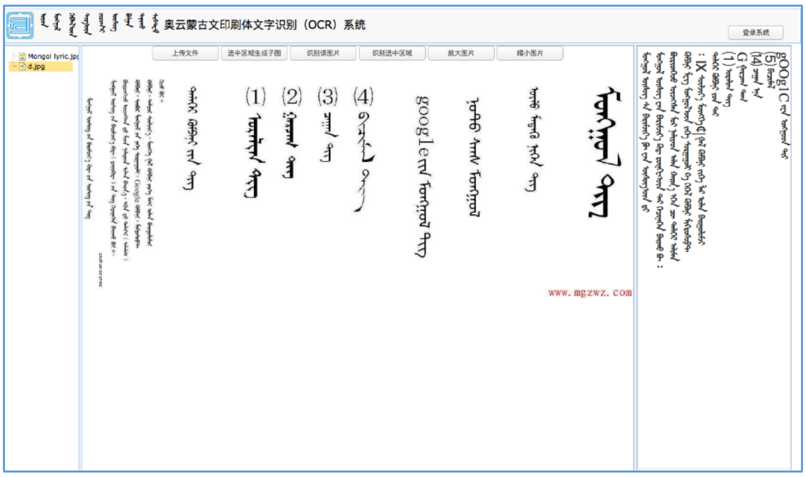


Figure 8: An example of an OCR test result of an image containing words written in different fonts

Appendix

Major Translated Websites in Vertical Mongolian

List of major websites in vertical Mongolian script translated or adapted from parallel Chinese-language websites. These sites could be used for developing a parallel Chinese-vertical Mongolian corpus

Website name (Chinese)	Websites in vertical Mongolian (URL)	Source website in Chinese
人民网	http://mongol.people.com.cn	http://www.people.com.cn
中国蒙古语广播网	http://www.mongolcnr.cn	http://www.cnr.cn
新华网	http://www.nmg.xinhuanet.com/mg/	http://www.xinhuanet.com
中国共产党新闻网	http://mongol.people.com.cn/306956/index.html	http://cpc.people.com.cn
兴安日报	http://xam.mgyxw.net	http://www.xingandaily.cn
内蒙古科技信息网	http://mengwen.nmnet.com.cn	http://www.nmsti.com/2014/

中国蒙古学信息网	http://mgl.surag.net	http://www.surag.net
内蒙古大学蒙古研究中心	http://mgxzx.imu.edu.cn/mgwb/m1.htm	http://mgxzx.imu.edu.cn
乌兰察布蒙古语新闻	http://uq.mgyxw.net	http://www.wlcbnews.com
赤峰市政府网	http://mgl.chifeng.gov.cn/U_index.html	http://m.chifeng.gov.cn
内蒙古自治区政府网	http://mgl.nmg.gov.cn/index.html	http://www.nmg.gov.cn/
新疆日报蒙文版	http://mongol.xjdaily.com	http://www.xjdaily.com.cn
中国蒙古语新闻网	http://www.mgyxw.net/	
新疆蒙古语广播网	http://www.xjmglr.com	
通辽日报	http://tl.mgyxw.net	