LREC 2018 Workshop

Belt & Road: Language Resources and Evaluation

PROCEEDINGS

Edited by

Erhong Yang, Le Sun

Organized by

Beijing Advanced Innovation Center for Language Resources

ISBN: 979-10-95546-29-0
EAN: 9791095546290

8 May 2018

语言资源高精尖创新中心
Beijing Advanced Innovation Center for Language Resources

Proceedings of the LREC2018 Workshop
"Belt & Road: Language Resources and Evaluation"

8 May 2018–MIYAZAKI

Edited by Erhong Yang, Le Sun

http://yuyanziyuan.blcu.edu.cn/art/2018/1/15/art_12642_1128625.html

Organising Committee

- Erhong Yang, Beijing Advanced Innovation Center for Language Resources, China
- Le Sun, Institute of Software, Chinese Academy of Sciences, China
- Nicoletta Calzolari, European Language Resources Association (ELRA),Italy
- Kam-Fai Wong, Chinese University of Hong Kong, China
- Jiahong Yuan, Linguistic Data Consortium

*:Main editors and chairs of the Organising Committee

Programme Committee

- Boxing Chen, National Research Council, Canada
- Yue Zhang, Singapore University of Technology and Design, Singapore
- Qun Liu, Dublin City University, Ireland
- Hitoshi Isahara, Toyohashi University of Technology, Toyohashi, Japan
- Xiaodong Shi, Xiamen University, China
- Yang Liu, Tsinghua University, China
- Chi Mai Luong, Vietnam Academy of Science and Technology
- Luvssandorj Dold, Academician of the Academy of Sciences of Mongolia, Mongolia
- Sharipbayev Altynbek Amirovich, Eurasian National University named after L.N.
- Gumlilyev, Kazakstan
- Nasun-urt, Inner Mongolia University-Mongolian, China
- Tse ring rgyal, Qinghai Normal University-Tibetan, China
- Turgun-lbrahim, Xinjiang University-Uighur, China
- Longyun Xuan, Yanbian University-Korean, China
- Lianfang Liu, Guangxi PingSoft-Zhuang language, China
- Gulila ALTENBEK, Xinjiang University-Kazakh, China
- Xiaobing Zhao, Minzu University of China, China
- Xiaohai Liu, Beijing Language and Culture University, China

Programme

Morning Session 1

09.00–10.30 Keynote Speech 1 (Tentative)

Weili Zhang, Xianpei Han, Le Sun and Ben He
A Geo-Tagged Chinese Poetry Corpus

Amel Fraisse, Quoc-Tan Tran, Ronald Jenn, Patrick Paroubek and Shelley Fisher Fishki TransLiTex
A Parallel Corpus of Translated Literary Texts

Kahaerjiang Abiderexiti, Ayiguli Halike, Maihemuti Maimaiti, Aishan Wumaier, Lulu Wang and Tuergen YIBULAYIN
Semi-Automatic Corpus Expansion for Uyghur Named Entity Relation based on a Hybrid Method

10.30–11.00 Coffee break

Morning Session 2

11.00–13.00 Keynote Speech 2 (Tentative)

Keynote Speech 3 (Tentative)

Huidan Liu, Long Congjun, Long-Long Ma, Jian Wu and Le Sun
CTTC: A Collection of Tibetan Text Corpora

Lili Bao
Study on the Textbook and Related Corpus Construction of the Mongolian language in Primary School

Sihui Fu, Nankai Lin, Gangqin Zhu and Shengyi JIANG Towards
Indonesian Part-of-Speech Tagging: Corpus and Models

13.00–14.00 Lunch break

Afternoon Session 1

14.00–16.00 Keynote Speech 4 (Tentative)

Xuan Zhou, Yinghao Chang and Deqing Cao
A Tentative Idea of Multi-modal Corpus Applied to National Minority Chinese Proficiency Test in China

Maihemuti Maimaiti, Aishan Wumaier, Kahaerjiang Abiderexiti, Lulu Wang and Tuergen YIBULAYIN
Construction of Uyghur named entity corpus

Programme

Lidia S. Chao, Derek F. Wong, Chi Hong Ao and Ana Luisa Leal UM-PCorpus
A Large Portuguese-Chinese Parallel Corpus

Cizhen Jiacuo and Sangjie Duanzhu
A Study on Machine Translation-oriented Parallel Corpus Construction Techniques for Tibetan, Chinese and English

16.00–16.30 Coffee break

Afternoon Session 2

16.30–18.00 Gaoqi RAO and Lung-Hao Lee
NLP for Chinese L2 Writing: Evaluation of Chinese

Na Sun:
A Word and Its Rules

Lianfang Liu, Jiakai Wen, Zixian Deng, Liangchun Lu, Yuanyuan Pan and Lixiang Zhao
The Characteristics of Southeast Asian Languages and Their Influence on Translation

Yanlu Xie, Xin Wei, Wei Wang and Jinsong Zhang
A Semi-manual Annotation Approach for Large CAPT Speech Corpus

Table of Contents

Table of Contents

A Geo-Tagged Classical Chinese Poetry Corpus

Weili Zhang[1], Xianpei Han[1], Le Sun[1], Ben He[2]
[1]Institute of Software, Chinese Academy of Sciences, Beijing China
[2]University of Chinese Academy of Sciences, Beijing, China
[1]{weili, xianpei, sunle}@iscas.ac.cn
[2]benhe@ucas.ac.cn

Abstract

The Chinese poetic tradition is the largest and longest continuous tradition in world literature. Classical Chinese poetry can be divided into certain standard periods or eras, in terms both of specific poems as well as characteristic styles. The knowledge about ancient civilizations can be learned from literature study on these poetry texts. However, there is little research focusing on building classical Chinese poetry resources for automatic natural language processing. In this paper, we take a preliminary step towards the above target and construct a geo-tagged Chinese poetry corpus. Specifically, an annotation criterion is first given to guide the tagging process for consistent annotation. Then we present details about the collecting, annotating and statistics about the data, from which a geo-tagged corpus of 5000 Chinese poems is built. Finally, the corpus is utilized to generate a geographic visualization, verifying its effectiveness on ancient civilization knowledge mining. Our corpus also provides a valuable resource for literature study, intelligent education, spatial data analysis, etc.

Keywords: Chinese poetry, geo-tagged poetry corpus, annotation criterion

1. Introduction

The Chinese poetic tradition is the largest and longest continuous tradition in world literature, which has a long history of more than 2,000 years. Classical Chinese poetry can be divided into certain standard periods or eras, in terms both of specific poems as well as characteristic styles. In ancient China, poetry is one of the most well-known and popular forms of literature, and nearly all scholars aspired to master poem composition. These classical poems help people to express their personal emotion, ambitions and thoughts. It also provides valuable literary texts for knowledge mining of ancient civilizations.

However, there was little research that focuses on building classical Chinese poetry resources for automatic natural language processing. Currently, most existing data sources about classical Chinese poetry are just raw texts. Due to the large size of classical Chinese poetry, and the genre diversity between ancient and modern Chinese language, it is difficult to analyze poetry texts using current natural language processing tools. To automatically understand a poem, a computing system must be able to extract different information about it, such as geographical locations, temporal information, related people and imagery in the poetry. Furthermore, all poems are correlated with each other based on different attributes, such as locations, poets, imageries, etc.

One of the most important knowledge for poem understanding is its geographic location. The geographic information provides background information where a poem was written, and the location itself provides comprehensive background about a poem. Furthermore, geographic information makes it easier to mine knowledge of authors, dynasties, and civilizations, by providing an important dimension for information integration, fusion and visualization. A simple example about the geo-labels of poetry is given in Figure 1.

In this paper, we construct a geo-tagged Chinese poetry corpus for automatically knowledge mining. Specifically, an annotation criterion is first given to guide the tagging process for consistent annotation. Then we present details about the collecting, annotating and statistics about the data, from which a geo-tagged corpus of 5000 Chinese poems is built. Finally, the corpus is utilized to generate a geographic visualization, verifying its effectiveness on ancient civilization knowledge mining. Our corpus also provides a valuable resource for literature study, intelligent education, spatial data analysis, etc.

2. Annotation Criterion

Classical Chinese poetry is written in ancient Chinese language, which is quite different from modern Chinese. Furthermore, China has a long history and most classical Chinese poems are written in ancient time, such as Qin dynasty (221–207 BC), Tang dynasty (AD 618–907) and Song dynasty (AD 960-1279). The above characteristics raise a number of unique challenges for geo tagging of classical Chinese poems. The main challenges are summarized as follows.

1) *Temporal variety of location names*. During the long history, many location names are changed. Therefore, it is quite common that the same location has different names in different times. For example, Soochow(苏州) has ancient names such as Gusu(姑苏), Wujun(吴郡) and WuXian(吴县);
2) *Location Name ambiguity*. Many location names are ambiguous. That is, the same location name may refer to different locations. For example, Han(韩) may refer to Han (state) or Han (Western Zhou state) in different dynasties. Therefore, the geo-tagging must distinguish ambiguous location names for down-stream applications.
3) *Location Role*. In classical Chinese poetry, some locations indicate where the poetry was written, while others are just mentioned in text, like the Hanshan Temple in Figure 1.

Based on the above observation, we geo-tag classical Chinese poems as follows:

1) Locations in poems are divided into two categories, one indicates where the poetry was written (e.g. Fengqiao and Gusu in Figure 1) -- *writing location*, and the other indicates locations mentioned in poetry text (e.g. the Hanshan Temple in Figure 1) -- *mention location*. Given a poem, annotators should label both categories of locations.
2) If more than one writing locations are annotated in a poem, *Part-Of* relations between them must be annotated. For example, in Figure 1, Fengqiao and

Gusu are both writing locations, then annotators must annotate that Fengqiao is *Part-Of* Gusu.

3) If a poem only contains one specific place, annotators should also give the city and the province this place belonging to by exploring the poem's context or background introduction. For example, given the verse 牧童遥指杏花村(The shepherd boy points at Xinghuacun), 杏花村 (Xinghuacun) must be annotated as a place, and the city it belongs to is Chizhou, and the province it belongs to is Anhui province.

4) All aliases, allusions of locations and place names should be labeled.

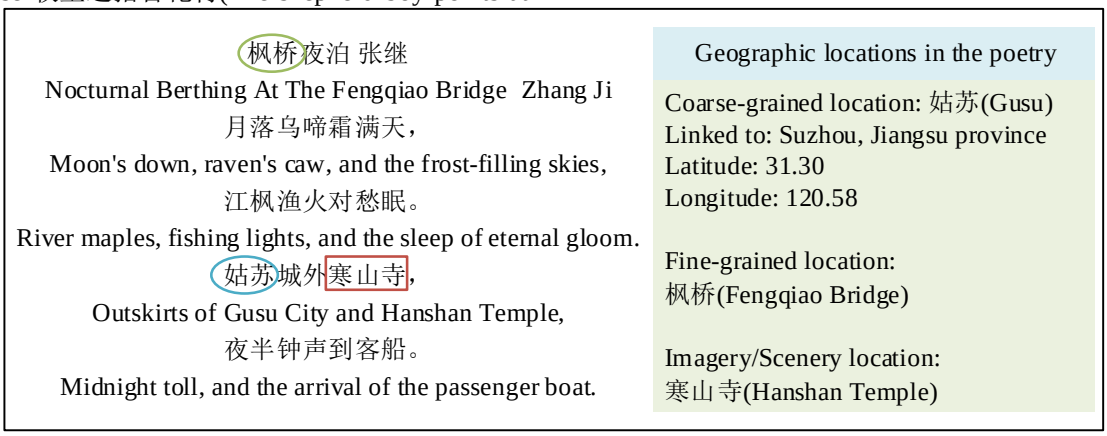

Figure 1: A Chinese poem and the related geographic locations in it

5) To resolve the temporal variety problem of location names, all tagged ancient locations should be linked to its corresponding modern ones. This process is based on a mapping gazetteer between ancient and modern Chinese locations.

6) We use the BMEO annotation schema, where B, M, E, O correspondingly indicate begin, middle, end and out of a location name. For example, 姑/WB 苏/WM 城/WE 外/O 寒/IB 山/IM 寺/IE, where W and I represent the writing location and the mention location respectively. For each location mention, we also annotate its details, including its complete hierarchical district name, latitude and longitude.

3. Corpus

3.1 Data Collection

We prepare a comprehensive collection of poetry for geo-tagging. Specifically, we crawled data from two poetry websites -- Souyun[1] and Gushiwen[2]. After data integration and duplication clean, a total of 750 thousand poems are obtained, which stretch from Pre-Qin period to Qing dynasty, and poems of diverse types and genres are included. Table 1 shows statistics of this poetry collection.

3.2 Data Annotation

In order to construct a representative geo-tagged classical Chinese poetry corpus, we perform a poem selection step for each poetry genre. Specifically, all poems that have explanation notes are selected, because these notes provide background for geo-tagging. Because only a small part of poems have explanation notes, we also randomly sample poems without notes for a balanced distribution, and the sample size is the same as that of the noted ones. Finally, totally 14 thousand poems are selected, and they are reshuffled before provided to annotators.

For each poem, three annotators majoring in literature are invited for labeling according to the criterion in Section 2. Besides poems, annotators are also provided with online resource of poetry allusions[3] for reference, as well as a name mapping gazetteer[4] between ancient and modern locations for linking location names to its current locations. The final tagging results are determined by majority voting from three annotators' results.

3.3 Data Statistics

Finally, our geo-tagged corpus contains 5000 poems. Among them 2500 poems have geo-labels, and about 80.9% and 61.9% have writing locations and mention locations respectively, and 33.3% have both types of locations.

Besides, Figure 2 shows the top 5 ancient places possessing the most poetry, as well as the top 5 ancient places mentioned most frequently by poetry in the corpus. It's not surprising to see that some ancient capitals are so popular that poets love to writing for these places, even when they weren't staying there.

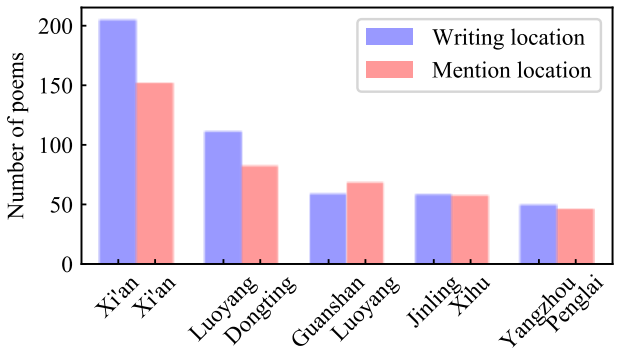

Figure 2: The top 5 writing locations and mention locations in poems of our corpus

[1] https://sou-yun.com/QueryPoem.aspx
[2] http://www.gushiwen.org/shiwen/
[3] http://cls.hs.yzu.edu.tw/orig/home.htm
[4] http://www.360doc.com/document/17/0902/01/11147672_684051406.shtml

4. Applications

The geo-tagged Chinese poetry corpus can be useful in many tasks. Basically, the corpus can provide training data for location extraction from ancient poetry text, which will further facilitate the automatically information extraction from these texts. Furthermore, our geo-tagging corpus provides valuable data resources for literature research, such as author profile, spatial text analysis, intelligent education, data visualization, etc.

We also demonstrate a simple application based on this corpus in Figure 3, which shows the top 11 places where authors write poems about Xi'an, the capital of Tang dynasty. We can see that, the capital of Tang dynasty has important influence on ancient poets: many poems are even from a foreign city -- Tokmak, as shown in Figure 3.

Types and genres	Shi Jing	Quatrain	Regulated verse	Iambic	Drama	Others
Total number	305	188,594	297,746	84,614	9,036	171,461
Number with notes	305	1,400	1,971	1,583	305	1,710

Table 1: Basic statistics of the raw poetry collection

Figure 3: Top 11 places where authors wrote the most poetry using Xi’an as a mention location

5. Conclusions

The Chinese poetic tradition is the largest and longest continuous tradition in world literature. This paper builds a geo-tagged corpus, which provides a valuable resource for knowledge mining, literature research, spatial data analysis, and data visualization. Concretely, we design a basic annotation criterion, and 5000 classical Chinese poems are manually annotated.

6. Acknowledgments

This work is supported by Projects of The Chinese Language Committee under Grants no. WT135-24. Moreover, we sincerely thank the reviewers for their valuable comments.

7. Bibliographical References

Su, J., Zhou C., Li Y. (2007). The Establishment of the Annotated Corpus of Song Dynasty Poetry Based on the Statistical Word Extraction and Rules and Forms. *Journal of Chinese Information Processing*, 21(2):52–57.

TransLiTex: A Parallel Corpus of Translated Literary Texts

Amel Fraisse[1], Quoc-Tan Tran[1], Ronald Jenn[1], Patrick Paroubek[2], Shelley Fisher Fishkin[3]
[1]University of Lille (France), [2]LIMSI-CNRS (France), [3]Stanford University (USA)
{amel.fraisse, ronald.jenn}@univ-lille3.fr, quoc-tan.tran@etu.univ-lille3.fr, pap@limsi.fr, sfishkin@stanford.edu

Abstract

In this paper, we present our ongoing research work to create a massively parallel corpus of translated literary texts which is useful for applications in computational linguistics, translation studies and cross-linguistic corpus studies. Using a crowdsourcing approach, we identified and collected 29 translations of Mark Twain's *Adventures of Huckleberry Finn* published in 23 languages including less-resourced languages. We report on the current status of the corpus, with 5 chapter-aligned translations (English-Dutch, two English-Hungarian, English-Polish and English-Russian). We evaluated the correctness of chapter alignment by computing the percentage of common words between the English version and the translated ones. Results show high percentages that vary between 43% and 64% proving the high correctness of chapter alignment.

Keywords: parallel corpus, comparable corpus, translated literary texts

1. Introduction

Parallel corpora are a valuable resource for linguistic research and natural language processing (NLP) applications. Such corpora are often used for testing new tools and methods in Statistical Machine Translation (SMT), where large amounts of aligned data are often used to learn word alignment models between two languages (Och and Ney, 2003). The most widely used parallel corpora in computational approaches are the Canadian Hansards (Roukos et al., 1995) which are bilingual (English and French), the United Nations Parallel Corpus (6 languages) (Ziemski et al., 2016), or the European Parliament proceedings (21 languages) (Koehn, 2005). These resources belong to the legal and political sphere.

Another source of parallel corpora that has recently attracted attention is religious texts such as the Bible. This line of research, which entailed the compilation of many parallel texts, has broken new ground and allowed computational linguistics to handle vast corpora. Cysouw and Walchli (2007) introduced the notion of 'massively parallel corpora' for texts that have translations into a great number of languages (100+). Although there are not many such texts, those that are available offer an incredibly rich source for computational linguistic researchers. That is why our project taps into relatively unexplored sources for massively parallel corpora: translated literary texts.

A growing number of those texts are now available in electronic form on the internet and they are indexed by public online catalogues such as Wikisource[1], Archive.org[2], Project Gutenberg[3], etc. In this paper, we will report on our ongoing research work to compile such a massively parallel literary corpus. This paper is structured as follows. The next section gives an overview on related work on the construction of parallel corpora. Section 3 describes the Mark Twain translation corpus. Section 4 presents our method for data collection. Section 5 outlines the corpus alignment. Section 6 describes the corpus evaluation and discusses the results. Section 7 mentions the expected use of the corpus and in the last section, we conclude and underline future work.

[1]https://wikisource.org
[2]https://archive.org
[3]https://www.gutenberg.org

2. Related work

Computational linguistics researchers have been exploring different sources for building parallel and comparable corpora. Resnik and Smith (2003) used the Web as parallel text to construct a significant parallel corpus for a low-density language pair.

In accordance with the fast growth of Wikipedia, many works have been published in the last years focused on its use and exploitation for construction of parallel corpora (Tomas et al., 2008; Tufiș et al., 2013; Labaka et al., 2016). Other research works used Twitter as comparable corpus to build multilingual linguistic resources (Fraisse and Paroubek, 2014; Vicente et al., 2016).

There have also been research works which show the potential of the Bible as a source to compile massively parallel corpora (Resnik et al., 1999). Mayer and Cysouw (2014), based on freely available resources, created a Bible corpus with over 900 translations in more than 830 language varieties. Christodouloupoulos and Steedman (2015) built a massively parallel corpus based on 100 translations of the Bible, emphasizing difficulties in acquiring and processing the raw material.

There are also parallel corpora related to translated literary works (e.g. Harry Potter, Le Petit Prince, Master i Margarita) or translations from the web, mostly available for a set of closely related languages (Mayer and Cysouw, 2014; Cysouw and Walchli, 2007). Most of these texts, however, cannot be regarded as massively parallel texts, they are not freely available online, and they mainly concern well-endowed largely known languages.

3. Mark Twain corpus

Mark Twain's books are some of the most well-travelled texts on the planet. As the UNESCO Index Translationum shows the American writer is ranked 15 in the top-50 of the most translated authors worldwide. His works have been

translated into almost every language in which books are printed (Rodney, 1982). The novel *Adventures of Huckleberry Finn* (Twain, 1885) is one of the most commonly translated of his books. Rodney (1982) identified 375 translations as of 1976. As UNESCO's Index Translationum[4] suggests, hundreds of additional translations have been published in the four decades since Rodney completed his survey. But these two sources are both significantly out of date and incomplete. (For example, UNESCO's Index Translationum lists 15 translations of the novel in Chinese, but Lai-Henderson (2015) documented 90 Chinese translations). The scores of language into which the book has been translated include Afrikaans, Albanian, Arabic, Assamese, Bengali, Bulgarian, Burmese, Catalan, Chinese, Chuvash, Czech, Danish, Dutch, Estonian, Farsi, Finnish, French, German, Georgian, Greek, Hebrew, Hindi, Hungarian, Indonesian, Italian, Japanese, Kazakh, Korean, Kirghiz, Latvian, Lithuanian, Macedonian, Malay, Malayalam, Marathi, Norwegian, Oriya, Polish, Portuguese, Romanian, Russian, Serbo-Croatian, Sinhalese, Slovak, Slovenian, Spanish, Swedish, Tamil, Tatar, Telugu, Thai, Turkish, Ukranian, and Uzbek. In many of these languages, there have been multiple translations over time, reflecting different moments in history, and different ideological perspectives on the part of the translators or publishers, as well as different attitudes towards the US, towards childhood, towards minorities and minority dialects, towards race and racism, etc. Of all the existing translations of Mark Twain's works, *Adventures of Huckelberry Finn* stands out because of its rich intercultural content and the great number of translations. That's why we decided to focus on this particular novel for our project.

4. Data collection

To collect our raw data, we proceeded in two steps. First, we built a seed corpus by crawling existing databases and digital archives such as Gutenberg Project , Unesco's Index Translationum, Wikisource, etc. For this seed corpus we collected the original English text of *Adventures of Huckelberry Finn* as well as the French, German, Polish, Russian, Dutch and Hungarian translations. Then, as the example of Chinese translations demonstrates (Lai-Henderson, 2015), the existing sources and databases show cracks that only the power of the crowd can help us fill. That is why we use a crowdsourcing-based approach to discover and collect translations from other languages that hadn't been indexed in those above-mentioned databases.

4.1. Crowdsourcing experiment

Due to the significant amount of existing translations and the growing number of digital versions made available online, the crowdsourcing allowed us to gather data that would have otherwise been beyond our reach. Crowdsourcing helped reduce the amount of time spent on the task, increase the variety and the range of the data covered (such as identifying translations which are not indexed in public databases). We used the CrowdFlower[5] platform. The parametrization of the experiment was as follows: as we are looking for translations over the world, we have not limited the geographic location of the contributors. Each task consisted in a set of 9 questions (i.e. units in the CrowdFlower terminology) and completing the task will earn 0.25 $ (instead of 0.15 $ recommended by CrowdFlower). In fact, the task that the workers had to go through to complete the job was complex. First we asked people to use search engines or online catalogs to look for existing translations in their native language. Then we asked them if they could find the translator's name, the first year of publication, the publishing house, the URL of the cover, the bibliographic record, and available public digital versions.
Because of the complexity of the task, the crowdsourcing approach did not look like the best option. We assumed that the cultural background of crowdsourcing workers would not allow them to complete the task efficiently but it turned out that they managed to provide us with valuable and reliable information. One week after launching the job on CrowdFlower, we received 710 judgements covering 31 different languages. On top came Spanish (163 responses), Arabic (76), Malay and Indonesian (47), German (46), French (45), Greek (43) and Turkish (39). After data cleaning we collected 29 translations in 23 languages of different formats (html, text, pdf, epub).

4.2. Full-text acquisition

Before collecting the full-text, we checked the reliability of the collected translations. First, we use a Python script and Google Translate to verify if the translated titles are equivalent to the original one. Among 710 given titles, we eliminated 25 incorrect responses (such as "The Adventures of Tom Sawyer").
Secondly, we verified the copyright by checking the full-text URLs in order to know whether they came from a national or public institution that has the right to distribute the digital versions. Based on information gathered by the crowd, we crawled further into national archives and digital libraries to get the full-text versions such as Wikisource[6], DBNL[7] (Digital Library for Dutch Literature), Archive.org, Lib.ru[8] (also known as Maksim Moshkow's Library and Russia's Project Gutenberg), MEK[9] (Hungarian Electronic Library), etc.

5. Corpus alignment

The original version of the novel as well as most of the collected translations are already structured by chapter and by paragraph. We kept the original structure for the alignment process. Each translation contains chapter (<CHAPTER>), and paragraph (<P>) mark-ups on separate lines (Figure 1). In this work, we performed an alignment at chapter level by using the mark-up <CHAPTER> as a marker to extract and align chapters of translations that have the same number of chapters as the English source version (43 chapters). In total, we aligned 5 translations (English-Dutch, two English-Hungarian, English-Polish and English-Russian). Transla-

[4] http://www.unesco.org/xtrans/
[5] https://www.crowdflower.com/
[6] https://fr.wikisource.org/
[7] http://www.dbnl.org
[8] http://az.lib.ru/
[9] http://mek.oszk.hu

tions with different numbers of chapters, will be aligned and included in a further version of the corpus. Table 1 describes the number of paragraphs and sentences for each aligned translation.

```
<chapter id="1" name="Civilizing Huck.–Miss Watson–
<p>You don't know about me, without you have read a
<p>Now the way that the book winds up, is this: Tom
<p>The widow she cried over me, and called me a poo
<p>After supper she got out her book and learned me
<p>Pretty soon I wanted to smoke, and asked the wid
<p>Her sister, Miss Watson, a tolerable slim old ma
<p>Now she had got a start, and she went on and tol
<p>Miss Watson she kept pecking at me, and it got t
<p>I set down again, a shaking all over, and got ou
</chapter>
```

Figure 1: Format of the released corpus. Extract from the chapter 1 of the English version.

6. Corpus evaluation and results

In order to evaluate the quality of our parallel corpus, we wanted to determine what degree of similarity between the original text and the translations was. As the alignment unit used for this version of the corpus was chapter–not paragraph or sentence–we evaluated the semantic similarity between two chapters as a whole. The goal of the evaluation is to find out how similar the parallel chapters are. We consider two aligned chapters as similar if they contain a significant percentage of words with the same semantic meaning. Firstly, we identified for each text the direction of translation–that is to say, whether they were directly translated from English or whether they went through another target language first. In fact, the source language has an important influence on the nature of its translation. A manual survey of the five translated versions studied in this work confirmed that they have English as their source language. (English-Polish, English-Russian, two English-Hungarian, English-Dutch). Secondly, we used Google Translate to acquire the English literal translation of each collected target text and compared it to the original. The comparison consists in computing the percentage of common words between each chapter of the literal and the original version (the stop-words are excluded). We used the Stanford tokenizer (Manning et al., 2014) to extract tokens from both texts.

The Figure 2 shows that the percentage of common words ranges between 43% and 64% according to chapters and target languages. For the Polish translation, the lowest score is 43% in chapters 13 and 43, and the highest score is 56% in chapter 1. For Russian, the lowest score is 46% in chapter 43 and the highest score is 60% in chapter 11. In the first Hungarian translation the lowest score is 49% in chapter 21 and the highest score is 64% in chapter 11. In the second Hungarian translation the lowest score is 46% in chapter 43 and the highest score is 60% in chapters 11 and 31. In the Dutch translation the lowest score is 51% in chapters 23, 29 and 43 and the highest score is 61% in chapter 11 .

Although these scores consider only literal and strict translation as common words, they show that the collected translations are similar and staying fairly to the original and could be considered as parallel in the strict sense of the term.

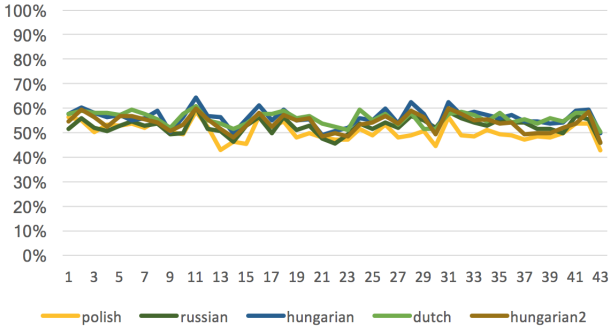

Figure 2: Percentage of common words with the English version by language and chapter.

7. Expected use and availability

One major achievement will be to provide statistical machine translation systems with a rich parallel corpus. This current version of our corpus displays 5 languages (English, Dutch, Hungarian, Polish and Russian) and other languages are being processed so that the corpus will grow over time. One of our ultimate goals is to reach out to the less-resourced languages such as Finnish, Latvian, Malay, Turkish, etc.

Another goal is to engage scholars in the field of digital humanities as well as languages and Translation Studies specialists to address a number of fundamental questions. What happen in translations? What is the impact of the linguistic and cultural transfer of the novel on its textual and iconic nature? An aligned digital corpus would allow them to evaluate the modifications and adaptations set up by translators and the translation process. It will make available to them a stable and reliable corpus to conduct their own research. This research work will raise awareness of corpora and how they can benefit academics both in their research and their teaching in various humanities areas.

The corpus is available online and accessible on Github at the URL: `https://github.com/amelfraisse/TransLiTex/releases/tag/v1.1`.

8. Conclusion and future works

In this work, we provided a parallel corpus of translated literary texts of Mark Twain's *Adventures of Huckleberry Finn*. The aim is to support interdisciplinary research that benefits from the convergence of knowledge in computational linguistics and Translation Studies. On the one hand, it explores new directions in which parallel corpora of literary texts can help produce statistically reliable results. On the other hand, it provides digital humanities scholars with materials for extraction and acquisition of new knowledge. For raw data collection, we resorted to crowdsourcing to discover translations that had not been indexed in public databases such as the UNESCO's Index Translationum and particularly when they were published in less-resourced

Version	Num. of Chapters	Num. of Paragraphs	Num. of Sentences
English	43	2155	6190
Dutch	43	2150	6134
Russian	43	2214	7486
Polish	43	2293	8339
Hungarian 1	43	2237	6503
Hungarian 2	43	2162	6608

Table 1: Characteristics of the realized parallel corpus: numbers of chapters, paragraphs and sentences in different translations.

languages. After verifying the copyright of full-text translations, we aligned them by chapter. We report on the current status of the corpus, with 5 aligned translations in 5 languages (English-Dutch, two English-Hungarian, English-Polish and English-Russian). We evaluated the semantic similarity between aligned chapters by computing the percentage of common words between each chapter of the literal translated and the original version. Results show that the percentage of common words ranges between 43% and 64% proving the high correctness of chapter alignment between the 5 translations. In a future work, we plan to perform the alignment at the paragraph and the sentence level and extend this version to other languages.

Acknowledgments

This research work is conducted within the framework of the "Global Huck" project funded by the MESHS (Maison Européenne des Sciences de l'Homme et de la Société) in Lille, France. It is a partnership with Stanford University. We are also grateful to the Center for Mark Twain Studies in Elmira, N.Y.

9. Bibliographical references

Christodouloupoulos, C. and Steedman, M. (2015). A massively parallel corpus: the bible in 100 languages. *Language Resources and Evaluation*, 49(2):375–395.

Cysouw, M. and Walchli, B. (2007). Parallel texts: Using translational equivalents in linguistic typology. *Sprachtypologie und Universalienforschung STUF*, 60(2):95–99.

Fraisse, A. and Paroubek, P. (2014). Twitter as a comparable corpus to build multilingual affective lexicons. In *Proceedings of the 7th International Workshop on Building and Using Comparable Corpora at LREC 2014*, pages 17–21.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Machine Translation Summit 2005*, page 79–86, Phuket, Thailand.

Labaka, G., Alegria, I., and Sarasola, K. (2016). Domain adaptation in mt using titles in wikipedia as a parallel corpus: Resources and evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Lai-Henderson, S. (2015). *Mark Twain in China*. Stanford University Press, Stanford, CA.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Mayer, T. and Cysouw, M. (2014). Creating a massively parallel bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Resnik, P. and Smith, N. A. (2003). The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.

Resnik, P., Olsen, M. B., and Mona, D. (1999). The bible as a parallel corpus: Annotating the 'book of 2000 tongues'. *Computers and the Humanities*, 33(1):129–153.

Rodney, R. M. (1982). *Mark Twain International: A Bibliography and Interpretation of his Wordwide Popularity*. Greenwood Press, Westport, CT.

Roukos, S., Graff, D., and Melamed, D. (1995). Hansard french/english. In *Philadelphia: Linguistic Data Consortium*.

Tomas, J., Bataller, J., and Casacuberta, F. (2008). Mining wikipedia as a parallel and comparable corpus. *Language Forum*, 34(1).

Tufiș, D., Ion, R., Ștefan Daniel, and Ștefănescu, D. (2013). Wikipedia as an smt training corpus. In *Proceedings of the 9th conference RANLP*.

Twain, M. (1885). *Adventures of Huckelberry Finn*. Charles L. Webster and Company, San Mateo, CA.

Vicente, I. S., Alegria, I., Espana-Bonet, C., Gamallo, P., Oliveira, H. G., Garcia, E. M., Toral, A., Zubiaga, A., and Aranberri, N. (2016). Tweetmt: A parallel microblog corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Ziemski, M., Junczys-Dowmunt, M., and Pouliquen, B. (2016). The united nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, page 3530–3534, Portorož, Slovenia.

Semi-Automatic Corpus Expansion for Uyghur Named Entity Relation based on a Hybrid Method

Kahaerjiang Abiderexiti[1,2] , Ayiguli Halike[1, 2], Maihemuti Maimaiti[1,2], Aishan Wumaier[1,2] , Wanglulu[1,2], Tuergen Yibulayin[1,2]

[1] School of Information Science and Engineering, Xinjiang University, Urumqi, Xinjiang 830046, China
[2] Xinjiang Laboratory of Multi-Language Information Technology, Urumqi, Xinjiang 830046, China
kaharjan@xju.edu.cn, 1506867752@qq.com, mahmutjan@xju.edu.cn,
hasan1479@xju.edu.cn, turgun@xju.edu.cn

Abstract

In order to address the issues that in Uyghur lack of relation extraction method and small size of relation annotated data, we present a semi-automatic way to expand existing Uyghur named entity relation corpus. Our method is based on Conditional Random Fields followed by rules. We integrate our relation extraction method into our annotation tool, with the help of human correction, we expanded existing corpus size by three times.

Keywords: Uyghur, named entity relation, conditional random field

1. Extended Abstract

Relation extraction is the prerequisite task of recognizing and characterizing a particular relationship between two or more entities in text. Depending on the languages in which annotated named entity relation corpora are available, relation extraction has been studied using different machine learning methods, including supervised, semi-supervised and even unsupervised method. Those studies focus on English language and other resources rich languages. However, relation extraction in Uyghur language, which is ethnic minority language that wildly used in Xinjiang Uyghur Autonomous Region of China, there are two problems: 1) there are no studies have reported regarding with relation extraction method. 2) the existing annotated Uyghur named entity relation corpus size is relatively small. To address these issues, we utilized the existing Uyghur named entity relation annotated corpus which only contains a small amount of annotated news articles to propose a hybrid semi-automatic method of expanding existing annotated corpus. Our method is based on Conditional Random Fields (CRFs) followed by some rule based post processing and manual correction. In this way, we expanded the corpus size by three times than the existing one.

2. Introduction

Relation extraction is the most important task in natural language processing, especially in the field of information retrieval, knowledge graph. The aim of relation extraction tasks is recognizing and characterizing a particular relationship between two or more entities in text automatically. There are several methods in relation extraction including supervised methods, semi-supervised methods, distant supervised methods and unsupervised methods. Supervise and semi-supervised methods require human annotated data depending on which methods are applied. Usually, supervised methods require more annotated data than semi-supervised methods. Unsupervised may not require annotated data but still need large amount of unannotated data. For Uyghur language, one of the official languages in Xinjiang Uyghur Autonomous Region in China, both annotated and unannotated data are scarce. However, relation extraction requires a certain amount of annotated corpus especially in supervised learning. The size of Uyghur named entity relation annotated data which is available on the Internet is relatively small. It is difficult to train Uyghur relation extractor using this small data. The data size has become one of the major limitations in studying relation extraction in Uyghur. And also there is not any report about relation extraction in Uyghur language. To solve these problems, we proposed the hybrid semi-automatic method that expanding existing annotated corpus based on conditional random fields (CRFs).

In this paper, first we describe related work in Uyghur corpus construction and relation extraction in other languages. Then describe our method which aims to expending existing corpus size by relation extraction. Finally, we show our results and discuss further improvements.

3. Related Work

Entity relationships are the key to building a knowledge graph, there are many methods for relation extraction. In the traditional approaches, entity recognition is considered as a predecessor step in a pipeline for relation extraction (Zelenko,etal.2003). However, it is ignored the dependence between the two tasks. and (Li Y, et al.2011) research entity relation descriptor based on linear-chain CRFs, and that reduce the space of possible label sequences and introduce long-range features. Recently, neural network based methods have become popular in natural language processing. In the relation extraction, there are also several methods which are based on neural network and also methods with using joint extraction with named entity. R kai, W Shi-Wen(2016) proposes an abbreviation disambiguation method based on the convolutional neural network (CNN) to solve the abbreviation disambiguation problem in the biomedical field when no labelled corpus exists and obtained an average of 90.1% accuracy. However, in this way the named entity recognition accuracy affects the relation extraction. Join extraction of entities and relations is the new way to address this

problem. (Zheng, et al.2017) use the novel tagging scheme whose annotations can be converted joint extraction task and study different end-to-end models to extract entities and relations . (Miwa M and Bansal M, 2016) propose end-to-end relation extraction method used BI_LSTM, and this model can extract jointly both named entities and relations with shared parameters.(Zhang M, Zhang Y,2017) also proposed global optimized neural model ,and achieving the best performances on two standard benchmarks.In Uyghur language processing, there are some works about constructing corpus, particularly, to solve shortness of Uyghur named entity and named entity relation corpus, (Kahaerjiang Abiderixiti, et.al, 2017) proposed the method for construction Uyghur named entity and relation corpus and release small size of annotated corpus. Wushouer J, et al. constructed "contemporary Uyghur grammatical Information Dictionary", which is provided a large amount of grammatical information and collocation features, and it is the basic resources of NLP. There is also a research about building Uyghur Dependency Treebank which is built from a public reading corpora (Aili M),and it is also important for linguistic researches.

4. CRFs Model

The task of relation extraction can be seen as a sequence labeling problem. For the small amount of data, conditional random fields (CRFs) model would be more effective. So we choose CRFs model as the main model expending our corpus by the help of relation extraction model. CRFs is one of the commonly used algorithms in natural language processing in recent years, which combine the maximum entropy and hidden Markov model, which is a typical non-directional pattern model of discriminant probability. The CRFs attempt to model the conditional probability of multiple variables after a given observed value. $X = \{x_1, x_2, x_3 \dots \dots x_n\}$ are the observed sequence, $Y = \{y_1, y_2 \dots \dots y_n\}$ are the corresponding tag sequences. Construction of conditional probability model is the goal of CRFs (Maimaiti, M.). The definition of CRF model (Lafferty,et al. 2001) is as follows:

Definition: set $G(V, E)$ as a node of V, and E as an undirected graph of a set with no edges.

$Y = \{Y_v | v \in V\}$, Each node in V corresponds to a random variableY_v, whose value range is a possible set of tags. If an observed sequence X for conditions, each random variable satisfy Markov characteristics as follows:

$$P(Y_V \backslash X, w \neq v) = P(Y_V \backslash X, Y_w, w \sim v) \quad (1)$$

Where W: V, denotes that two nodes are adjacent nodes in graph G. Then, (X, Y) as a conditional random field. The CRFs model calculates the conditional probability model of the output node as the condition of given input nodes. For a chain with parameter $\theta = \theta_1 \theta_2 \dots \theta_k$,the conditional probability of a state sequence obtained for a given sequence x is defined as:

$$P_\theta(Y|X) = \frac{1}{Z_x}\{\sum_{n=1}^{N} \sum_k f(y_{n-1}, y_n, X, n)\} \quad (2)$$

Among them, the denominator Z_x normalized factor:

It enables that the sum of the probabilities of all possible state sequence of the given input to be 1. $f_k(y_{n-1}, y_n, X, n)$for the whole observing sequence X, the feature functions, which are located at n and n-1, may be 0, 1, or any real number. In most cases, the characteristic function is a binary representation function; When the characteristic condition is satisfied, the value is 1, otherwise it is 0. The definition of characteristic function as shown below:

$$f(y_i, x, i) = \begin{cases} 1 & if\ y_{i-1} = y` and\ y_i = y \\ 0 & other \end{cases} \quad (3)$$

$$g(y_i, x, i) = \begin{cases} 1 & if\ x_i = x\ and\ y_i = y \\ 0 & other \end{cases} \quad (4)$$

$\theta = \theta_1 \theta_2 \dots \theta_k$ is the corresponding weight for the feature function. For X, in the next step is to search the $Y* = \text{argmax}\ P(X\backslash Y)$ with the highest probability.

5. Uyghur Relation Extraction Model

On the basis of analysis in the grammatical and semantic features of Uyghur named entity relation, we firstly use the CRfs model to study Uyghur relation extraction method. The reason of using CRFs is that it is better neural network models when the data size is relatively small. In the design of features, we use different Uyghur language features such as words, syllables, part-of-speech tagging and distributed word vector representations.

5.1 Task Definition

As we said above , relation extraction task can be seen as sequences labeling problem. As shown on figure 1, annotation for relation extraction from raw texts. it only consists of relation extraction tags, which recognizes valid relations over entity pairs. According to the characteristics and difficulties of Uyghur Named Entity and relations, we construct feature sets between the different entity categories. The features of our method are divided into word related features and the dictionary features. Firstly, word related features include Uyghur words itself, part-of-speech tagging, syllable, word length and syllable length, etc. Because Uyghur words stemming is complicated and there are no good stemming tools publicly available., so we didn't use stem characteristics. Secondly, the dictionary features, It`s feature include common dictionaries, person name dictionaries, place name dictionaries, organization name dictionaries, and the similarity dictionaries based on word vectors, etc. The following Table 1 shown as, the feature information of a every word in the Uyghur sentence.

5.2 Feature Template

The influence of combining different features on the named entity relations extraction can't be ignored. Therefore, the selection of feature templates plays an extremely important role in relation extraction. Named entity relations extraction need to consider the context, whereas the CRFs model synthesizes contextual information as well as external features. Named entity relation recognition needs to consider the context, whereas the CRFs model synthesizes contextual information as well as external features.

In this paper, we use CRFSharp open source tools to build Uyghur named entity relation recognition model, the use of defined characteristics acquired template feature to learn. In the model, not only the atomic feature (unary feature) template, but also the composite feature template has to be defined. The definition of feature templates common to throughout in this paper is given in Table 2.

words	$feature_1$	$feature_2$	$feature_{...}$	$feature_n$	Final Tag
شىنجاڭ	Arg1	Arg2	…….	Arg_n	B_Org-Aff.Employment_1
ئۇنۇۋېرسىتتى	Arg1	Arg2	…….	Arg_n	E_Org-Aff.Employment_1
مەكتەپ	Arg1	Arg2	…….	Arg_n	O
مۇدىرى	Arg1	Arg2	…….	Arg_n	O
ۋەلى	Arg1	Arg2	…….	Arg_n	B_Org-Aff.Employment_2
ياقۇپ	Arg1	Arg2	…….	Arg_n	E_Org-Aff.Employment_2
.	Arg1	Arg2	…….	Arg_n	O

Table 1: Sequence labeling tags for relation extraction task

Feature type	Template	Meaning
Atomic feature	$w_i(-2 \le i \le 2, i \in Z)$	The current word w_i and its upper and lower two window words, the word's window size is 5
	$F_i(-1 \le i \le 1, i \in Z)$	The characteristics of the current word F_i and the words of its upper and lower windows, ie the window size is 3
Composite feature	$w_{i-1} \mid w_i(0 \le i \le 1, i \in Z)$	The combination of the current word and its upper words feature
	$F_{i-1} \mid F_i((0 \le i \le 1, i \in Z))$	The combination feature of the current and upper words
	$w_i \mid F_i(i = 0)$	The current word and its combination features
	$F_{i-1} \mid F_i \mid F_{i+1}(i = 0)$	The characteristics of the current word and compound features of one window`s upper and below

Table 2: feature template

In Table 2, w represents the first column of the corpus, that is, a column of words, and F denotes other characteristic columns except words; in which the $F_{i-1}|F_i|F_{i+1}(i = 0)$ in the composite feature represents a combination of three features, and the other three features represent the binary feature combinations.

5.3 Rules

The relation is different from the Named Entity. And Uyghur relation extraction is more difficult and more challenging. Because the annotated data size is small and Uyghur language is morphologically complex. We simplified the task and assume that the named entity is given. So we just tagged the relationship between the named entities represented in a sentence. However, the result still did not help much annotator. So after relation extraction based on CRFs. we use some rules to posted it the extraction result. Relation annotation has some rules. What we are considering here is the rules for the annotated relations.

5.3.1 Physical.Located

Physical location relations: The first argument of this relationship must be a person. The second argument can be for facilities (FAC), location (LOC) and geographical social and political entities (GPE) shown by Table 3.

Relation type	Arg1	Arg2
Physical.located	PER	FAC,LOC,GPE

Table 3: Candidate Arguments for Physical.located

In the example below, گۈزەلنۇر (Guzalnur) is the Arg1 (PER) and it is must be a person and شاڭخەيدە (in Shanghai) is Arg2 and it is belonging to LOC.

گۈزەلنۇر شاڭخەيدە ئوقۇۋاتىدۇ .

Guzalnur is studding in Shanghai

5.3.2 Physical.Near

The rules for Physical.near relation is also shown in Table 4.

Relation type	Arg1	Arg2
Physical.Near	PER,FAC, GPE,LOC	FAC,LOC,GPE

Table 4: Candidate Arguments for Physical.Near

In the example below, قەشقەر ۋىلايىتى (Kashagar prefecture) is the Arg1 (GPE), ئاقسۇ ۋىلايىتىنىڭ (Aksu prefecture)is the Arg2 (GPE).

قەشقەر ۋىلايىتى ئاقسۇ ۋىلايىتىنىڭ جەنۇبىغا جايلاشقان .

Kashgar prefecture is located the sought of Aksu prefecture.

As described above, all the subtypes of named entity relations have certain rules, which rules similar to previous example. And all types and subtypes are classified as follows:

I. GenPart-Whole (Geographical, Subsidiary)
II. Personal-Social (Business, Family, Role, other)
III. Physical (Located, Near)
IV.ORG-Affiliation (Employment, investor-Shareholder, Student-Alum, Owner, Founder)
V. General-Affiliation
(PersonAge, Organizationwebsite).

6. Results

After we add rules for post edit which descried section 5, our suggestion for relation annotation is improved. We got positive feedback from our human annotator.

As the result, we have mainly completed the following two tasks. The first one is, we integrate our relation extraction model into our annotation tool.

The second one is, when annotating the named entity relationship corpus, our model provides a suggestion for annotation, thereby alleviate the burden of human annotators. . Thus, the scale of our annotated corpus is increasing rapidly. The annotation result of this tool is shown as below:

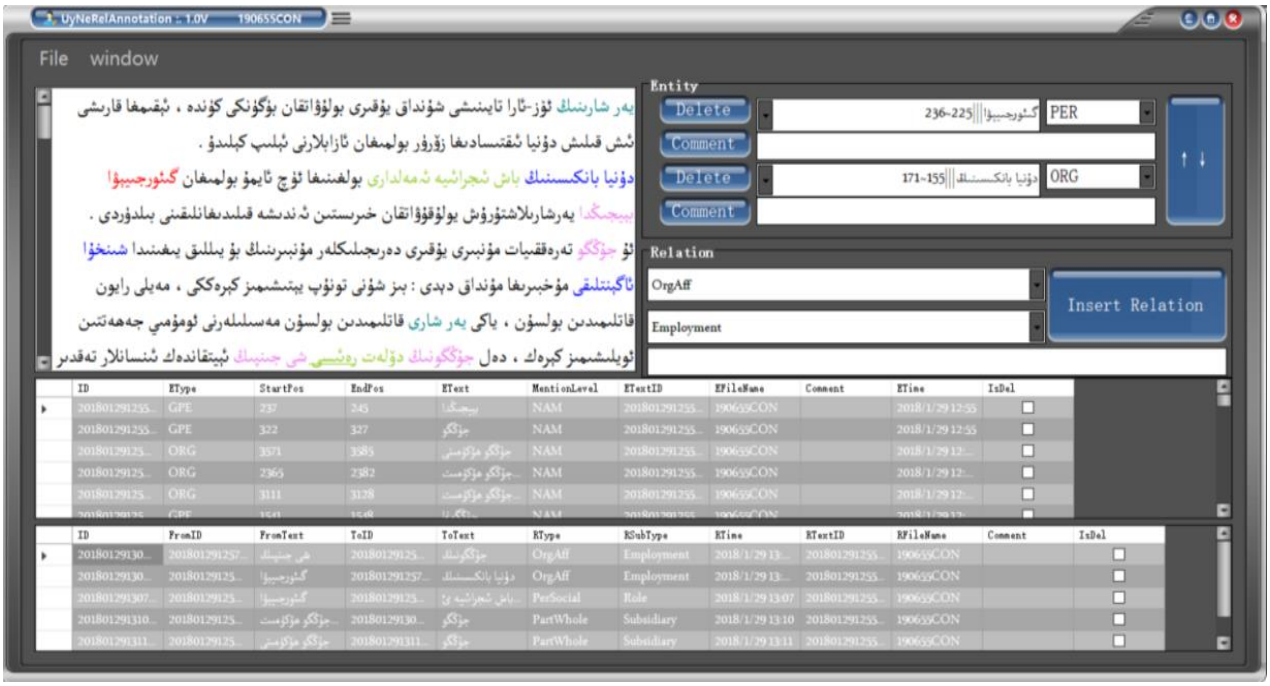

Figure 1: Interface of the UyNeRel Annotation Software

We estimated that our human annotation speed is speedup more than three times. Two annotators who are already familiar with relation annotation annotate 21 documents per day before integrating our relation extraction model into our relation extraction tool. After we add our model to annotation tool their annotation speed is increased and they have annotated 62 documents per day on average. In this semi-automatic way, the existing 500 documents increased to 1500 in 77 days in average. As the result the relation annotation corpus size is expanded by three times in this semi-automatic way.

7. Conclusion

In this study, we described our work on expending Uyghur Named Entity Relation, the main purpose of this article is to provide the extension annotated corpus which is needed in the study of automatic relation extraction. In the immediate future, we plan to focus on the relation extraction task based on neural network in Uyghur.

8. Acknowledgements

This work has been supported as part of the NSFC (61462083, 61762084, 61463048，61262060), 973 Program (2014cb340506), and 2017YFB1002103, ZDI135-54.

9. References

Guodong, Z., Jian, S., Jie, Z., & Min, Z. (2002). Exploring Various Knowledge in Relation Extraction. ACL 2005, Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 25-30 June 2005, University of Michigan, Usa (419--444). DBLP.

Zheng, S., Wang, F., Bao, H., Hao, Y., Zhou, P., & Xu, B. (2017). Joint extraction of entities and relations based on a novel tagging scheme. 1227-1236.

Wingendcr, M. (2007). Standardisation tendencies in an expanded europe - a corpus-based study of the anglicisms in polish. Welt der Slaven-Halbjahresschrift fur Slavistik, 52(1), 1-20.

Rimkus, C. D. M., Junqueira, T. D. F., Callegaro, D., & Leite, C. D. C. (2013). Segmented corpus callosum diffusivity correlates with the expanded disability status scale score in the early stages of relapsing-remitting multiple sclerosis. Clinics, 68(8), 1115-1120.

Abiderexiti, K., Maimaiti, M., Yibulayin, T., & Wumaier, A. (2017). Annotation schemes for constructing Uyghur named entity relation corpus. International Conference on Asian Language Processing. IEEE.

Wushouer, J., Abulizi, W., Abiderexiti, K., Yibulayin, T., Aili, M., & Maimaitimin, S. (2015). Building Contemporary Uyghur Grammatical Information Dictionary. Revised Selected Papers of the Second International Workshop on Worldwide Language Service Infrastructure (pp.137-144). Springer-Verlag New York, Inc.

Aili, M., Xialifu, A., Maihefureti, & Maimaitimin, S. (2015). Building Uyghur Dependency Treebank: Design Principles, Annotation Schema and Tools. Worldwide

Language Service Infrastructure. Springer International Publishing.

Maimaiti, M., Wumaier, A., Abiderexiti, K., & Yibulayin, T. (2017). Bidirectional long short-term memory network with a conditional random field layer for uyghur part-of-speech tagging. Information, 8(4), 157.

Li, Y., Jiang, J., Hai, L., Ming, K., & Chai, A. (2011). Extracting relation descriptors with conditional random fields. Asian Federation of Natural Language Processing, 392-400.

Zhang, M., Zhang, Y., & Fu, G. (2017). End-to-End Neural Relation Extraction with Global Optimization. Conference on Empirical Methods in Natural Language Processing (pp.1730-1740).

Miwa, M., & Bansal, M. (2016). End-to-end relation extraction using lstms on sequences and tree structures.

CTTC: A Collection of Tibetan Text Corpora

Huidan Liu[a], Congjun Long[b], Longlong Ma[a], Jian Wu[a], Le Sun[a]
a.Institute of Software, Chinese Academy of Sciences, Beijing, China, 100190
b.Institute of Ethnology and Anthropology, Chinese Academy of Social Sciences, Beijing, China, 100081
huian@iscas.ac.cn,longcj@cass.org.cn,wujian@iscas.ac.cn,sunle@iscas.ac.cn

Abstract

The Chinese Academy of Sciences launched the Multi-Layer MultiLingual Resource Database (MLLRD) project which aims to collect language resources for natural language processing tasks for low resource languages used in China, such as Mongolian, Tibetan, Uyghur and so on. Tibetan text corpus building is one of the sub projects, in which we have built a Collection of Tibetan Text Corpora(CTTC), including: (1) Tibetan web article corpus which has 440,900 documents. (2)Tibetan text classification corpus. (3) Chinese-Tibetan parallel text corpus which has 773,068 sentence pairs. (4) Part-Of-Speech tagged corpus which has 52,041 sentences. (5) Tibetan tree bank which has 6,040 trees. The paper reports the methods to build these corpora, the contents and scales of each corpus, and applications of them.

Keywords: Tibetan, Corpus, Machine Translation, Tree Bank

1. Introduction

Corpora are the basic and necessary materials for natural language processing. The Chinese Academy of Sciences launched the Multi-Layer MultiLingual Resource Database (MLLRD) project which aims to collecting language resources for natural language processing tasks for low resource languages used in China. The project collects resources generally for three tasks, namely machine translation, speech recognition, hand written character recognition. For machine translation task, bilingual sentence level aligned parallel text are collected for Chinese-Tibetan, Chinese-Uyghur and Chinese-Mongolian. There are more than 300 thousand sentence pairs for any of the three language pairs. For speech recognition task, text-speech aligned sentences are collected for Tibetan, Uyghur and Mongolian. There is 360GB speech data in total. For hand written character recognition, hand written characters from 300 writers are collected for each of the three languages.
Tibetan text corpus building is one of the sub projects. In the sub project we have built a Collection Tibetan Text Corpora(CTTC), including: (1) Tibetan web article corpus. (2)Tibetan text classification corpus. (3) Chinese-Tibetan parallel text corpus. (4) Part-Of-Speech tagged corpus. (5) Tibetan tree bank. We introduce these corpora in the following sections.

2. Tibetan Web Article Corpus

2.1. Sources of the Corpus

Previously Liu et al. (2012b) proposed an approach to build a large scale text corpus for Tibetan natural language processing. We adopt the method to build our corpus. We use a web crawler initialized with a seed URL list, which includes some well-known Tibetan websites. Then we check the crawled web page whether it contains Tibetan text with a Tibetan examiner, and if a page has Tibetan text in it, all URLs which it links to are appended to the fetching list of the crawler. The procedure continues until no new Tibetan web pages are found. After that we know where to get Tibetan text. For Tibetan web article corpus, we crawled articles from 19 Tibetan websites which mainly focus on news and broadcastings(Table 1).

1	http://blog.amdotibet.cn
2	http://epaper.chinatibetnews.com
3	http://tb.chinatibetnews.com
4	http://tb.tibet.cn
5	http://tb.xzxw.com
6	http://ti.gzznews.com
7	http://ti.kbcmw.com
8	http://ti.tibet3.com
9	http://tibet.cpc.people.com.cn
10	http://tibet.people.com.cn
11	http://tibetan.qh.gov.cn
12	http://www.amdotibet.cn
13	http://www.qhtb.cn
14	http://www.tbmgar.com
15	http://www.tibet3.com
16	http://www.tibetcnr.com
17	http://www.tibetology.ac.cn
18	http://www.vtibet.com
19	http://xizang.news.cn

Table 1: Sources of Tibetan web article corpus.

2.2. URL Filtering

Web pages can be classified into two kinds, namely "topic" and "hub". A topic page contains long text in it while a hub page contains many links to the topic pages. As our target is to extract Tibetan web articles from the web pages, We only care about the topic pages rather than the hub pages. Topic pages rather than hub pages are selected with a rule based method by checking the url.
Table 2 and Table 3 show some URLs of topic pages and hub pages of the three Tibetan web sites respectively. Comparing tens of thousands of URLs of the three web sites, we find the following rules:

- The topic URLs of "Tibetan Web of China" have the pattern of "{host}/{column}/{year}-

Site	Example URLs
China Tibet Online	http://tibet.people.com.cn/141101/15137028.html http://tibet.people.com.cn/141101/15199715.html http://tibet.people.com.cn/15143391.html
Tibetan's Web of China	http://ti.tibet3.com/economy/2011-01/14/content_370366.htm http://ti.tibet3.com/folkways/2008-12/10/content_3541.htm http://ti.tibet3.com/medicine/2009-10/27/content_99171.htm

Table 2: Example URLs of topic pages.

Site	Example URLs
China Tibet Online	http://tibet.people.com.cn/140827/141059/index3.html http://tibet.people.com.cn/96372/125163/index.html http://tibet.people.com.cn/141101/index11.html
Tibetan's Web of China	http://ti.tibet3.com/culture/index.htm http://ti.tibet3.com/tour/node_701.htm http://ti.tibet3.com/economy/index.htm

Table 3: Example URLs of hub pages.

{month}/{date}/content_{articleid}.htm". Everyone of them contains the string "content_".

- The hub URLs of "Tibetan Web of China" contain the string "index" or "node".
- The topic URLs of "China Tibet Online" have the pattern of "{host}/{columnid}/{articleid}.html". Characters between the host URL "{host}" and the file suffix name "html" are numbers or slash.
- The hub URLs of "China Tibet Online" contain the string "index".

With these rules, we make text extraction only on the topic pages.

2.3. Text extraction

We analysed the layout structure of the web pages from each web site and get clues to build templates to extract topic title, publishing date, author, topic content and some other topic related informations. Figure 1 shows the structure of a web page[1]. From the figure, we see that there are some HTML tags giving the boundaries of different text blocks, and we can find the corresponding HTML tags of the article title, publishing date, author, article content and so on.

2.4. Content of the Corpus

At present, the corpus has about 440.9 thousands documents, 9.50 million sentences, 228 million syllables in total. Each Article is saved as an XML file. Figure 2 shows an article from the corpus.

2.5. Quality of the Corpus

Some predefined rules are used to check whether there are spelling errors in a syllable in a previous of the corpus. The statistical data show that there are 9700 misspelt ones out of the 20743 Tibetan syllables occurred in the corpus, which shares 46.7628%. But their occurrence is only 27,427 in the 93 million syllable in total, sharing only 0.0308%(Liu et al., 2017), which shows that the corpus has a very high quality.

[1] http://tibet.people.com.cn/15260188.html

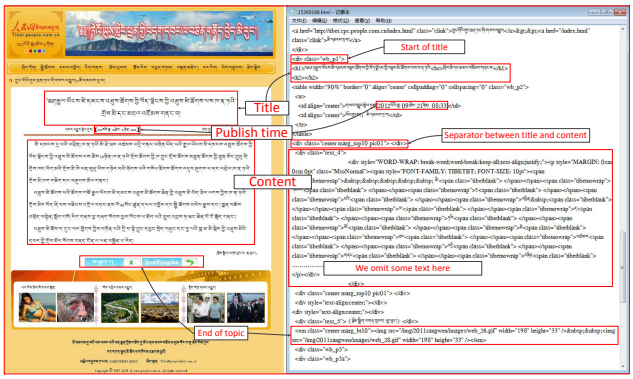

Figure 1: The structure of a web page from "China Tibet Online".

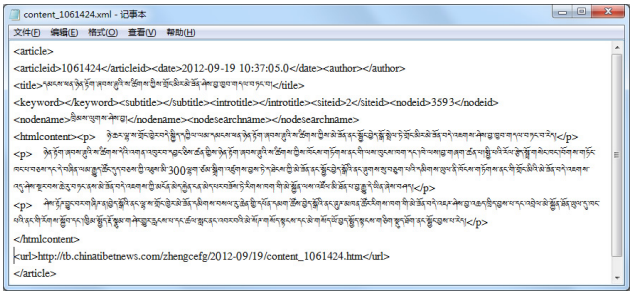

Figure 2: The text extracted from a page from "China Tibet news", in XML format.

3. Tibetan Text Classification Corpus

The web article corpus is further processed to build a text classification corpus. It's a heavy task to manually classify those document into domains. However, we can get the domain information for a certain subsets of the web article corpus. For some web sites listed above, we can get the domain information from the URL of each web page. For instance, the URL "http://tb.chinatibetnews.com/xzmeishi/2011-12/05/content_831210.htm" shows it belongs to a column called "xzmeishi". so it must be a page about Tibetan foods, because "xz" is the abbreviated form of Chinese word "xizang" (西藏), which means the Tibetan Autonomous Region, while "meishi" means "delicious food". So we can classify the documents in the corpus into domains. Table 4 and 5 list the domains of two subsets of the articles from two web sites named "China Tibet News" and "Tibetan's web of China" respectively. Obviously, a large part of the documents in the corpus are news as expected, because the two web sites are both hold by news agencies.

4. Chinese-Tibetan Parallel Corpus

4.1. Sources of the Corpus

We get documents for the Chinese-Tibetan parallel corpus from two types of sources. The first source of the corpus is translation agencies. A large part of documents in our corpus are collected from several translating agencies. As most of them are translated from Chinese to Tibetan, we

	Domain	♯doc	(%)	♯sent	(%)
1	Art	3,240	4.76	112,642	8.71
2	Economy	712	1.05	12,477	0.96
3	History	2,897	4.25	19,627	1.52
4	News	25,247	37.08	576,842	44.59
5	Picture	12,732	18.70	51,088	3.95
6	Politics	3,230	4.74	63,437	4.90
7	Rural Life	2,402	3.53	35,535	2.75
8	Social Life	1,153	1.69	9,881	0.76
9	Specials	9,986	14.67	268,003	20.72
10	Technology	1,988	2.92	38,321	2.96
11	Buddhism	1,983	2.91	48,832	3.77
12	Food	215	0.32	2,963	0.23
13	Medicine	720	1.06	36,676	2.84
14	Tour	1,588	2.33	17,296	1.34
Total		68,093	100.00	1,293,620	100.00

Table 4: Domains of a subset of the documents from "China Tibet News".

Order	Domain	♯doc	(%)	♯sent	(%)
1	Art	92	0.35	3,021	0.45
2	Culture	885	3.40	109,749	16.18
3	Economy	78	0.30	7,749	1.14
4	Education	15	0.06	695	0.10
5	Music	323	1.24	3,169	0.47
6	News	24,055	92.45	519,576	76.61
7	Photo	80	0.31	2,548	0.38
8	Policy	116	0.45	7,062	1.04
9	Politics	124	0.48	7,668	1.13
10	Medicine	107	0.41	11,417	1.68
11	Tour	145	0.56	5,563	0.82
Total		26,020	100.00	678,217	100.00

Table 5: Domains of a subset of the documents from "Tibetan's web of China".

know the correspondence between the Chinese part and the Tibetan part when we got them. We have nearly 600 thousand sentence pairs from the first source.

The second source of the corpus is the web. We collected articles from two web sites as listed in Table 6 which publish articles in both Chinese and Tibetan. They mainly focus on news and broadcastings. We have about 202 thousand sentence pairs from the second source.

	Host	Language
1	http://tb.xzxw.com	Tibetan
2	http://www.xzxw.com	Chinese
3	http://ti.tibet3.com	Tibetan
4	http://www.tibet3.com	Chinese

Table 6: Sources of the bilingual corpus.

4.2. Document Alignment

We use a feature based method to find the Chinese correspondence for each Tibetan article. Three kinds of features are used: numbers, common punctuations and geographic names in the context of each document. Numbers and some punctuations have same presentation in Chinese and Tibetan while geographic names are translated fixedly. Thus we regard them as good clues to make the document alignment.

4.2.1. Number Extraction

Table 7 shows three ways to present Tibetan numbers. In our method, we extract first two forms of numbers in Tibetan documents, and transfer the numbers presented as Tibetan symbol digits to Arabic numbers.

Form	Description	Example
Arabic numbers	consist of Arabic number (0 to 9)	"2012"
Tibetan digital numbers	alike Arabic numbers, consist of Tibetan digital character	``༢༠༡༠'' (2010)
Tibetan syllable numbers	consist of one or several Tibetan syllables	``བཅུ་ལྔ'' (15)

Table 7: Three ways to present Tibetan numbers

Table 8 shows two ways to present Chinese numbers. In our method, we extract first form of numbers in Chinese documents.

Form	Description	Example
Arabic numbers	consist of Arabic digit (0 to 9)	'2012'
Chinese digital numbers	consist of one or several Chinese syllables	``十五'' (15)

Table 8: Two ways to present Chinese numbers

4.2.2. Punctuation Extraction

Chinese and Tibetan have their own punctuation marking system respectively. However, Tibetan borrows some Chinese punctuations, such as parentheses "()", book title mark "" and double quotation mark. In our method, we extract these three punctuation marks as features for they will be preserved in the same form when an article is translated from Chinese to Tibetan or from Tibetan to Chinese.

4.2.3. Geographic Names Extraction

We use a bilingual dictionary of Chinese and Tibetan, which consist of most of place of interest and administrative division in Tibet to extract geographic names in articles with maximum matching method. Tibetan geographic names are translated to Chinese which is taken as the features to make document alignment.

4.2.4. Candidate Document Pair Generation

In the Internet, there are millions of Chinese and Tibetan documents, so it's necessary to filter document pairs that are impossible to be parallel. As the number of extracted Chinese articles are much larger than that of Tibetan ones,

we try to find the translation for each Tibetan article. For each Tibetan article, if the publishing date of a Chinese article from the same web site is less than 15 days before or after the publishing date of the Tibetan article, it is taken as a candidate, and will be further computed whether it's a translation for the Tibetan article.

4.2.5. Document alignment

If two documents are parallel, the occurrence and orders of the numbers, geographic name and punctuations in each documents will be basically the same. So, we use normalize edit distance of feature vectors of two documents as the measure to identify whether they are parallel. The range of normalize edit distance is 0.0 to 1.0, and the smaller distance means the occurrence and orders of the features in two documents are more coincided and the more likely they are parallel. The formula of normalized edit distance between document a and b is:

$$NED(a, b) = \frac{ED(a, b)}{max\{|a|, |b|\}} \quad (1)$$

ED(a,b) is the edit distance between document a and b, and mathematically given by $ed_{a,b}(|a|, |b|)$ where:

$$ed_{a,b}(i, j) = \begin{cases} max\{i, j\}, \ \ if \ \ min(i, j) = 0 \\ min \begin{cases} ed_{a,b}(i-1, j) + 1 \\ ed_{a,b}(i, j-1) + 1 \\ ed_{a,b}(i-1, j-1) + [a_i \neq b_j] \end{cases} \end{cases} \quad (2)$$

If the normalize edit distance of two documents is less than threshold 0.2, the document pair will be recorded for manual judgement.

4.2.6. Manual judgement

No matter two documents are transliterate or paraphrase, as long as two documents tell the same story, they will be judged as parallel documents. The parallel documents will be archived into parallel corpus.

4.3. Sentence Alignment

After we get bilingual articles, we make sentence level alignment. The Chinese part of the corpus is processed by some rules and open source tools. Some rules are used to segment Chinese text into sentences. A Chinese word segmentation tool named "ICTCLAS" is used to segment Chinese sentence into words. We also segment Tibetan texts into words with a Tibetan word segmentation tool(Liu et al., 2011; Liu et al., 2015). A Chinese-Tibetan dictionary with 137,873 items was collected by combining several published dictionaries(Liu et al., 2011; Liu et al., 2012a). Bilingual articles are respectively segmented into monolingual sentences. They are further segmented into words. As the correspondences of some words in Tibetan sentence to their Chinese translations in Chinese sentence exist, a dynamic programming algorithm is applied to find the correspondence of the sentences in each pair of Bilingual articles. A previous study shows that the aligning precision of this approach is 84.8%(Yu et al., 2011). We implemented a tool (Figure 3) for further proofreading to correct alignment errors.

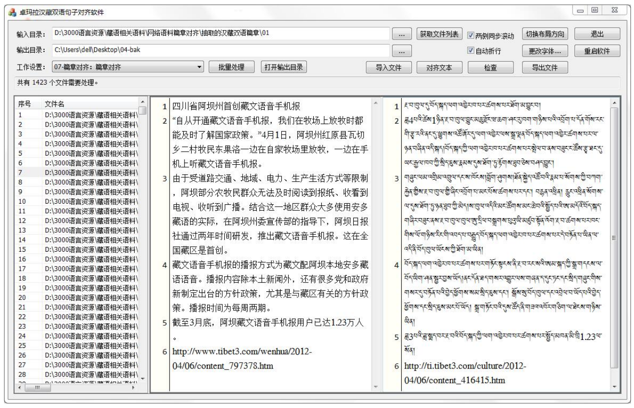

Figure 3: Chinese Tibetan sentence alignment tool.

4.4. Content of the Corpus

The corpus has 771 thousand bilingual sentence pairs. As shown in Table 9 and Table 10, two versions are included. Set A is a long sentence version and Set B is a short sentence version. Each sentence pair in Set A has a complete Chinese sentence which ends with a period, while each sentence pair in Set B has a shorter Chinese sentence which may end with comma. Both of the sets are used in the machine translation system.

Domain	♯Sent (Set A)	♯Sent (Set B)
Law text	115,299	68,535
Leader's works	53,292	96,181
News	228,613	4,270
Government reports	72,849	102,795
Dictionary	31,234	
Total	501,287	271,781

Table 10: Bilingual sentence level parallel corpus.

5. Tibetan Part-Of-Speech Tagged Corpus

We collected Tibetan sentences from some textbooks used in primary school and middle school. Sentences are segmented into words by a dictionary based Tibetan word segmentation tool named "SegT"(Liu et al., 2012a) and tagged with Part-Of-Speech tags manually. This is the first version of the corpus. A Tibetan lexical analyser was trained with CRFs model using this version of the corpus.

After that, a subset of the web article corpus is used to build a larger Tibetan Part-Of-Speech tagged corpus. The Tibetan lexical analyser is applied on the raw text and wrong segmented words and POS tags are corrected manually. Then, we get a newer version of the corpus, and trained newer version of Tibetan lexical analyser. This procedure is implemented iteratively. An annotation inconsistence checking tool is implement to proofread the corpus. Figure fig:AnnotationChecker shows it.

The present version of the corpus has 52,041 sentences, 731,716 words in total. The following sentences are samples from the corpus.

- ནམ་མཁའ་/ng སྔོ་/a ཞིང་/c དྭངས་/a པ་/h ལ་/c །/xp

Long version	Chinese	西藏自治区面积122万平方公里, 平均海拔在4000米以上,有着独特的自然生态和地理环境。
	Tibetan	བོད་རང་སྐྱོང་ལྗོངས་ཀྱི་ས་ཁྱོན་ལ་སྤྱི་བཞིན་ས་ཁྲི་122ཡོད་པ་དང་། རྒྱ་མཚོའི་རིས་ལས་མཐོ་ཚད་ཆསྟེམས་ས་སྤྱི་4000ཡན་ཆེན་པ་དང་། རང་བྱུང་གི་སྐྱེ་ཁམས་དང་ས་གཤིས་ཀྱི་ཡུལ་ཁམས་དཔེགས་བསལ་ཡིན།
Short version	Chinese	西藏自治区面积122万平方公里,
	Tibetan	བོད་རང་སྐྱོང་ལྗོངས་ཀྱི་ས་ཁྱོན་ལ་སྤྱི་ལེ་ཤུ་བཞིན་ས་ཁྲི་122ཡོད་པ་དང་།
	Chinese	平均海拔在4000米以上,
	Tibetan	རྒྱ་མཚོའི་རིས་ལས་མཐོ་ཚད་ཆསྟེམས་ས་སྤྱི་4000ཡན་ཆེན་པ་དང་།
	Chinese	有着独特的自然生态和地理环境。
	Tibetan	རང་བྱུང་གི་སྐྱེ་ཁམས་དང་ས་གཤིས་ཀྱི་ཡུལ་ཁམས་དཔེགས་བསལ་ཡིན།

Table 9: Long version and the corresponding short version of the sentence pairs.

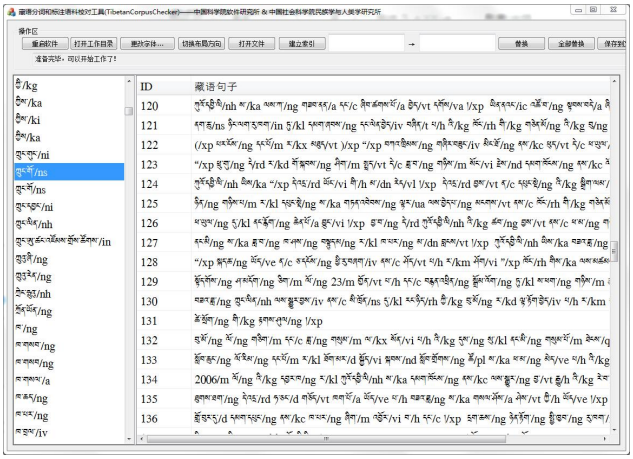

Figure 4: Tibetan annotation inconsistence checking tool.

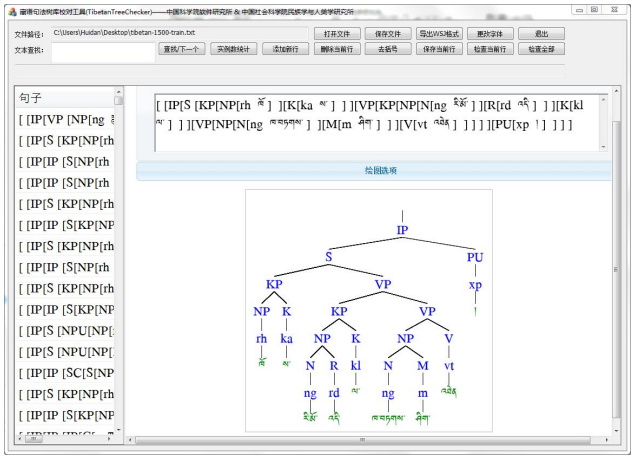

Figure 5: Tool for building Tibetan tree bank.

- སྒྲིན་དཀར་/ng རི་མོ་/kl ཐིས་/vt ཡ་/h ནང་བཞིན་/ua མཇོས་/a །/xp
- བར་སྣང་/ng མཐོ་རས་/ng ཀུ་/kl ཁྲུང་ཁྲུང་/ng ཁྱུ་/q གཅིག་/m བྱལ་བསྒྲིགས་/iv ནས་/c སྡེ་སྡོམ་གས་/ng ཀུ་/kx འཕྲུར་/vi ཕྱིན་/t འདུག་/ve
- ས་གཞི་/ng ར་/kl གསེར་ཁུ་/ng ཐུགས་/vt ཡ་/h ནང་བཞིན་/ua སེར་ལམ་ལས་/ia ཅིད་/vt །/xp
- ཞིང་ས་/ng ཚོ་/pl ས་/ka ཐེལ་འཚོག་/ng ཆེན་པོ་/a ས་/ki བཅོམ་མ་/ng རྡ་/vt ཞིང་/c །/xp
- གདོང་/ng ནས་/kc ལོ་ཤེགས་/a ཐུང་/vi བ་/h ཅི་/kg དགའ་སྤོ་/ng ཅི་/kg འཚོམ་མདངས་/ng ཐུང་/vt །/xp

6. Tibetan Tree Bank

The aforementioned Tibetan word segmented and POS tagged corpus is taken as the basis to build Tibetan tree bank. As there is much manual work to do to add parentheses which indicate boundaries of phrases in different granularities, a software tool as show in Figure 5 is implemented to show a tree visually and to find errors in the trees. The following errors will be checked and people will be asked to correct them:

- The leaf node has no word or no POS.
- Parentheses don't match.
- A node has more than two children.
- Two trees are found in an item.

The present version of the corpus has 6,040 trees, including 51,429 words in total. The following trees are samples from the corpus.

- (IP(VP(NP(ng ཆུང་ཟུང་)) (V(vl རེད་))) (PU(xp ।)))
- (IP(S(KP(NP(rh ཁོ་)) (K(ka ས་))) (VP(KP(NP(N(ng རིམ་)) (R(rd འདི་))) (K(kl ལ་))) (VP(NP(N(ng ཁ་བཏགས་)) (M(m ཤིག་))) (V(vt འཛེན་))))) (PU(xp ।)))
- (IP(IP(S(NP(rh ང་)) (VP(TP(ng ཉིན་ཆུན་)) (VP(KP(NP(ng སྨོད་ལྡ་)) (K(kx ར་))) (V(vi འགྲོ་))))) (I(T(h གི་)) (E(ve ཡོད་)))) (PU(xp ।)))
- (IP(IP(S(NP(rh ང་)) (VP(KP(NP(N(ng ཆ་བདང་)) (NP(N(ng སོད་ཡིག་)) (N(ng སློབ་འབྲིང་)))) (K(kx ཏུ་)) (vi འགྲོ་)))) (I(T(h གི་)) (E(ve ཡོད་)))) (PU(xp ।)))
- (IP(S(KP(NP(rh ཁོང་)) (K(ka ཀྱིས་))) (VP(NP(NNP(KP(NP(ng རྒྱ་མཚོ་)) (K(kg ནི་))) (N(ng གནས་ཚུལ་))) (R(rw ཅེ་ཙམ་))) (V(vt ཤེས་)))) (PU(xp ।)))

- (IP(S(KP(NP(rh ར)) (K(ka ས))) (VP(VPH(VP(NP(ng ཉིན་ཁ)) (V(ve ཡོད))) (H(h ལ))) (V(vt ཚོམས)))) (PU(xp །)))
- (IP(IP(S(NP(rh ཁོ)) (VP(VPC(VP(KP(NP(ng ཡུལ་མི)) (K(ks ར)))) (V(iv ཁ་བྱས))) (C(c ནས))) (V(vi བསྡད)))) (I(E(ve ཡོད)))) (PU(xp །)))
- (IP(IP(S(KP(NP(rh ར)) (K(ka ས))) (VP(KP(NP(rh ཁྱོད)) (K(kd ལ))) (VP(V(iv རོགས་བྱེད)) (AUX(va འགོས))))) (Y(yi ས))) (PU(xp །)))

7. Application of the Corpora

7.1. Tibetan Language Modelling

The web article corpus is used to train a Tibetan language model which is further used in a Chinese-Tibetan machine translation system.

7.2. Word Embedding

The web article corpus is also used to get the distributed representation of (sub)syllables and words to implement NLP systems with deep neural networks for several tasks such as Tibetan word segmentation, machine translation and so on.

7.3. Tibetan Text Classification

A subset of the corpus from “China Tibet News” is used in the research of Tibetan news classification. We studied the methods of text classification of Tibetan news web documents. We used bag of word model to represent Tibetan documents, and implemented four kinds of models for classification, such as the K-nearest neighbour, logistic regression, multi-layer perception and support vector machine. A training set of 4718 documents and a test set of 500 documents with 8 categories were constructed according to the URL of Tibetan web documents. Experiment shows the multi-layer perception achieves the topmost accuracy rate of 84.6%.

7.4. Machine Translation

We built a Chinese-Tibetan machine translation system based on the encoder-decoder structure with attention mechanism. The encoder encodes a source sentence into a fixed-length vector by using recurrent neural network. The decoder generates a translation word by word and allows a model to automatically search for parts of a source sentence that are relevant to predicting a target word by using the mechanism of attention. With this approach, experiments show that the method achieves a NIST score and BLEU score of 6.39 and 0.296 on a sub set of the corpus with 390 thousand sentence pairs as the training set and 1,000 sentence pairs as the test set. Figure 6 shows the demo page of the machine translation system.

7.5. Assistant Translation

We also built a Chinese-Tibetan assistant translation tool based on the machine translation system as the translation server. The assistant translation tool gets machine translation result from the server. Meanwhile, it gets translations and bilingual sentence pair instances for the phrase or word which is currently being translated. Figure 7 shows the assistant translation tool.

Figure 6: Chinese-Tibetan machine translation system.

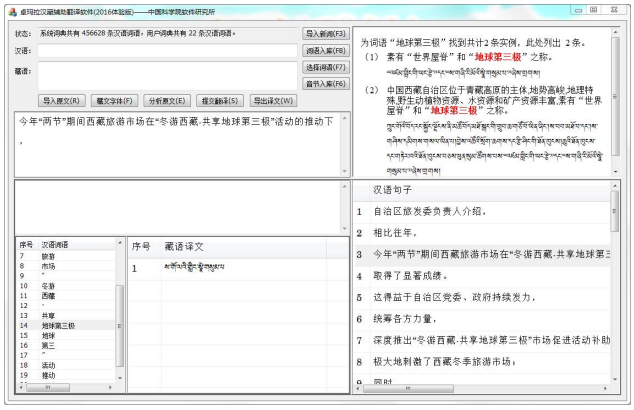

Figure 7: Chinese-Tibetan assistant translation tool.

7.6. Tibetan Lexical Analysis

The Tibetan POS tagged corpu is used to train a Tibetan Word Segmentation and POS tagging tool (Tibetan lexical analyser) with the conditional random fields model. The CRF++ toolkit 0.58 [2] by Taku Kudo is used. About $1/5$ of the corpus are randomly selected as the test set, 3,983 sentences (47,332 words) in total. The remaining 15,931 sentences (191,852 words) forms the training set. The OOV rate of the test set is 5.34%. Sub syllable tagging method is used to reformulate the word segmentation and POS tagging into a universal tagging task, and a machine learning model is trained to predict both the word position and the POS of each sub syllable. Thus we get the first version of the Tibetan lexical analyser. It gets F1 score of 94.43% on the test set(Liu et al., 2015).

[2] http://taku910.github.io/crfpp

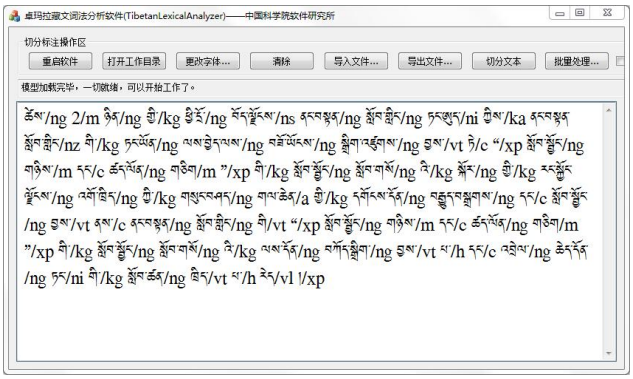

Figure 8: Tibetan lexical analyser.

7.7. Tibetan Parsing

We use the Berkeley parser to train a Tibetan parser with a former version of the Tibetan tree bank. The training set and the test set include 3,746 and 354 trees respectively. Experiments show that if the POS tags are provided, the parser achieves a better performance. The precision, recall and F1 are 0.9251, 0.9273 and 0.9262 respectively.

8. Conclusion

As a low-resource language, Tibetan language processing is facing a big challenge. In this paper, we introduced our work on building a collection of Tibetan text corpora(CTTC), including: (1) Tibetan web article corpus. (2)Tibetan text classification corpus. (3) Chinese-Tibetan parallel text corpus. (4) Part-Of-Speech tagged corpus. (5) Tibetan tree bank. The corpora are applied in many research tasks such language modelling, machine translation, lexical analysis, text classification, parsing and so on.

In the future, we will collect more web text to increase the scales of these corpora, especially for Tibetan tree bank as its scale is still small. We will also make more annotations based on the existing corpora to build corpora for other Tibetan NLP tasks. CTTC is available for academic researches by contacting the authors.

9. Acknowledgements

We thank the reviewers for their critical and constructive comments and suggestions that helped us improve the quality of the paper. The research is partially supported by Informationization Project of the Chinese Academy of Sciences (No.XXH12504-1-10) and Research Project of the National Language Committee(ZDI135-17).

10. Bibliographical References

Ai, J., Yu, H., and Li, Y. (2009). Statistical analysis on tibetan shaped structure. *Journal of Computer Applications*, 29(7):2029–2031.

Gao, D. and Gong, Y. (2005). A statistically study on the qualities of all modern tibetan character set. *Journal of Chinese Information Processing*, 19(1):71–75.

Jiang, D. and Dong, Y. (1994). Statistical analysis on linear processing of tibetan clustered structures. *Chinese Information Processing*, (4):44–46.

Jiang, D. and Dong, Y. (1995). Research on property of tibetan characters as information processing. *Journal of Chinese Information Processing*, 9(2):37–44.

Jiang, D. and Kong, J. (2006). *Advances on the Minority Language Processing of China*. Social Sciences Academic Press, Beijing, China.

Jiang, D. and Long, C. (2010). *On Characters of Tibetan Writing System: Alpabetic Characters, Pronunciations, ISO Codes, Frequencies, Sorting Orders, Picture Symbols and Transliterations*. Social Sciences Academic Press, Beijing, China.

Jiang, D. (2006). History and development of tibetan text information processing. In *Frontiers of Chinese Information Processing - Proceedings of the 25th Anniversary Conference of Chinese Information Processing Society of China*, pages 83–97. Tsinghua University Press, Beijing, China.

Liu, H., Nuo, M., Ma, L., and et al. (2011). Tibetan Word Segmentation as Syllable Tagging Using Conditional Random Fields. In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation (PACLIC 2011)*, pages 168–177.

Liu, H., Nuo, M., Ma, L., and et al. (2012a). Segt: A pragmatic tibetan word segmentation system. *Journal of Chinese Information Processing*, 26(1):97–103.

Liu, H., Nuo, M., Wu, J., and He, Y. (2012b). Building large scale text corpus for tibetan by extracting text from web pages. In *Proceedings of the 10th asian language resources at COLING 2012*, pages 8–17.

Liu, H., Long, C., Nuo, M., and Wu, J. (2015). Tibetan word segmentation as sub-syllable tagging with syllables part-of-speech property. In Maosong Sun, et al., editors, *Chinese computational linguistics and natural language processing based on naturally annotated big data (LNAI 9427)*, pages 189–201.

Liu, H., Hong, J., Nuo, M., and Wu, J. (2017). Statistics and analysis on spell errors of tibetan syllables based on a large scale web text corpus. *Journal of Chinese Information Processing*, 31(2):61–70.

Lu, Y., Ma, S., Zhang, M., and Luo, G. (2003). Researches of calculations of tibetan characters, pieces, syllables, vocabulary and universal frequency and its applications. *Journal of Northwest Minorities University(Natural Science)*, 24(48):32–42.

Yu, X., Wu, J., and Hong, J. (2011). Research and realization of dictionary-based chinese-tibetan sentence alignment. *Journal of Chinese Information Processing*, 25(4):57–62.

Study on the Textbook and Related Corpus Construction of the Mongolian language in Primary School

Lili Bao
School of Mongolian Studies, Inner Mongolia University,Huhhot,010021,China

The textbook of the Mongolian language in primary school is an important part of education of Mongolian students. As a kind of carrier to popularize or propagate the national compulsory education and national culture, it accumulates rich national language resources. As Mongolian language course is important for students, constructing the corpus of textbooks of the Mongolian language, exploring the situation of text-selecting and the layout of words are the main steps in designing and implementing systematic teaching. It will also be one of the important means to improve the quality of the textbook.

In this paper, we use computer-aided data processing techniques to develop the Mongolian corpus, language classroom corpus, test corpus, extracurricular reading corpus and student composition corpus of 2015-2016 school year in primary school, and automatically analyze vocabulary, grammar and semantic features such as descriptive statistics, text legibility score, reference cohesion, latent semantic analysis, lexical diversity, conjunctions, situational patterns, syntactic complexity, syntactic pattern density, vocabulary information and readability. It is to reveal the connotation of national culture and characteristic information. The study will not only fill the gaps in the quantitative study of words in Mongolian language, but also provide a basic material and scientific basis for the compilation of student dictionary and objective evaluating texts.

Towards Indonesian Part-of-Speech Tagging: Corpus and Models

Sihui Fu[1], Nankai Lin[1], Gangqin Zhu[2], Shengyi Jiang[1(✉)]
[1] School of Information Science and Technology
Guangdong University of Foreign Studies, Guangzhou, China
[2] The Faculty of Asian Languages and Cultures
Guangdong University of Foreign Studies, Guangzhou, China
sihuifu93@gmail.com, neakail@outlook.com,
199210621@oamail.gdufs.edu.cn, jiangshengyi@163.com

Abstract

As a member of the Malayo-Polynesian languages, Indonesian is spoken by a large population. However, language resources and processing tools for Indonesian are quite limited. Part-of-speech (POS) tagging aims to assign a particular POS to a word, concerning its distribution and function in the context, which can provide valuable information for most natural language processing tasks. This work introduces our work on designing an Indonesian part-of-speech (POS) tagset, including 29 tags, and constructing a large Indonesian POS corpus comprised of over 355,000 tokens. During the design and annotation processes, we make judgments mostly from a typological perspective, following the specifications of Universal Dependencies, while not missing those language-specific phenomena. In addition, we try to utilize several state-of-the-art sequence labeling models, trained on the proposed corpus, to implement automatic POS tagging, and the experiment results are favorable, with the accuracies higher than 94%.

Keywords: Indonesian, part-of-speech tagging, corpus

1. Introduction

The part-of-speech (POS), also referred to as the grammatical category of a word, signifies the morphological and syntactic behaviors of a lexical item. Some common ones include verbs, nouns, adjectives and adverbs. POS tagging is the process of assigning a particular POS to a word based on both its definition and its context. Since POS can provide valuable linguistic information, POS tagging is an underlying step for most natural language processing (NLP) tasks, such as chunking, syntactic parsing, word sense disambiguation, and machine translation.

Bahasa Indonesia (Indonesian for 'language of Indonesia') is a member of the Malayo-Polynesian branch of the Austronesian language family. Unlike English or other high-resource languages, although spoken by over 198 million people (Simons and Fennig, 2017), Indonesian possesses quite limited language resources, which also leads to the limited development of language technology applied to Indonesian.

Some previous studies have presented their efforts on the construction of Indonesian POS corpora or automatic POS taggers (Dinakaramani et al., 2014; Pisceldo et al., 2009; Rashel et al., 2014), but to the best of our knowledge, either the size of the corpora they used is not big, or the taggers do not attain satisfactory performance.

In this paper, we report our work on designing an Indonesian POS tagset and building a large manually-tagged Indonesian POS corpus comprised of over 355,000 tokens, under the instructions of Universal Dependencies[1]. Furthermore, we attempt to achieve automatic POS tagging using state-of-the-art models.

In the remaining parts, section 2 will briefly review previous work on Indonesian POS tagging. The design and construction processes are described in section 3. The models we employ to build POS taggers are introduced in section 4. Section 5 gives experiment setups and results and section 6 concludes.

2. Related Work

In terms of Indonesian POS tagging, only few corpora are available and relevant processing tools are not mature enough. Pisceldo et al. (2009) employed two POS tag schemes (containing 37 and 25 tags respectively) to manually annotate two Indonesian corpora[2], and intended to develop an Indonesian POS tagger based on conditional random fields (CRF) and maximum entropy (ME). However, the size of their corpora is small (40,513 tokens in total). Also, the experiment results on corpus 1 are not ideal (The highest accuracy is 77.36% for 37 tags and 85.02% for 25 tags).

Wicaksono and Purwarianti (2010) built a Hidden Markov Model (HMM) based POS tagger on a 15,000-token Indonesian corpus, which was proposed in Pisceldo et al. (2009). Affix tree, succeeding POS tags and additional dictionary lexicon were used to improve the performance of vanilla HMM. Subsequently, they extended their work to develop an Indonesian Mind Map Generator (Purwarianti et al., 2013), which includes several Indonesian NLP tools such as POS tagger, syntactic parser and semantic analyzer. The POS tagger was built based on the methods mentioned in their 2010 work, while a decision tree was also used to handle the empty score of emission probability.

Dinakaramani et al. (2014) explored the design of a linguistically motivated Indonesian POS scheme,

[1] http://universaldependencies.org/
[2] For each corpus, they tried both POS tag schemes.

and manually tagged a corpus containing 10,000 sentences (over 250,000 tokens). Furthermore, they developed a rule-based POS tagger by combining several language resources, including closed-class tagging dictionary, multi-word expression dictionary, MorphInd (Larasati et al., 2011) and disambiguation rules, and then applied this tagger to their previously proposed corpus, obtaining an accuracy of 79% (Rashel et al., 2014).

In this work, we propose our own Indonesian POS tagset and present a larger Indonesian POS corpus, compared with previous work. Moreover, we attempt to build an automatic POS tagger based on state-of-the-art models.

3. The Corpus

3.1 The Design of Indonesian Tagset

Both Indonesia and English employ the Latin alphabet as their writing system and (almost) contain the same letters. On the other hand, many Indonesian words are derived from English words, such as *komputer* 'computer', *halo* 'hello', *mesin* 'machine', etc. Given these similarities, we first based our design of Indonesian POS tagset on the existing English ones, and chose the Penn Treebank tagset (Santorini, 1990) for its maturity and popularity. Furthermore, by virtue of the guidelines of Universal Dependencies, we regulated our initial tagset to achieve cross-linguistically consistent annotation, while attending to language-specific phenomena. In addition, during the manual annotation, we also consulted previous work on Indonesian tagsets (Dinakaramani et al., 2014; Pisceldo et al., 2009) to see if any revision was required.

One of the guiding principles is simplicity. Different corpora adopt different tagging schemes, which leads to varying sizes of tagset. For example, there are 87 tags in the Brown Corpus tagset, 45 in the English Penn Treebank tagset, whereas 137 tags in the UCREL CLAWS7 tagset. Considering that manual annotation of a large scale corpus is labor-intensive, a tagset consisting of lots of tags will increase annotators' cognitive load, and therefore we should propose a small tagset, while maintaining useful linguistic information for later natural language processing tools. Meanwhile, we want to develop a corpus which can describe the common properties and structural diversities of multiple languages, from the perspective of linguistic typology. Hence, following the instruction of Universal Dependencies, we abstracted those widespread grammatical categories found cross-linguistically (universality), but did not ignore those specific ones in Indonesian (particularity).

3.2 Data Source

To obtain attested Indonesian data, we crawled substantive news articles from the website detik.com [3], Indo-Asia-Pacific Defense Forum [4], BBC Indonesia[5], etc., whose content covers various topics including politics, finance, society, military, etc. After separating paragraphs into individual sentences, we randomly picked out over 20,000 sentences as our dataset to be annotated. Altogether, the corpus has over 355,000 lexical tokens.

3.3 The Annotation Process

This section will briefly introduce the annotation process of our dataset. In general, the process contains five steps and seven human annotators are engaged in.

(1) The first 2000 sentences are manually annotated, according to the initial tagset constructed on the basis of the Penn Treebank tagset. Annotators need to annotate each word in a sentence by means of its syntactic function and definition in the KBBI dictionary [5]. In this step, considerable issues are put forward, such as the adequate tags for abbreviations, combinations of digits and letters, book titles, and website links. Solutions are presented after discussions and agreement of all annotators and the tagset was revised accordingly.

(2) Referring to the specifications in Universal Dependency, we further regulated our tagset. Thus, the definition of a grammatical category in Indonesian is more consistent with that in other languages. However, specific properties could not be neglected, such as the pronominal suffix in the preposition-object structure, *olehnya* 'by him/her'. Therefore, several language-specific POS tags are also proposed.

(3) The first 2000 sentences were retagged. Annotators should make their judgments based on the specifications in Universal Dependencies and syntactic information. Issues were welcomed to raise and solved by joint discussions.

(4) The remaining sentences were manually tagged, in accordance with the procedure described in (3).

(5) We manually evaluated and revised the tagged sentences with the help of annotators. At the same time, some previous work was reviewed to make comparisons and revisions.

3.4 Problematic Cases

Unlike English, it seems that Indonesian does not have a standardized grammar system by far, which has brought about plentiful confusions and disputes to our design of POS tags and the process of data annotation. On the one hand, different people have different opinions on a word as to its grammatical category. For instance, *sudah* 'already' is regarded as an adverb in KBBI, but as a modal verb in Dinakaramani et al. (2014); *sekarang* 'now' is regarded as an adverb in Pisceldo et al. (2009) but a common noun in Dinakaramani et al. (2014). According to Tallerman (2015), three important linguistic criteria for identifying a word's class is to check its morphosyntax, distribution and function in

[3] https://news.detik.com/

[4] http://apdf-magazine.com/id/

[5] http://www.bbc.com/indonesia

a phrase or sentence. However, Indonesian is not an inflectional language, which means morphosyntax may not help. Therefore, though annotators would rely more on Indonesian dictionaries at the beginning, in the subsequent stages we required them to make judgments based on a word's distribution and function in its context, instead of being bound to the POS given by dictionaries.

On the other hand, it is difficult to achieve the balance between universality and particularity. In Indonesian grammar, there exist some special grammatical categories, which we might find their alternatives in the universal framework. In such case, whether to retain the original terms or to incorporate them in the unified framework is a problem, since rough incorporation may lose the traits of the individual language. A typical example is those indefinite numbers in Indonesian, including *beberapa* 'some', *semua* 'all', *banyak* 'many', etc., which may be regarded as indefinite pronouns in English. However, we noticed other indefinite numbers like *belasan* 'eleven to nineteen' and *ratusan* 'hundreds' share more similarities with numbers. Thus, we at last preserve the category 'indefinite number'.

In addition, since Indonesian is an agglutinative language, many of its complex words are formed by stringing together multiple morphemes (including stems and affixes) without changing their spellings. One case is those words with pronominal suffixes, such as *namamu* 'your name' (*nama* 'name', *-mu* 'your'), *olehnya* 'by him/her/them' (*oleh* 'by', *-nya* 'him/her/them'), etc. Figure 1 lists three cases concerning the use of the pronominal suffix *–nya* ('INJ' suggests interjection). Some previous work separates such words into the stems and suffixes and tags them respectively. However, we insist to maintain a word's integrity, and therefore propose three unique tags: SP (subject-predicate relation), VO (verb-object relation) and PO (preposition-object relation), corresponding to the three cases in Figure 1 respectively. One might argue that such words should be tagged according to the grammatical categories of the heads in these structures. We will not deal with it in this work, leaving it open for discussions.

(1) "Tapi saya tidak marah kok", **katanya**.
But I not angry INJ said.she
"But I am not angry!", she said.
(2) Orang tuanya **mengusirnya** dari rumah.
People old.his expelled.him from home
His parents threw him out of the house.
(3) Mesin yang rusak ini diperbaiki **olehnya**.
machine that broken this repaired by.him
The broken machine was repaired by him.

Figure 1. Several cases of the pronominal suffixes *–nya* in Indonesian

3.4 Indonesian POS tagset

The final version of our Indonesian POS tagset is presented in Table 1, consisting of 29 tags.

Tag	Description	Example
CC	Coordinating conjunction	dan, tetapi, atau
CD	Cardinal number	satu, dua, tiga, 79, 2017, 0.1
DT	Determiner	para, sang, si, sebuah, seorang
FW	Foreign word	poetry, technology, out, world
ID	Indefinite number	puluhan, segala, 30-an, beberapa
IN	Preposition	di, ke, oleh, untuk, dari, antara
JJ	Adjective	besar, tinggi, manis, cerdik
JJS	Adjective, superlative degree	terdekat, terbesar, terpenting, terbaik
MD	Auxiliary verb	harus, perlu, boleh, adalah, mau
NN	Common noun	buku, pipi, rupiah, km, sekarang
NNP	Proper noun	Indonesia, MH370, Li Li, SBY
OD	Ordinal number	pertama, ketiga, ke-6
P	Particle	pun, -lah, -kah
PO	Preposition-object structure	untuknya, antaranya, olehku, padamu
PRD	Demonstrative pronoun	ini, itu, sini, sana
PRF	Reflexive pronoun	sendiri, diri
PRI	Indefinite pronoun	siapapun, apapun
PRL	Relative pronoun	yang
PRP	Personal pronoun	saya, kamu, dia, kami, kalian
RB	Adverb	sudah, tidak, sangat, juga
SC	Subordinating conjunction	baik...maupun..., sebelum, kalau
SP	Subject-predicate structure	katanya, sebutnya, tuturnya, imbuhnya
SYM	Symbol	+, %, @, $, 15/2/2017, 13:00
UH	Interjection	oh, hai, ya, sih, mari
VB	Verb	ada, melihat, gagal, menyoroti, main
VO	Verb-object structure	meningkatnya, terbentuknya
WH	Question	apa, siapa, mana, bagaimana
X	Unknown	yagg, busaway, saaat
Z	Punctuation	",.?"()

Table 1. Indonesian POS tagset

4. Models

POS tagging can be regarded as a sequence labeling problem. Given an input sequence $X = \{x_1, x_2, \dots, x_n\}$, where n is the length of X, the prediction model should output a sequence $Y = \{y_1, y_2, \dots, y_n\}$, in which each y_i is the label of x_i. In the case of POS tagging, Y refers to the POS sequence of an input sentence. State-of-the-art supervised models for handling the sequence labeling problem include conditional random fields (CRFs) (Lafferty et al., 2001), long short-term

memory (LSTM) (Huang et al., 2015; Lample et al., 2016; Reimers and Gurevych, 2017), etc. In this paper, we explore three models to achieve automatic POS tagging, namely CRFs, Bidirectional LSTM (Bi-LSTM) and sequence-to-sequence learning (seq2seq).

4.1 CRFs

CRFs are a type of discriminative undirected graphical model for labeling sequential data. For a linear chain CRF, given an input sentence s, the score of one of its possible label sequence l can be calculated through Equation 1:

$$sc(l|s) = \sum_{j}^{m} \sum_{i}^{n} \lambda_j f_j(l_{i-1}, l_i, i, s) \quad (1)$$

where i is the position of a word in the sentence, l_i is the label of the current word, l_{i-1} is the label of the previous word, f_j is the feature function, and λ_j is the feature weight. After having the scores of each possible label sequence, we can obtain the probabilities of these label sequences by exponentiation and normalization:

$$p(l|s) = \frac{\exp\left[sc(l|s)\right]}{\sum_{l'} \exp\left[sc(l'|s)\right]} = \frac{1}{Z(s)} \exp\left[sc(l|s)\right] \quad (2)$$

where $Z(s)$ is usually called the normalization factor.

4.2 Bi-LSTM

LSTM networks have been widely used in many sequence labeling tasks and show state-of-the-art performance. In this work, we employ the Bi-LSTM network with a CRF classifier (Fig.2, slightly adapted from (Reimers and Gurevych, 2017)). Its character representation is also derived from a Bi-LSTM network (Fig.3). A detailed explanation of the Bi-LSTM model can be found in (Huang et al., 2015; Lample et al., 2016).

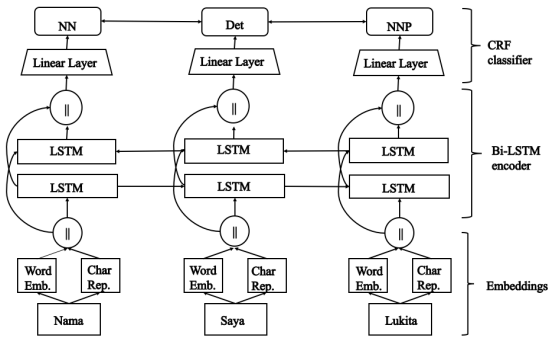

Figure 2. The Bi-LSTM network with a CRF classifier. (||) means concatenation.

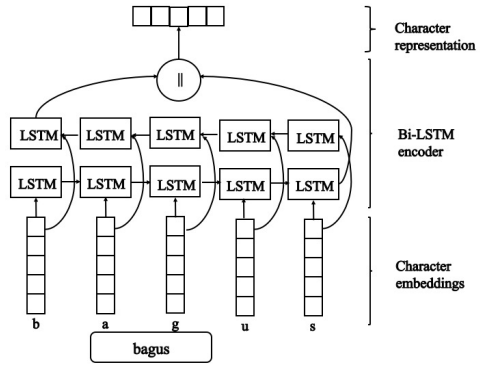

Fig.3 Character-based representation derived from the Bi-LSTM network. (||) means concatenation.

4.3 Seq2seq

Sequence to sequence learning has been successfully applied to machine translation (Wu et al., 2016) and text summarization (Nallapati et al., 2016). A popular approach is to encode an input sequence into a distributed representation with a bi-directional recurrent neural network (RNN) and decode the representation with another RNN, while the encoder and decoder are usually linked by the attention mechanism (Ghader and Monz, 2017), as Figure 4 shows. In this work, we also attempt to use the sequence-to-sequence architecture to perform sequence labeling.

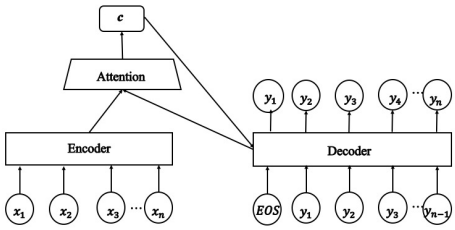

Figure 4. The sequence-to-sequence framework

5. Experiment

5.1 Setup

For CRFs, we used CRF++[6], a simple and open-source implementation of CRFs. Table 2 lists the feature set which obtained the best performance in our experiments, and we report the experiment result based on this feature set. Except for the *-c* (which was set as 3), other parameters are in accordance with the default settings.

Type	Feature	Description
Unigram	w_n $(n = -1,0,1)$	The previous n, current, and next n words
Prefix	$p_n(w_0)$, $n = 2,3,4$	The first n letters in the current word
Suffix	$s_n(w_0)$, $n = 2,3,4$	The last n letters in the current word
Bigram	$t(w_{-1})$	The predicted tag of the previous word

Table 2. The defined feature sets used in CRFs

[6] http://taku910.github.io/crfpp/

To implement the Bi-LSTM network, we used TensorFlow[7] version 1.2. For the setting of hyper-parameters, we referred to the suggestions of Reimers and Gurevych (2017). The pre-trained word embedding, with 200 dimensions, was trained on the Indonesian text (about 170 million tokens) that was crawled from several Indonesian news websites, using GloVe (Pennington et al., 2014). If a token does not occur in the vocabulary of the pre-trained word embedding, we would assign it a random word embedding (subject to a Gaussian distribution). In addition, the pre-trained word embedding is trainable during the training process. The dimension of character embedding is 100. The number of recurrent units for the Bi-LSTM layer which produces character representation (Fig.3) is 100, while for another Bi-LSTM layer (Fig.2) is 300. Adam was chosen as the optimizer. The dropout rate is 0.5. Also, we used a mini-batch size of 32 and employed the early stopping strategy if the score for development set does not increase for more than 3 training epochs. We report the result from the run with the highest score on development set.

As for the seq2seq architecture, we used NeuralMonkey [8], a convenient tool for quickly building sequential neural network models. It has implemented a framework for tagging. Therefore, the SentenceEncoder module was employed as the encoder, and the SequenceLabeler module as the decoder. Hyper-parameters are in accordance with the default settings.

5.2 Result

To compare the performance of different models, we employed the 10-fold cross validation. The corpus was divided into 10 folds. In each experiment, we used one fold for test and the remaining for training. Accuracies of different models can be calculated by comparing the manual tagging and the automatic tagging realized by these models. Table 3 shows the results. For each model, we report the average accuracy of 10 experiments.

Models	Avg Acc.
CRFs	95.12%
Bi-LSTM+CRF	95.68%
Seq2seq	94.14%

Table 3. The performance of different models

The highest average accuracy is produced by the Bi-LSTM network with a CRF classifier. CRFs perform slightly worse. It seems that Seq2seq is not competitive with the other two methods, but it takes the least time to train the model. Next, we will consider utilizing different encoders and decoders, and adding the attention mechanism to improve the performance of the Seq2seq architecture.

6. Conclusion

This paper describes our work on designing an Indonesian POS scheme and building a considerable Indonesian POS corpus using the text collected from multiple sources. In the design process, it is important to make a trade-off between universality and particularity. We put emphasis on those grammatical categories found cross-linguistically, following the specifications of Universal Dependencies, but would not miss those specific ones in Indonesian. During the annotation process, to tag a word, annotators need to consider its distribution and function in the context, not only the POS given by dictionaries. Finally, we propose an Indonesian POS tagset comprised of 29 tags and an Indonesian POS corpus of over 355,000 tokens, which could contribute to Indonesian language resources and provide support for further Indonesian NLP. Furthermore, we tried to achieve automatic POS tagging by using several state-of-the-art models trained on our corpus, and the experiment results are quite promising.

In future work, we intend to build a high-performance Indonesian POS tagger. Moreover, we would like to use the corpus to aid other Indonesian NLP tasks, such as chunking, syntactic parsing, etc.

Acknowledgements

This research was substantially supported by the National Natural Science Foundation of China (No. 61572145) and Department of Education of Guangdong Province. We greatly thank our annotators for their excellent wok.

References

Dinakaramani, A., Rashel, F., Luthfi, A., and Manurung, R. (2014). Designing an Indonesian part of speech tagset and manually tagged Indonesian corpus. In *Proceedings of the International Conference on Asian Language Processing 2014, IALP 2014* (pp. 66–69).

Ghader, H., and Monz, C. (2017). What does Attention in Neural Machine Translation Pay Attention to? *arXiv preprint arXiv: 1710.03348*.

Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional LSTM-CRF Models for Sequence Tagging. *arXiv preprint arXiv: 1508.01991*.

Kamus Besar Bahasa Indonesia. (2008). Kamus Pusat Bahasa, Jakarta, 4th edition.

Lafferty, J., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning* (Vol. 8, pp. 282–289).

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural Architectures for Named Entity Recognition. In *Proceedings of NAACL 2016*. San Diego, California.

Larasati, S. D., Kuboň, V., and Zeman, D. (2011). Indonesian morphology tool (MorphInd): Towards an Indonesian corpus. In *Communications in Computer and Information Science* (Vol. 100 CCIS, pp. 119–129).

[7] https://www.tensorflow.org/

[8] https://github.com/ufal/neuralmonkey

Nallapati, R., Zhou, B., Santos, C. N. dos, Gulcehre, C., and Xiang, B. (2016). Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)*.

Pennington, J., Socher, R., and Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1532–1543).

Pisceldo, F., Adriani, M., and Manurung, R. (2009). Probabilistic Part of Speech Tagging for Bahasa Indonesia. In *Proceedings of the 3rd International MALINDO Workshop, Colocated Event ACL-IJCNLP*.

Purwarianti, A., Saelan, A., Afif, I., Ferdian, F., and Wicaksono, A. F. (2013). Natural language understanding tools with low language resource in building automatic indonesian mind map generator. *International Journal on Electrical Engineering and Informatics*, *5*(3), 256–269.

Rashel, F., Luthfi, A., Dinakaramani, A., and Manurung, R. (2014). Building an Indonesian rule-based part-of-speech tagger. In *Proceedings of the International Conference on Asian Language Processing 2014, IALP 2014* (pp. 70–73).

Reimers, N., and Gurevych, I. (2017). Optimal Hyperparameters for Deep LSTM-Networks for Sequence Labeling Tasks. *arXiv preprint arXiv: 1707.06799*.

Santorini, B. (1990). Part-of-Speech Tagging Guidelines for the Penn Treebank Project (3rd Revision). http://doi.org/10.1017/CBO9781107415324.004.

Simons, G. F., and Fennig, C. D. (Eds.). (2017). *Ethnologue: Languages of the World*. SIL International, Dallas, Texas, 20th edition.

Tallerman, M. (2015). Understanding Syntax. Routledge, Taylor & Francis Group, London, 4th edition.

Wicaksono, A. F., and Purwarianti, A. (2010). HMM Based Part-of-Speech Tagger for Bahasa Indonesia. In *4th International MALINDO (Malay and Indonesian Language) Workshop*.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., … Dean, J. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv preprint arXiv: 1609.08144*.

A Tentative Idea of Multi-modal Corpus Applied to National Minority Chinese Proficiency Test in China[1]

Zhou Xuan, Chang Yinghao, Cao Deqing
Beijing Language and Culture University
No. 15, Xueyuan Road, Haidian District, Beijing
luoxiting2012@163.com

Abstract

To the masses of Minorities in the frontier areas of China, proficiency in the Chinese language tool is not only conducive to the common development and prosperity of all nationalities in our country, but also an important cornerstone for ensuring the smooth running of the "Belt and Road" strategy. However, for a long time, the primary means of monitoring Chinese learning in frontier minority areas is the Chinese Minority Chinese Proficiency Test (MHK). The disadvantage of this method lies in the fact that the test scores of written and oral exams are mostly rough and monotonous. They fail to fully assess the mastery of Chinese in minority areas and accurately reflect the integration of minority languages and Chinese. Therefore, this paper argues that there is an urgent need to establish a dynamic text corpus of Chinese Minorities in frontier areas, which can not only be used to monitor the use of Chinese and bilingual integration in local areas, but also be used in test selection, difficulty control, vocabulary development, and provide reference for the Minority Chinese test, provide more targeted suggestions for Chinese teaching in the "Belt and Road" region, and provide data support for the security of national language and character at the same time.

Keywords: Multi-modal Corpus ,Education Measurement, Language Testing

China is a multi-ethnic country which has many ethnic minorities. The Chinese learning situation of ethnic minorities in the border region matters whether the policy of national integration is successful. In recent years, with the promotion of China's international status, there are more and more researches on the popularization of Chinese.But most studies have focused on the spread of Chinesein foreign countries, rather than systematic monitoring of the use of Chinese in border areas.The promotion of Chinese is an important indicator of China's improvement in soft power and the mastery of Chinese of the border ethnic minority is a necessary condition for the internal political stability of the country.With the implementation of One Belt And One Road strategy, the languages of ethnic minorities near the border are more important since Chinese is the cornerstone of One Belt And One Road's economic and trade communication. In addition, once the One Belt And One Road strategy was carried out, the language security needs to be real-time monitored since opportunities of external exchanges for ethnic minorities will increase.

At present, the main means of measuring the mastering situation of Chinese for minority is a series of tests represented by National Minority Chinese Proficiency Test(MHK). MHK is a national standardized test built under the guidance of second language teaching theory and combined with the characteristics of ethnic minorities of Chinese learners in China. It mainly examines a test taker's ability of using Chinese in communication, especiallythe ability of study, work and social communications. The test items of MHK include listening comprehension, reading comprehension, written expression and oral expression. By testing different language skills, MHK can comprehensively test the ability of candidates to communicate in Chinese.Language testing is an important means to evaluate language ability, but a single report score does not specify the problem.The shortcoming of this approach is that the test scores can neither comprehensive evaluation of ethnic minority areas to master Chinese, also can't accurately reflect the fusion of the ethnic languages and Chinese. Therefore, this paper argues that minority nationalities Chinese corpus needs to be established, which can be used to monitor the local Chinese and bilingual fusion status and can also be used to helptest material selection, difficulty control, glossary revision and oral evaluation standarddevelopment for the National Minority Chinese Proficiency Test.

1. Corpus Linguistics Bibliometrics Analysis in Recent Ten Years

Bibliometrics is a quantitative method which studied the external characteristics of the literature and the output must be quantized information content. As a branch of library information science, mathematical and statistical methods are used by bibliometrics to describe, evaluate and predict the current situation and development trend of a subject. (QiuJunping, Wang Yue Fen, 2008: 1). In the past, bibliometric methods were mostly used in the field of natural science, and then gradually radiate to the humanities and social sciences. From the perspective of literature metrology, we take a quantitative research approach to the research of the subject area. From the metrology point of view, we can calculate the theoretical indicators of each subject in the field of literature and describe the developmenton a certain dimension of one subject.

1.1 The Object of Bibliometric Analysis andHighly Cited Papers

We can get hundreds of search results by selecting the "advanced search method" on online literature retrieval platform, taking "Philosophy and Humanities", "Social Science", "Information Technology" as the subject area,

[1] Foundation project: Beijing Language and Culture University's hospital level scientific research project (funded by the Central University's special research fund), "test credit scoring model based on big data test" (17YJ050008).

selecting CSSCI as a journal source, choosing the time span from 1998 to 2017 andtaking "corpus" as the key words to carry on the accurate retrieval. We finally retrieve 270 articles through the artificial screening, removing the non-disciplinary research articles such as "conference essay, meeting review, book review, notice and notice" in the result. With the help of bibliometrics, this paper analyzes the annual changes of the published papers and the data of the research topics, and calculates the frequency and proportion of different dimensions. In addition, according to the quoted frequency sort, we select the most cited citations from 1998 to 2017, and make a review of representative papers related on language test, thus outline the hot spots and trends of the research in the field of Chinese corpus in the past 20 years.

1.2 Changes of Yearly Quantity Published Articles

Quantity of published articles is the number of essays published under the specific conditions.Taking a year as a node of time and counting the number of published papers in each year in 10 years, we can see the intensity of the discipline research in the field of domestic Chinese corpora in recent years. Changes of yearly quantity published articles between 1999 and 2018 are as follows.

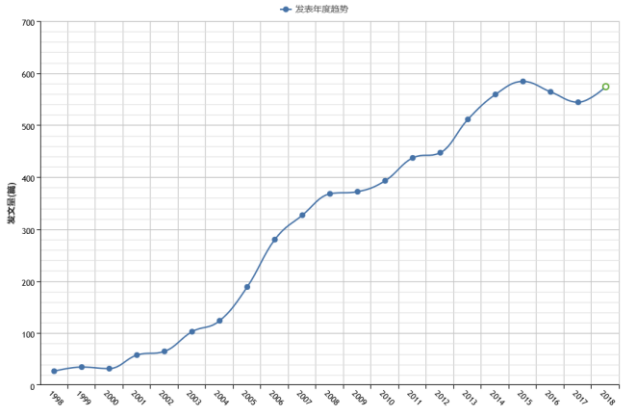

Figure 1: Changes of yearly quantity published articles between 1999 and 2018

From Figure 1, we can see that, the volume of articles involved in corpus research is on the rise during the 20 years from 1998 to 2017on the whole. Among them, there was a decrease from 2015 to 2017, an upward trend from 1998 to 2015 and a downward trend from 2015 onwards, indicating that 2009 is a watershed in the field of corpus linguistics. From the correlation between the number of published papers and the development of disciplines, we can roughly infer that the corpus linguistics volume increased steeply from 2015 to 2015, indicating that the discipline flourished during this period and was constantly making new breakthroughs period. After 2015, the number of published papers has picked up. It doesn't mean that corpus linguistics is no longer important. Actually, it is due to corpus linguistics has made breakthroughs in recent years and the fundamental difficulties have been overcome. As a result, the number of published papers has slowed down.

Among them, it is noteworthy that quantity published journal articles in 1998 was 4 and the number in 2017 and 2014 was about 540. The number of papers published during the six years from 2004 to 2009 increased by leaps and bounds every year, with an average increase of nearly 100 per year. Finally, the first small peak was ushered in in 2009, reflecting that these years are the spring of corpus research. However, it started to grow again for the second time in 2013. In 2015, the number of published papers on linguistic test reached its peak,which shows that the study of corpus linguistics still shows great vitality after 20 yearsdevelopment.

1.3 The Change of Hot Spots and Trends in Corpus and Language Testing Field

The distributive situation of the research topics can reflect the concentrated hot spots and development trends of a subject. By focusing on the concentrated distribution of the research topics over the past 20 years, we can find the research hot spots in the field of the corpus in recent years. From the vertical perspective, we can outline the development trend of corpus linguistics.

We used literature visual analysis tools to analyze 5950 retrieved essays which were used as samples. K-means clustering analysis was used to analyze and the threshold value was set as 50, and "key words" were analyzed for papers from 1999 to 2018. Finally, we get clustering results of class topics. By selecting the high volume of text from the amount of distribution of keywords and time extension analysis, we hope to outline the research hot topics and vein in the field of corpus focus.

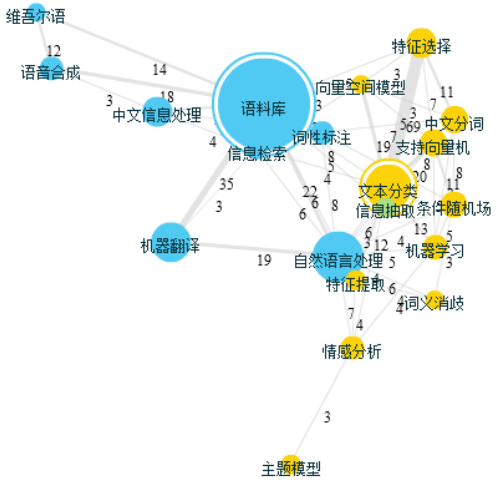

Figure 2: Clustering results of class topics of corpus

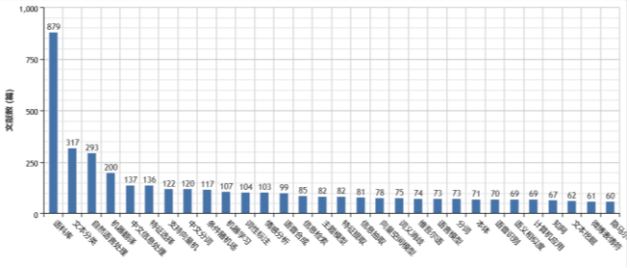

Figure 3: Research hot spots on corpus research

Since corpus linguistics belongs to a branch of natural language processing and computational linguistics, topics of natural language processing and Chinese information processing are removed fromretrieved essays. From figure 2, it can be found that the research focus of the corpus research mainly involves text classification, machine translation (more than 200 papers) in the past 20 years, followed by machine translation, machine learning, emotional analysis, speech recognition and other fields.

This phenomenon shows that corpus linguistics not only focus on the text corpus, but also focus on voice and other non-text corpus. Technically speaking, with the development of machine learning, more and more machine learning is used to deal with the urgent problems in corpus. Among them, there are 74 articles in Uyghur language, which indicate that the corpus has been explored in the minority languages, and Uyghur language is a very important part.

From the corpus bibliometrics analysis, we can see that the corpus researchhas made great progress in recent years, but the main concern is language ontology and corpus construction technology application. Scholars pay more attention to how language resources are collected,the way to collect, how to set up the corpus, the development of computer application technology.The research of application of corpus on teaching and language testing is relatively rare.Due to the particularity of the subjective test of language testing, the application of corpus technology is needed.

2. Application of Corpus in Language Testing

2.1 Corpus Auxiliary Language Test Proposition

As the first process of test formation, proposition often requires a lot of time and labor, is a very important but difficult part in language testing. Language test includes many sections, reading section is the longest and long-lasting part. In the proposition process of MHK, the collection of reading texts and the proposition of the item has become an indispensable part. However, in the long term item-constructed process we find that most of the corpus material used by the proponents is collected from the Internet. Although the network is more convenient and faster than the paper media collection, the quality of the network material is uneven. Finding the high quality corpus that meets the requirements in a short period of time is not an easy task. The propositional personnel often need to spend a lot of time to filter and modify the corpus. Compared with the proposition, the collection of corpus will take more time for the proposition, which makes the proposition inefficient and low efficiency.

In addition, compared to the collection of reading text in the process oftraditional language test proposition, the collection of the listening text is much more difficult, which makes the following questions are very common. Firstly, how to produce a text that meets the needs of everyday situations and the output of natural language flow, this would make the proposition staff often feel a headache. Secondly, if the text of the listening all rely on proposition personnel’s original creation, it would lead to dialogue’smismatch withthe logic of daily conversation and the output of natural language flowdue to the lack of mental capacity. Because of the shortage of high-quality audiometry texts and titles, the high repetition rate of conversation and dialogue content is a common problem in listening test texts, which may increase the exposure of topics in an untruthful manner, which is not conducive to the long-term and effective implementation of large-scale examinations. Third, it increases the workload of initial examination and the difficulty of work, resulting in the low utilization rate of language test talents.

In view of this, if multi-modal corpus for daily conversation of multiple scenes in accordance with linguistic norms can be established, annotated and analyzed in many aspects of sound and body posture, stored in different categories for retrieval and screening, it will greatly reduce the burden on the proposition stuff. At the same time, effectively improve the quality and speed of the proposition, reduce the basic trial links, and improve the utilization rate of language test talents.

2.2 Language Test Validity Argument and the Difficulty of Expansion

Validity is the most important indicator to ensure the quality of the test questions and test the validity of the test. For years, the validity of the language test is validated using the theoretical framework of validity test. For example, the validity of the occupational proficiency test demonstrates the validity of the test question by using the correlation between the test scores and academic equivalence criteria. Regardless of the validity model, the validity of the test questions is validated from the "post-test correlation" between the post-test score and a conventionally established standard. If the validity is not high, there is no way to retrieve. If we can control the validity of a test in a certain way before the question test, this is not only the expansion of the validity argument research in language testing, but also is a great help to grasp the validity of the test questions in advance, especially beneficial for the ethnic Chinese proficiency test.

Test difficulty is to ensure that the parallel test fairness of an important regulatory indicators. In the current MHK test, essay, oral and other subjective questions is to take a parallel test paper to examine the candidates. Candidates who take the test may do essay or oral exams on different topics, but it is difficult to assess the consistency of test questions. At present, the control of the difficulty of exam questions can only be controlled by experts and experienced proposition stuff based on experience which cannot be quantified. If we build a corpus based on the MHK language test, we can make statistics on the ability distribution of participants who participated in the MHK language test, observe the candidates' knowledge of knowledge points in different languages, and have an intuitive data support for the difficulty of the difficulty of the test questions Subjective questions difficult to predict and test the equivalent of the problem.

2.3 Promote the Basic Unity of Language Proficiency Evaluation Criteria

The study of language ability in the field of language testing has undergone three different stages: the stage of skill/component speaking, the stage of speaking of overall ability and the stage of establishing communicative competence model. However, the academic community has not formed a unified understanding and evaluation criteriaso far. As an unavoidable problem in the field of language testing, we all see each other's own ways to make language testing has long been unable to obtain a more authoritative test system of global recognition, but also to test the candidates caused great distress.

It seems that our discussion does not seem to make much sense if we just stay on verbal extermination. However, different scholars' ideas, disagreements and debates may be able to obtain fair judgment of third-party fair referees in this era. That is, using the platform of multi-modal corpus and utilizing the application of high penetration electronic terminals to collect and analyze high and low

level language proficiency people's written information and video information, through the actual observation of the differences in performance, to speculate on the essence of language ability and the appropriate evaluation criteria, the resulting language proficiency evaluation criteria will be more scientific. Moreover, the research results supported by the actual data will be approved by more experts and scholars and the general public, and promote the basic unification of the language proficiency evaluation standards.

The basic unification of language proficiency evaluation criteria and the requirement of quantitative assessment meeting the level of adaptive test capability will also provide theoretical basis and data support for the adaptability of language test.

2.4 Integration of Measurement, Study and Research

For many years, there has been a phenomenon of the test disconnect between language testing and language teaching. After a long period of learning and preparation, the improvement of language proficiency of many candidates cannot be tested by the test.The test results may reflect the candidate's language abilityin some aspects, may also ignore some aspects of the language skills of candidates, the test is not targeted.

This long-term state separation of study and measurement greatly reduces the effectiveness of teaching and testing, and the establishment of a contiguous and consistent multi-modal corpus will help to promote the consistency of learning tests, improve the effectiveness of language tests and enhance learning Testability of the study can be landed. We can establish a multi-modal corpus including image, sound and other forms of multi-modal corpus through teachingto increase students 'interest in learning, discover candidates' problems through testing, and promote teaching pertinence through questions.We can also collect information from multi-modal corpus about the correct and wrong data of students. The consistency oftest and teaching can help the test effectively distinguish between candidates with different levels of competence, reflect the improvement and decline of students. In this way, we can form a highly integrated and efficient test mode.

3. Suggestions for establishing corpus based on MHK

3.1 Collection of Language Resources Should Be True and Wide

The collection of the real corpus resources can not only come from the real corpus of the students in the test, but also from the first-line students' homework during the teaching. The resources should be extensive in order to ensure the effective application of the later corpus. In addition, it should be based on the different levels of difficulty and the order of collection should be successively.

3.2 Corpus Construction Must Be Targeted

Today, the formed corpus has been built, but the corpus that can be applied directly to the Chinese language test, especially the minority Chinese language test corpus, does not exist. The corpus construction must be well-targeted and able to solve the practical problems faced by ethnic minority Chinese proficiency testing. Otherwise, there is no need for construction.

3.3 Constructed Corpus Should Be Open

The use of many corpora is not open, which is not conducive to the development and updating of the corpus. Therefore, we suggest that the nature of the corpus is open and shared, and all the corpora need to be used by corpora in all fields. At the same time, the corpus should be dynamic to facilitate the timely updating of the corpus and the timely updating of the corpus can be realized automatically by the corpus individual users, so you can save the cost of lasting maintenance.Professional corpus maintenance staffs only need to be responsible for compliance with norms of audit and technical maintenance.

4. Conclusion

The application of corpus in the field of language testing is a multidisciplinary research subject. It requires experts in many fields such as linguistics, computer science and language testing to discuss together in order to clarify the needs and construction direction of the corpus. Researchers in the field of language testing provide concepts and requirements for corpus construction, computational linguistics researchers provide technical and ontological research support in order to ensure that the corpus is scientific and advanced.

At the same time, the corpus construction is not an overnight thing and requires scientific planning and effective implementation. Many problems such as how to obtain the operating cost of open corpus and how to coordinate the corpus development patent are all discussed. This article hopes to play a valuable role for later researches.

References

LijunChen,FanzhuHu (2010).Language Resource: A Tourism Resource urging Development.24(6), pp. 22-27.

Daming Xu(2008), Language Materials Management And Planning For Language Material Discussion. Journal of Zhengzhou University, pp. 12-15.

Dong Yongyi(2016). Study On The Relationship With Corpus And Language Testing. Language Construction, pp 87-88.

YunDuanNong(2008). How Corpus Be Applicated in Modern Language Testing. Examination Weekly, pp 218.

Liang Bo(2013). The Application of Mini-text in Compiling Of English Language Testing Materials. Journal of Wuhan Institute of Shipbuilding Technology, pp 86-89.

Lin Lin(2016). Study on the Application Of Corpus Linguistic and Corpus—Comment on the series of teaching practice series and corpus of foreign language teachers in national colleges and universities. News and writing, pp. 117.

Qin Peng(2007).Development and Application of Software Tools for Monitoring Language Resource of Print Media, Beijing Language and Culture University.

Wang Hui, Wang Yalan. 2016. Language Situation of "the Belt and Road" Countries. Language Strategy Research(2), pp. 13-19.

Construction of Uyghur named entity corpus

Maihemuti Maimaiti[1,2], Aishan Wumaier[1,2], Kahaerjiang Abiderexiti[1,2], Wanglulu[1,2], Wuhao[1,2], Tuergen Yibulayin[1,2]

[1] School of Information Science and Engineering, Xinjiang University, Urumqi, Xinjiang 830046, China
[2] Laboratory of Multi-Language Information Technology, Urumqi, Xinjiang 830046, China
mahmutjan@xju.edu.cn, hasan1479@xju.edu.cn, kaharjan@xju.edu.cn,wanglulu@stu.xju.edu.cn, 840315259@qq.com, turgun@xju.edu.cn

Abstract

The research of named entity recognition plays an important role in natural language processing (NLP) and can improve the performance of high-level NLP tasks, such as Machine Translation, Question Answering System, syntactic analysis and so on. Uyghur is a morphologically rich language with lack of manually created language resources. In this paper, we present the construction process of Uyghur named entity annotated corpus. This process composes of three steps. In the first step Chinese named entity recognition is used to select sentences from sentence aligned corpus and to extract Chinese named entities. In the second step Chinese-Uyghur named entity dictionary is automatically constructed using Chinese-Uyghur machine translation system, for the automatic pre-annotation of Uyghur named entities, and all annotations are corrected manually using an annotation tool. In the final step, corpus annotation quality is improved by using multiple strategies. As a result, four different sentence-level annotated corpora, person name annotated corpus, location name annotated corpus, organization name annotated corpus and personal, location, organization names annotated corpus, are constructed separately. To our knowledge, this is the first large-scale Uyghur named entity annotated corpus(UNEC) which is very valuable for the further researches.

Keywords: Uyghur, named entity, corpus

1. Introduction

With the development of the Internet, more and more text data appear. Information extraction has become an important direction in the field of natural language processing. Among them, named entity recognition (Coates-Stephens S. 1992; Thielen C. 1999) is a hot spot of information extraction proposed by MUC-6 conference (Sundheim B M. 1996), which is an important part of natural language processing. There are a large number of entities in the text, such as person names, location names and organization names. With the increase of text data, there are constantly appearing new named entities, and some of the named entities may be eliminated. It is almost impractical to construct a dictionary containing all named entities. Therefore, automatic recognition of named entities is an important task, other entities are easily treated as unknown words in the processing of natural language processing. Thus, affecting the performance of machine translation, knowledge map construction, Question Answering System, syntax analysis and other application areas. There are a large number of named entity annotation corpora in languages such as English (Sang E F T K et al. 2003) and Chinese (Walker C et al. 2006) at present. The named entity recognition technologies in many languages are relatively mature (Polifroni J et al. 2010; Savary A et al. 2010; Desmet B et al. 2013). But resource-deficient languages, like Uyghur, so far, no publicly available named entity corpus has yet to appear. On the one hand, it limits the research of Uyghur named entity recognition, on the other hand, it has some influence on the development of Uyghur information extraction technology.

Therefore, this paper firstly collected a large number of bilingual corpora in the field of news. It explores how to quickly establish a named entity corpus, for a resource-deficient language, by using cross-language named entity recognition technology. Firstly, named entities are automatically labeled for resource-rich languages. Secondly, sentence pairs with named entities are selected and pre-labeled for resource-deficient language sentences using bilingual named entity dictionaries. Finally, corrections and supplements were made manually, and annotation memory technology was used to further improve the efficiency and quality of annotation. Thus, Uyghur language location name annotation corpus, organization name annotation corpus, person name annotation corpus and Uyghur named entity annotation corpus are constructed respectively.

2. Related work

Compared with Chinese and English, the construction of the Uyghur Named entity corpus is very backward. In the study of Uyghur named entity and the construction of Uyghur annotated corpus, researchers have already made some contributions. Here are some important jobs, according to the features of Uyghur person name, this work built Uyghur person name annotated corpus that contained 5258 sentences (Rozi A et al. 2013). In the integrated recognition task which includes person name, organization name and location name, they constructed the a comprehensive named entity corpus that consists of 11257 annotated sentences (Tashpolat N et al. 2017). In addition, researchers constructed Uyghur music entity corpus which contains 2400 sentences in the research of Uyghur music entity recognition (Adila Ahmat et al.2017). The above are all Uyghur named entity corpus that have been reported. At present, the number of Uyghur entity annotated corpora is very small, and most of researchers used the data from the internet to research by rule-matching (Arkin M et al. 2013; Maihefureti et al. 2014; Mahmoud A et al. 2017). However, the research of the named entity recognition by machine learning (Yang Y et al. 2011; Jiazheng L I et al. 2011; Abiderexiti K et al. 2017) can't be separated from the standard data resource. So it is necessary to build standard data resource of named entity.

3. Creating the data resource

The construction of a corpus requires not only a large amount of data resource, but also expensive manpower. The details of construction are as follow.

3.1 Data Source of corpus

Based on the Uyghur-Chinese parallel corpus of the 13th China Workshop on Machine Translation (CWMT2017), we also use the Uyghur-Chinese parallel corpus provided by the Laboratory of Xinjiang Multilingual Information Technology for manual annotation

3.2 Annotation specification

In the processing of manual annotation, for all corpora, we use three kinds of tags, Person (PN, personal name), Location (LN, place name) and Organization (ON, organization name). Due to one sentence may contain two or more tokens, this paper adopts annotation specification named BIO (begin-in-out) proposed by (Ramshaw L A et al. 1995). The tagging sets contains 7 kinds of labels. Table 1 shows this tag sets:

label	Meaning
O	Non entity words
B-PER	The first word of person name or person name of a single word
I- PER	Not the first word in the person name
B-LOC	The first word of location name or location name of a single word
I-LOC	Not the first word in the location name
B-ORG	The first word of organization name or organization name of a single word
I-OR	Not the first word in the organization name

Table 1: Tag sets of named entity in Uyghur

This is an example to describe an original sentence and tagged sentence , as Table 2 shown:

Original sentence	Shi Jinping Bëyjingda Birleshken Döletler Teshkilatining bash katipi Ban Kimon bilen körüshti . (Translation : Xi Jinping met with UN Secretary-General Ban Ki-moon in Beijing.)
Tagged sentence	【Shi Jinping PN】【Bëyjingda LN】【 Birleshken Döletler Teshkilatining ON】 bash katipi 【Ban Kimon PN】 bilen körüshti .
After tagging	Shi/B-PER Jinping/I-PER Bëyjingda/B-LOC Birleshken/B-ORG Döletler/I-ORG Teshkilatining/I-ORG bash/O katipi/O Ban/B-PER Kimon/I-PER bilen/O körüshti/O ./O

Table 2: An examle of tagged sentence

3.3 Annotation method and processing

In order to reduce the work of manual annotation, we used a method of man-machine combination to improve the speed of tagging and guarantee the quality of corpus at the same time. Processing of construction incorporated three steps: preprocessing, tagging, post-processing. The whole processing is showed as Figure 1.

3.3.1 Pre-processing

Bilingual sentence deduplication processing: Due to the bilingual sentence alignment corpora is very large, there may be repeated sentences. To avoid re-labeling the same sentence, it is necessary to remove the repeated sentences. There are two steps, the first step is to remove all the punctuation, only to reserve Chinese character, then removing the repeated sequence of Chinese characters. The second step is to do the same operation in Uyghur.

Chinese named entity annotation: After the processing of bilingual sentence deduplication, the NLPIR[1] system was used for construction of the location name annotated corpus, the other named entity corpus used Pyltp[2] system which from Harbin Institute of Technology. The functions of these two systems incorporate Chinese word segmentation and Named entity annotation.

Entity extraction and sentence filtering: In order to reduce the manual checking time and improve the speed of tagging, we filter out sentences which not incorporate person name, location name and organization name. Then, corpus only contained the Uyghur sentences which prepare to tag and the corresponding Chinese named entity list which include entity labels.

3.3.2 Name entity tagging

Chinese and Uyghur name entity dictionary automatic construction: First, we redo the Chinese named entity list and construct the entity dictionary in descending order by the occurrence frequency. Our group's Chinese-Uyghur machine translation system had been used to translate this dictionary and generated a Bilingual entity dictionary. There are some problem in the dictionary, such as incorrect translation, empty translation result and translation result include a variety of additional components. Therefore we manually review the translation result and correct the entity dictionary.

Named entity automatic annotation: According to the established Chinese-Uyghur named entity dictionary, all Uyghur sentences had been tagged for the first time. In order to prevent the spread of errors, the Uyghur entities corresponding to the Chinese entity have been tagged based on Chinese entities list. Since named entity in Uyghur sentences may be variants that attached some additional component, we used a method of fuzzy matching.

Named entity manual proofreading: Through automatic tagging, part of Ughur sentences has been tagged well. There are still many problems in the result of automatic annotation, such as error of Chinese named entity recognition, the source sentence alignment problem (the content is not aligned, the translation longer or less than the original sentence, etc.) and variants of the named entity result in matching error. Due to these problems, all the automatic annotated Uyghur sentences were imported in manual annotation system and conducted manual

[1] https://github.com/NLPIR-team/NLPIR

[2] https://pypi.python.org/pypi/pyltp

proofreading and correction. The same named entity would be ignored to improve the speed of tagging and ensure the consistency of tagging.

3.3.3 Post-processing

Since multiple people tag at the same time, it is difficult to avoid mistakes just like annotation error, annotation inconsistency, leakage of annotation, etc. Based on the manual annotated corpus, we established Uyghur named entity dictionary and proofread each entity by using source sentences. After that, the CRF annotation machine was trained with all the language materials, and the corpus has been automatically tagged. Other possible error would be corrected by comparing with manual annotation and automatic annotation. Finally, we derive the named entity corpus in the format of the above tagged sentence.

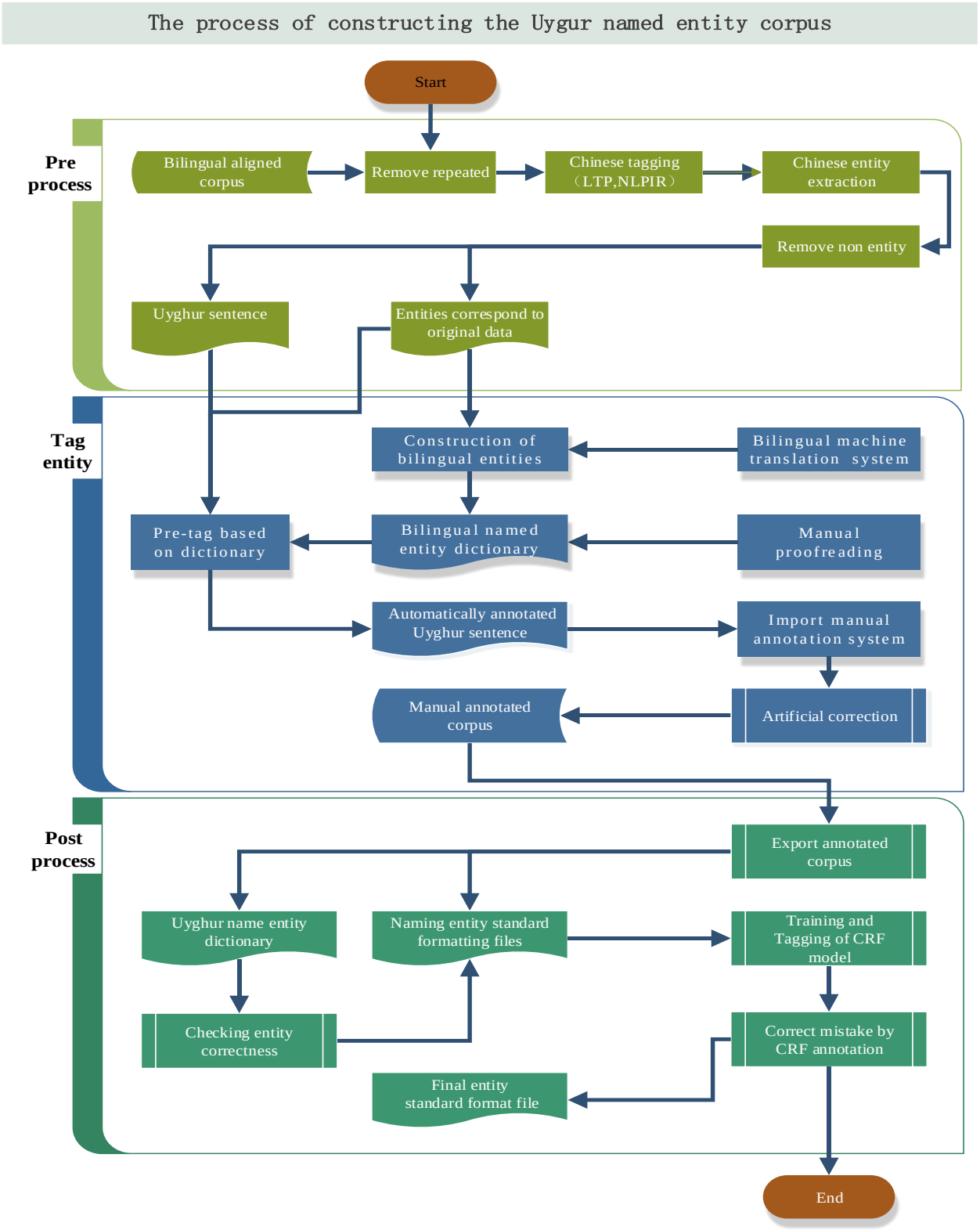

Figure 1: Flow chart of the construction of Uyghur named entity corpus.

4. Data analysis

The whole corpus resources contains location name annotated corpus, person name annotated corpus, organization name annotated corpus and a comprehensive named entity corpus include person name, location name, organization name.

4.1 Location name annotated corpus

Location name annotated corpus contains 13385 sentences with 20218 place names, and the number of location names appeared 41009 times. The statistics are shown in Figure 2. Most of location names are consist of one or two words, total occupying 93% of the location name number, and their number are 8506 and 10296 respectively, shown in Figure 3.

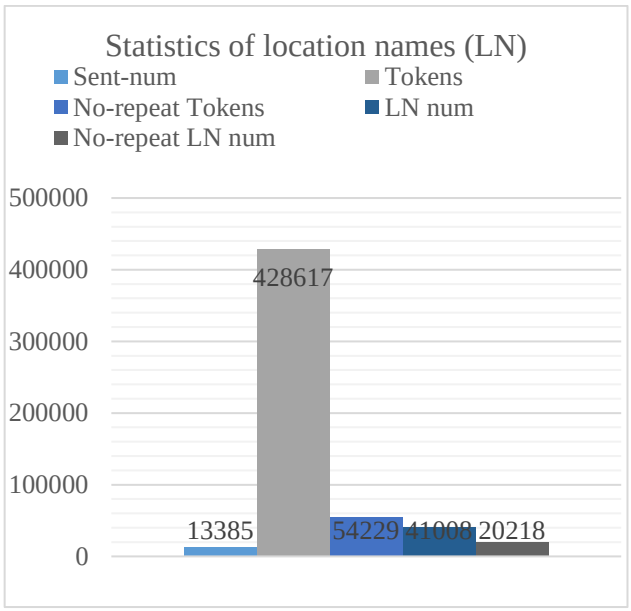

Figure 2: Location name annotated corpus statistics.

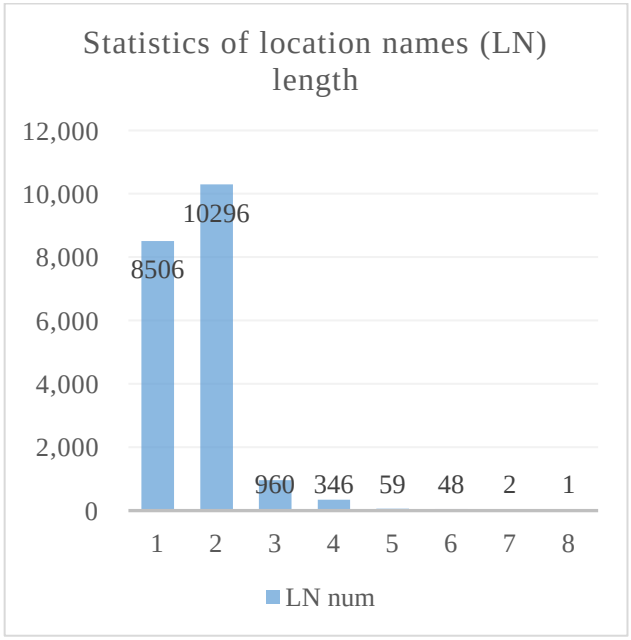

Figure 3: Location name length statistic.

4.2 Organization name annotated corpus

Organization name annotated corpus contains 11337 sentences with 9733 organization names, and the number of organization names appeared 13436 times. The specific statistics are shown in Figure 4.

The statistic of organization name annotated corpus found that the length of most organization name is between 2 and 11 words. Compared with the location name, the length of organization name is longer and the distribution is wider. The specific statistics are shown in Figure 5.

4.3 Person name annotated corpus

Person name annotated corpus contains 21078 sentences with 18066 person names， and the number of person names appeared 34598 times. As shown in the Figure 6:

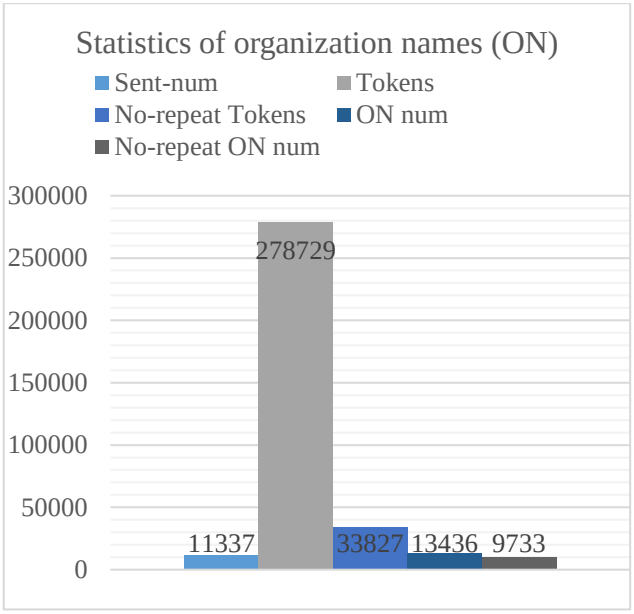

Figure 4: Organization name annotated corpus statistics.

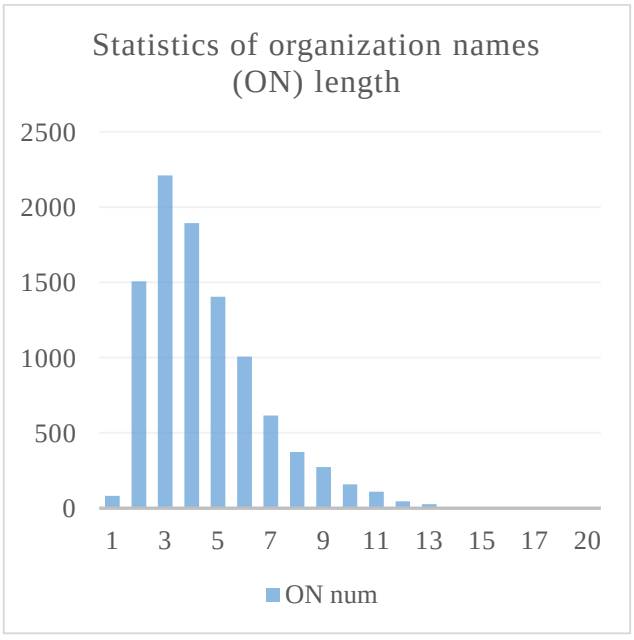

Figure 5: Organization name length statistic.

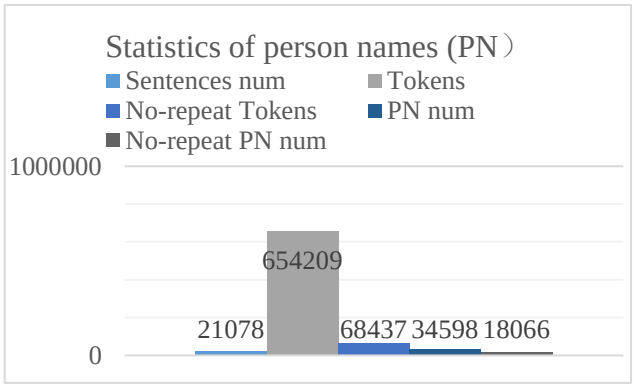

Figure 6: Person name annotated corpus statistic.

Most of person name are consist of one or two words, their proportion is high up to 99%. The Figure 7 show this situation.

4.4 Named entity annotated corpus

Uyghur named entity corpus which is comprehensive corpus contains 39027 sentences. The number of entities is 102360(include repeating entities). The person name appeared 28469 times, account for 27.8%, the location name appeared 42585 times, account for 41.6%, and organization names appeared 31306 times, account for 30.6%. The specific statistics are shown in Figure 8.

Figure 9 shows the length of different named entity in corpus. As can be seen from the statistics, the length of person name more than two words is rare; the length of location name is almost between 1 to 6 words, the number of location name was decrease with increase entity length; the length of organization name is mainly between 2 to 12 words, its length is relatively long.

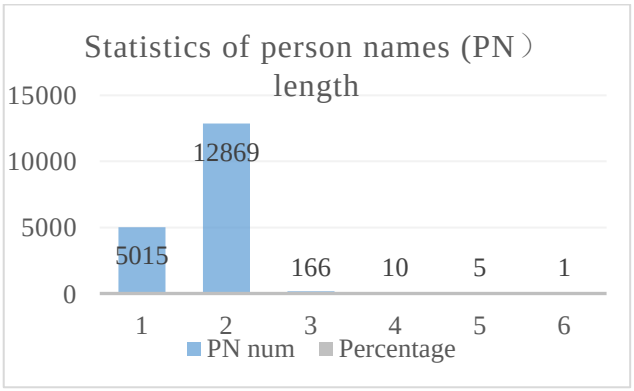

Figure 7: Person name length statistic.

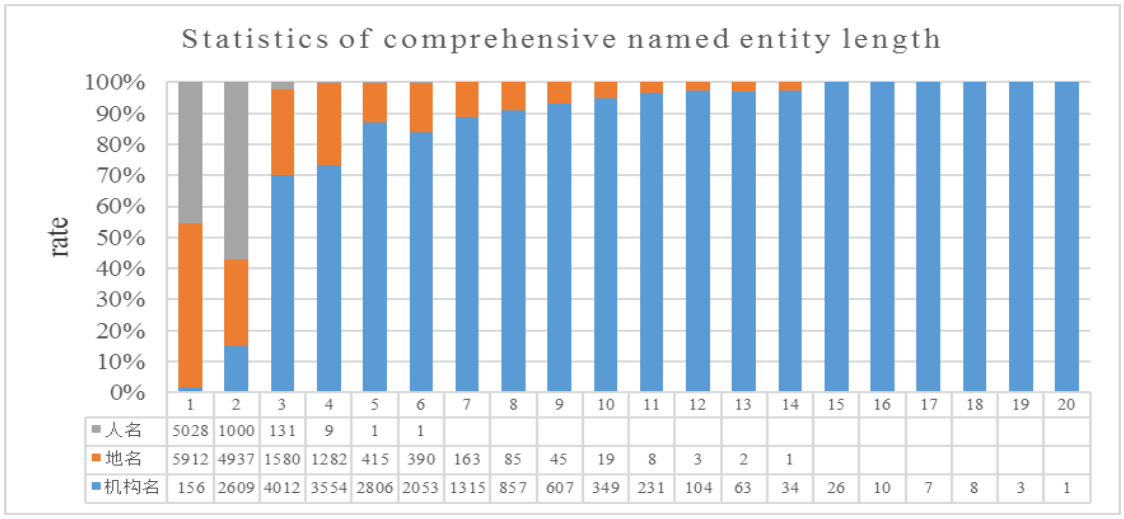

Figure 8: Statistics of Uyghur named entity corpus.

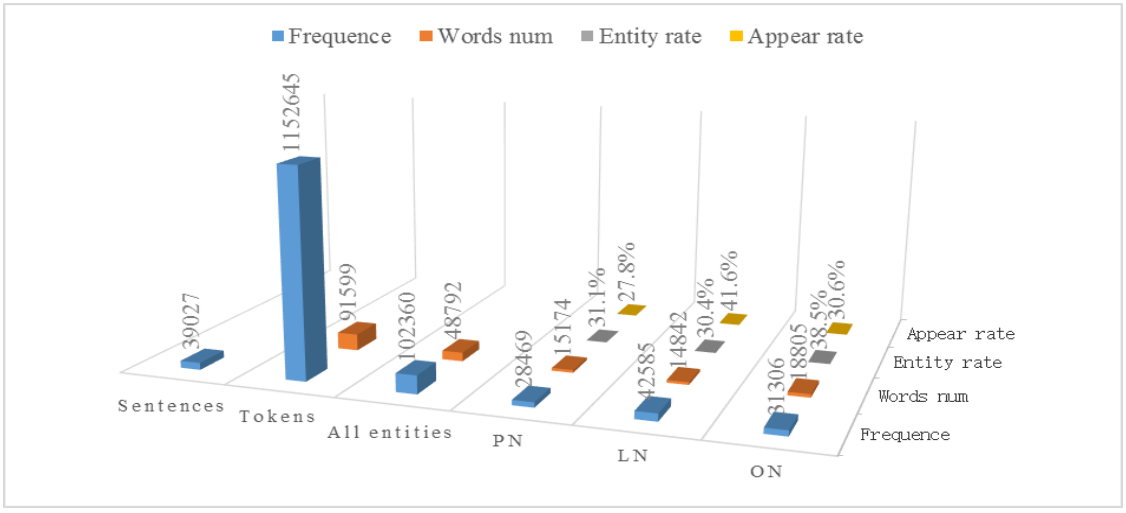

Figure 9: Uyghur named entity corpus length statistic.

5. Conclusions

Our paper used bilingual sentence alignment corpus and Chinese named entity recognition technology to establish person name annotated corpus, location name annotated corpus, organization name annotated corpus and named entity annotated corpus in Uyghur. During the construction of corpus, it is very useful to build bi-lingual entity dictionary, automatic annotation, annotation memory, error analysis and so on. By using a human-machine combination method which greatly reduces the cost of human resources and effectively guarantees the quality of the corpus.

With the completion of constructing corpus, our work has laid a solid foundation for the next research on Uyghur named entity recognition, and it can also play a positive role in the further research of Machine Translation, information extraction, syntactic analysis, semantic analysis and so on.

6. Acknowledgments

This work has been supported as part of the NSFC (61462083, 61762084, 61463048, 61262060), 973 Program (2014cb340506), and 2017YFB1002103, ZDI135-54.

7. References

Coates-Stephens S. 1992. The analysis and acquisition of proper names for the understanding of free text. Computers & the Humanities, 26(5-6), 441-456.

Thielen C. 1999. An approach to proper name tagging for german.

Sundheim B M. 1996. Overview of results of the MUC-6 evaluation. A Workshop on Held at Vienna, Virginia: May (pp.423-442). Association for Computational Linguistics.

Sang E F T K, Meulder F D. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. Conference on Natural Language Learning at Hlt-Naacl(Vol.21, pp.142-147). Association for Computational Linguistics.

Walker C, Strassel S, Medero J, Maeda K. 2006. Ace 2005 multilingual training corpus. Progress of Theoretical Physics Supplement, 110(110), 261-276.

Polifroni J, Kiss I, Adler M. 2010. Bootstrapping Named Entity Extraction for the Creation of Mobile Services. International Conference on Language Resources and Evaluation, Lrec 2010, 17-23 May 2010, Valletta, Malta. DBLP.

Savary A, Waszczuk J, Przepiórkowski A. 2010. Towards the Annotation of Named Entities in the National Corpus of Polish. International Conference on Language Resources and Evaluation, Lrec 2010, 17-23 May 2010, Valletta, Malta. DBLP.

Desmet B, Hoste V. 2013. Towards a Balanced Named Entity Corpus for Dutch. International Conference on Language Resources and Evaluation, Lrec 2010, 17-23 May 2010, Valletta, Malta (pp.535-541). DBLP.

Rozi A, Zong C, Mamateli G, Mahmut R, Hamdulla A. 2013. Approach to recognizing uyghur names based on conditional random fields. Journal of Tsinghua University, 53(6), 873-877.

Tashpolat N, Wang K, Askar H, Palidan T. 2017. Combination of statistical and rule-based approaches for uyghur person name recognition. Zidonghua Xuebao/acta Automatica Sinica, 43(4), 653-664.

Adila Ahmat, Feng Xiangping. 2017. Cognition of Uyghur musical named entity based on condition random Field. [j]. Intelligent Computer and Applications.

Arkin M, Hamdulla A, Tursun D. 2013. Recognition of uyghur place names based on rules. Communications Technology.

Maihefureti, MiriguRouzi, MaierhabaAili, TuergenYibulayin, Department T A. 2014. Uyghur organization name recognition based on syntactic and semantic knowledge. Computer Engineering & Design.

Mahmoud A, Yusuf H, Zhang J, Zong C, Hamdulla A. 2017. Name recognition in the uyghur language based on fuzzy matching and syllable-character conversion. Qinghua Daxue Xuebao/journal of Tsinghua University, 57(2), 188-196.

Yang Y, Xu C, Gulimire A. 2011.uyghur named entity identification basing on maximum entropy model[C]// 3rd International Conference on Information Technology and Computer Science.

Jiazheng L I, Liu K, MairehabaAili, Yajuan L V, Liu Q, TuergenYibulayin. 2011. Recognition and translation for chinese names in uighur language. Journal of Chinese Information Processing, 25(4), 82-87.

Abiderexiti K, Maimaiti M, Yibulayin T, Wumaier A. 2017. Annotation schemes for constructing Uyghur named entity relation corpus. International Conference on Asian Language Processing. IEEE.

Ramshaw L A, Marcus M P. 1995. Text chunking using transformation-based learning. Text Speech & Language Technology, 11, 82--94.

UM-PCorpus: A Large Portuguese-Chinese Parallel Corpus

Lidia S. Chao[†], Derek F. Wong[†], Chi Hong Ao[†], Ana Luísa Leal[‡]
[†]NLP^{2}CT Lab / Department of Computer and Information Science
University of Macau, Macau SAR, China
[‡] Department of Portuguese, University of Macau, Macau SAR, China
{lidiasc, derekfw, analeal}@umac.mo, nlp2ct.benao@gmail.com

Abstract

This paper describes the creation of a high quality parallel corpus for Portuguese and Chinese that extracted from parallel and comparable documents. The corpus is constructed using an on-line alignment platform, UM-pAligner. The UM-pAligner consists of two main alignment components, parallel sentence identification and classification model, for acquiring the parallel sentences from either the parallel or comparable texts in a semi-automatic manner. The extracted parallel sentences are manually verified. The resulting corpus is composed of the parallel sentences covering the texts of the newswire, legal, subtitle, technical and general on-line publications, around 6 million parallel sentences. About 1 million parallel sentences are compiled and made available for download at the NLP^{2}CT website.

Keywords: Portuguese-Chinese, parallel corpus, machine translation, alignment platform, UM-pAligner

1. Motivations

Parallel corpora are valuable resources for linguistic research (McEnery and Xiao, 2007) and natural language processing, in particular a sentence-aligned parallel data has been the main source for the development of neural machine translation (NMT) systems (Sutskever et al., 2014; Yang et al., 2017; Xu et al., 2017). Despite many corpora have been created and published (Koehn, 2005; Steinberger et al., 2006; Smith et al., 2013; Tian et al., 2014; Ziemski et al., 2016), the construction and exploiting of parallel corpora that paired with English still dominate the research for machine translation (MT) (Bojar et al., 2014). Corpora of other language pairs are relatively rare (Post et al., 2012; Chu et al., 2014) and not always available in large enough quantities to build an MT system with good quality (Kolachina et al., 2012). This is typically referred as low-resource language pair characterized by the amount of parallel data available for training an MT model.[1] Hence, the creation of parallel corpora is an important step to drive the MT research for a language pair. Despite Portuguese and Chinese are two of the top ten most influential languages in terms of populations (Weber, 1999) and information production (Lobachev, 2008), they are categorized as a low-resource language pair in the field of MT. According to our knowledge, there is no a high quality and large parallel corpus publicly available for Portuguese-Chinese. One notable parallel corpus is the OpenSubtitles (Lison and Tiedemann, 2016) that has been released through the Open Parallel Corpus (OPUS) project.[2] The OpenSubtitles is mainly compiled from the movie and TV subtitles, consisting of 2.6 billion sentences for more than 60 languages, including around 6.7 million of parallel sentences for Portuguese and Chinese. However, the construction of the OpenSubtitles corpus is completely automatic and the extracted parallel sentences are not manually verified by the bilingual expert. In addition, the corpus is a kind of spoken data, making the content "too narrow". Another parallel corpus regarding Portuguese-Chinese is the parallel treebank released by the University of Macau Xing et al. (2016). The corpus consists of 500 texts of the newswire. The parallel sentences are syntactically annotated. The alignments of inter-nodes between the pair of syntactic structures are linked at both the word and phrase level. However, the corpus is relatively small and is not suitable for training an MT. In this work, we intend to construct a large parallel corpus for Portuguese and Chinese. The corpus is complied from different text domains and genres, including newswire, legal, subtitle, technical as well as the official publication of Macau government agencies.[3] The corpus contains 6 million parallel sentences, and about 1 million of which are released to the public for research purposes.[4] The remainder of this paper is organized as follows. Section 2 presents the models, methods and the platform for the construction of parallel corpus. The content and analysis of the created parallel corpus are described in Section 3. The MT experiments on the parallel corpus are conducted in Section 4, followed by the conclusions to end the paper.

2. Methods

In the present work, the UM-PCorpus is designed to be a multi-domain parallel corpus, which embraces texts of different genres (or domains). This serves as an important mean to the research of domain adaptation in MT (Wang et al., 2014; Wong et al., 2016). In this version, the corpus consists of newswire stories, subtitles, legal articles, IT documents and the official publications of Macau government departments and agencies. Another concern regarding this construction is the alignment quality of the parallel data. The data sources are manually identified and carefully selected. In crawling the data, a number of checks are performed in order to ensure the parallelism. The articles are removed if either their length ratio (or the ratio

Corresponding author: Derek F. Wong

[1] We consider the language pairs with parallel datasets less than 1 million sentences as low-resource.

[2] http://opus.nlpl.eu/

[3] https://www.gov.mo/en/

[4] http://nlp2ct.cis.umac.mo/um-corpus/index2.html

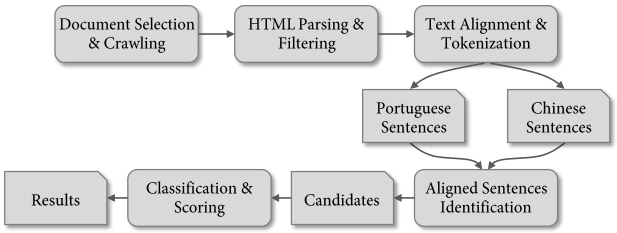

Figure 1: The construction procedure of UM-PCorpus.

of sentences) is beyond the threshold. Figure 1 depicts the work flow of the construction of the parallel corpus (Tian et al., 2014). The whole process consists of the selection and crawling of the on-line bilingual documents, parsing the HMTL files and performing heuristic checks to discard any unaligned files, splitting the text into sentences (Wong et al., 2014) and tokenizing those of the Chinese text (aka Chinese word segmentation (Zeng et al., 2013a)), the identification of parallel sentences and finally scoring the candidates to determine the final parallel sentences.

2.1. Neural Network Based Alignment Identification

The parallel documents are valuable sources for inducing the aligned sentences, however, it is relatively very rare for Portuguese-Chinese when comparing with English-Chinese, English-French and those between European languages (Callison-Burch et al., 2012). In contrast, comparable documents are more readily available in larger quantities than the parallel documents. To this end, we first propose a parallel sentence identification model based on semi-supervised orthogonal denoising autoencoder (Ye et al., 2016; Leong et al., 2018), under the framework of multi-view learning, to retrieve the possible aligned sentences based on their semantic meaning (i.e. distributed representation) (Wong et al., 2016) instead of the symbolic

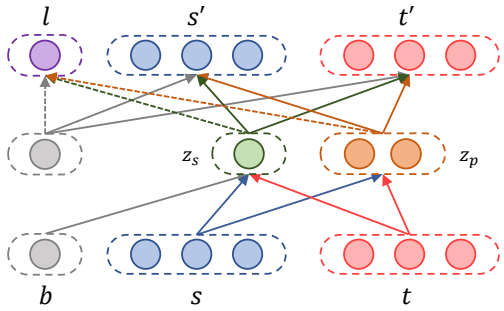

Figure 2: Semi-supervised orthogonal denoising autoencoder. The representations of source sentence s and target sentence t are being treated as different input views. The private and shared latent spaces, z_p and z_s represent the common features shared by both sentences and the private features owned by individual sentence. The s' and t' are the reconstructed representations of the source and target sentences, while l is the prediction label.

features (i.e. dictionary, word alignments, etc.) (Zamani et al., 2016). The architecture of the proposed model is depicted in Figure 2.

Formally, given a concatenated representation vector $\mathrm{x} = \{x_1, \ldots, x_m, x_{m+1}, x_n\}$ of a source sentence x_s and its target sentence x_t, an autoencoder aims to transform it to a hidden space $h = s(Wx + b)$, and the hidden representation h is subsequently transformed back to its reconstructed vector $\mathrm{x}' = g(W'h + b')$ through the activation functions $s(\cdot)$ and $g(\cdot)$ with the weight matrices W and W', and the bias b and b'. The objective is to learn the model parameters that minimizes the reconstruction error $\ell(\mathrm{x}, \mathrm{x}')$, where $\ell(\cdot)$ is a loss function to measure how good the reconstruction performs. To accommodate the shared and private latent spaces in the context of multi-view learning, the autoencoder model is revised to connect only the private latent space z_p to its original input view, and disconnect it from the other views, such that the private latent spaces are independent from each other. While the shared space z_s is connected to all of the input views, i.e. the representations of the source and target sentences (Leong et al., 2018). To maintain the orthogonality of the private spaces, the bias is disconnected from the private spaces (Ye et al., 2016). Formally, $I(A|B)$ is defined to denote the indices of columns of matrix A in terms of the matrix B if A is a submatrix of B. The orthogonal constraints on weights are defined as follows:

$$W_{I(z_p^{v_2}|[z_s,z_p]),I(\mathrm{x}^{v_1}|\mathrm{x})} = 0$$

$$W'_{I(\mathrm{x}^{v_1}|\mathrm{x}),I(z_p^{v_2}|[z_s,z_p])} = 0,$$

where $\mathrm{v} = \{v_1, ..., v_k\}$ denote the different views of an input x, z_s is the shared latent space and $\mathrm{z}_p = \{z_1, ..., z_k\}$ are the corresponding private spaces of different views v. The denoising autoencoder was originally proposed to enforce the autoencoder in learning robust features (Ye et al., 2016). In our task, we want the model to be able to learn the latent features which are best to distinguish if a pair of sentences are the translations of each other. To this extend, we further modify the model to guide the training towards this objective. The latent spaces are leveraged by adding a feed-forward NN layer in addition to the reconstruction layer, and defined as:

$$l = \sigma(W_l[z_s, z_p] + b_l),$$

where $\sigma(\cdot)$ is the sigmoid function, W_l and b_l are the weight matrix and the bias. The model parameters are optimized by minimizing the loss function:

$$J = \alpha J_{rec} + (1 - \alpha) J_{label},$$

where J_{rec} and J_{label} are the reconstruction and cross-entropy loss. The hyper-parameter α is used to weight the reconstruction and cross-entropy error in controlling the preference of the learned model:

$$J_{label} = \frac{1}{n} \sum [l' \log(l) + (1 - l') \log(1 - l)]$$

$$J_{rec} = \frac{1}{2n} \sum ([\mathrm{x}_s; \mathrm{x}_t] - [\mathrm{x}_{s'}; \mathrm{x}_{t'}]).$$

Language	Avg. Length	Tokens	Vocabulary
Chinese	14.39	88,197,691	334,223
Portuguese	16.41	100,581,355	425,300

Table 1: Statistics of the UM-PCorpus

2.2. Maximum Entropy Based Classification

To complement the autoencoder model that uses continuous real-valued embeddings to represent sentences, we also develop a conventional maximum entropy (MaxEnt) classification model which uses the discrete features, either the symbolic or numeric features. Previous works have also shown the effectiveness of using a MaxEnt model in parallel corpus construction (Munteanu and Marcu, 2005) and many natural language processing applications (Berger et al., 1996; Wong et al., 2009; Zeng et al., 2013b). For our classification problem, the model is defined as:

$$p(c|\mathrm{s},\mathrm{t}) = \frac{\exp(\sum \lambda_i f_i(l,\mathrm{s},\mathrm{t}))}{Z(\mathrm{s},\mathrm{t})},$$

where $p(c|\mathrm{s},\mathrm{t}) \in [0,1]$ is the probability where a value close to 1.0 indicates that the paired sentences are translations of each other, $l \in (0,1)$ is a class label representing where the sentences (s,t) are parallel or not parallel, $Z(\mathrm{s},\mathrm{t})$ is the normalization factor, f_i are the feature functions, and λ_i are the feature weights to be learned. The features we considered in this task include the length-based features (Gale and Church, 1993), alignment-based features (Munteanu and Marcu, 2005) and the anchor texts (Patry and Langlais, 2011).

2.3. UM-pAligner Platform

We integrate the proposed models and implement an on-line parallel sentence alignment platform, UM-pAligner. The platform currently supports Portuguese and Chinese languages only, and it is publicly available at the NLP^{2}CT website.[5] Besides the underlying proposed methods, one notable function of the alignment platform is that the alignments between sentences of the inputs are presented in terms of an alignment matrix. The entry is indexed by the a pair of source and target sentences. The score in an entry is the weighted model score given by the orthogonal denoising autoencoder and the MaxEnt models. The GUI interface allows a data annotator to easily identify and verify the alignments between the sentences.

3. The UM-PCorpus

The constructed corpus consists of texts that collected from various sources of different text genres, covering different topics. According to the text genres, we categorize the types of texts into five different domains in a more general way:

- **News**: This data contains the stories of Macau news. Those are good sources of high quality text for Portuguese and Chinese. The data are collected from the on-line Macau newspapers and the news articles published by the Macau government agencies,[6] from 2005 to 2015. More than 30,000 articles have been collected, consisting of approximately 296,000 sentences.

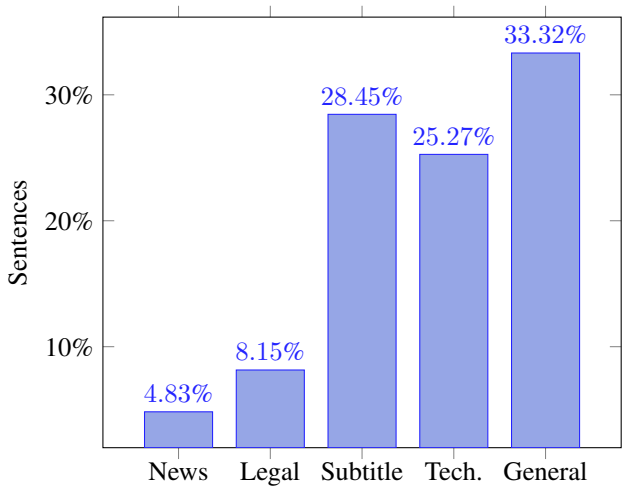

Figure 3: Distribution of data among different domains.

- **Legal**: This collection of texts is compiled from the ordinances and subsidiary legislation of Macau Special Administrative Region (SAR), and other relevant instruments published by the Macau Legal Affairs Bureau[7] and the Macau Printing Bureau.[8] There are 16,689 articles, consisting of about 0.5 million sentences.

- **Subtitle**: The subtitles are the transcriptions of spoken languages. This data is mainly extracted from the subtitles of the TED Talks[9] and the movie subtitles of the OpenSubtitles[10]. After a number of checks and proofreading, around 1.7 million high quality parallel sentences are selected and included in this corpus.

- **Technical**: The technical data is composed of the documents regarding computer software and hardware instructions, as well as the content collected from the technical forums of IT companies. This type of data constitutes about 25% of the corpus.

- **General**: For those of texts that cannot be put into one of the above domains are categorized as general domain, due to their very different sources and the small amount of parallel data it has. This collection of texts is comprised of the websites and the official publications of the Macau government departments and agencies, as well as the high quality parallel sentences extracted from the Wikipedia,[11] using the proposed methods (as described in Section 2.1.)

[5] https://nlp2ct.cis.umac.mo/NMT/aligner

[6] The Macau SAR Government Portal: https://www.gov.mo/

[7] http://www.dsaj.gov.mo/

[8] http://www.io.gov.mo/

[9] https://www.ted.com

[10] https://www.opensubtitles.org

[11] https://www.wikipedia.org/

		Chinese			Portuguese		
Domain	Sentences	Average Length	Tokens	Vocabulary	Average Length	Tokens	Vocabulary
News	146,095	28.40	4,148,669	69,691	36.00	5,259,712	65,462
Legal	173,420	18.92	3,280,904	77,081	21.22	3,680,346	77,701
Subtitle	250,000	9.16	2,289,436	48,842	10.79	2,698,296	70,461
Tech.	250,000	22.06	5,514,523	53,717	24.41	6,102,664	64,262
General	250,000	21.54	5,385,459	87,707	26.37	6,592,183	121,074
Total	1,069,515	19.28	20,618,991	200,163	22.75	24,333,201	224,481

Table 2: Statistics of the released 1M UM-PCorpus

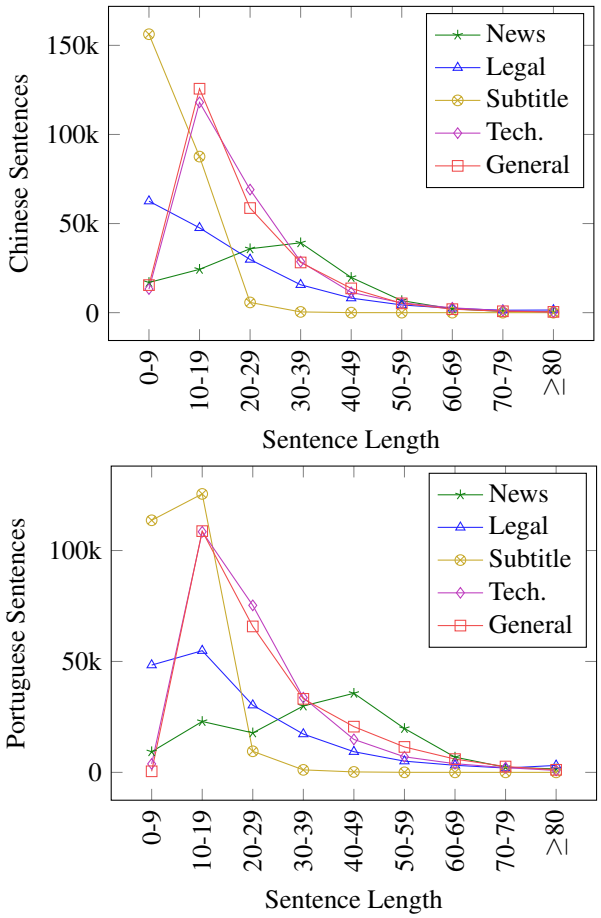

Figure 4: Distribution of sentence lengths in different domains.

The constructed UM-PCorpus contains more than 6.1 million parallel sentences. Table 1 reports the statistics of the corpus in terms of average sentence length, number of words and vocabulary size. The distribution of data among different domains is illustrated in Figure 3. The data across different domains is imbalanced. However, in the released corpus of 1 million sentences, we carefully adjust the content to embrace sentences from different domains in similar proportion. The statistics of the released UM-PCorpus are reported in Table 2, and the distribution of sentence lengths in different domains for Portuguese and Chinese data is presented in Figure 4. For machine translation evaluation purpose, we additionally prepare five test sets and each of which consists of 1000 parallel sentences extracted from each domain. The final test set contains 5000 parallel sentences in total.

4. Machine Translation Experiments

In this section, we present the translation results on the test sets using the NMT systems that trained on the whole UM-PCorpus for Portuguese↔Chinese translation. The test set sentences are excluded from the training data. The Chinese and Portuguese texts are respectively tokenized using the Chinese word segmentation toolkit of *NiuTrans* (Xiao et al., 2012) and the *tokenize.perl* script of Moses.[12] The case-insensitive BLEU is used as the evaluation metric (Papineni et al., 2002). All the models are trained with the following settings. We use our in-house encoder-decoder NMT model which has a deep LSTM network with 2 encoder and 2 decoder layers equipped with a local attention and feed-input model (Luong et al., 2015). The encoder-decoder with LSTM units (Hochreiter and Schmidhuber, 1997) is trained via the back-propagation through time algorithm (BPTT) (Werbos, 1990). All the models use 1024 LSTM nodes per encoder and decoder layers. The size of the source and target word embeddings is 1024. We use the vocabulary size of 60K for both source and target languages. Words are segmented into sub-word units using the byte-pair-encoding algorithm (Sennrich et al., 2016). The size of the mini-batches is set to 80 and the maximum sentence length is 50. We clip the gradient norm to 5.0. The parameters are uniformly initialized in $[-0.08, 0.08]$. The models are trained for 15 epochs using stochastic gradient descent (SGD). The training starts with a learning rate of 0.70 and begins to halve the learning rate every epoch after 7 epochs. We use the dropout rate of 0.2 for our LSTMs (Zaremba et al., 2015).

Test set	Sent.	Avg. Length		BLEU	
		zh	pt	zh→ pt	pt→ zh
News	1,000	27.63	34.09	**22.83**	20.60
Legal	1,000	28.56	31.78	**38.39**	33.40
Subtitle	1,000	8.71	9.92	**22.69**	19.27
Tech.	1,000	22.47	24.86	50.29	**53.75**
General	1,000	22.13	26.02	**38.03**	32.04
Average		21.90	25.33	**35.69**	33.42

Table 3: Translation results on the five test sets

[12] http://www.statmt.org/moses/

Table 3 presents the translation results of the test sets. It is observed that in general the zh→pt translation gives a better BLEU score than that of the pt→zh translation. This is quite different from the conclusions drawn by Belinkov et al. (2017). One possible explanation is that Portuguese is morphologically richer than Chinese. In the training data, the Portuguese exhibits a larger vocabulary size than that of the Chinese one. This can be observed from the statistics reported in Table 1. That results in introducing a large number of Out-of-Vocabularies (OOVs) when we use a vocabulary size of 60K, and consequently the model does not work well for parameter learning when we have insufficient vocabularies. However, we believe this phenomena is worth to explore further.

5. Conclusions

In this paper, we present the construction of a large and high quality Portuguese-Chinese parallel corpus, UM-PCorpus, for machine translation research. The corpus is comprised of texts of news stories, legal articles, video and movie subtitles, technical documents and the general on-line publications. Among the 6 million parallel sentences, about 1 million of which are compiled and made available to the community for research purposes. The released corpus is licensed under the Creative Commons BY-NC-ND 4.0.[13]

6. Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grant No. 61672555), the Joint Project of Macao Science and Technology Development Fund and National Natural Science Foundation of China (Grant No. 045/2017/AFJ) and the Multiyear Research Grants from the University of Macau (Grant Nos. MYRG2017-00087-FST, MYRG2015-00175-FST, MYRG2015-00188-FST and MYRG2016-00109-FAH).

7. Bibliographical References

Belinkov, Y., Durrani, N., Dalvi, F., Sajjad, H., and Glass, J. R. (2017). What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 861–872.

Berger, A. L., Pietra, S. D., and Pietra, V. J. D. (1996). A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., Soricut, R., Specia, L., and Tamchyna, A. (2014). Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2012). Findings of the 2012 workshop on statistical machine translation. In *the Seventh Workshop on Statistical Machine Translation*, pages 10–51.

Chu, C., Nakazawa, T., and Kurohashi, S. (2014). Constructing a chinese-japanese parallel corpus from wikipedia. In *LREC*, pages 642–647.

Gale, W. A. and Church, K. W. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Kolachina, P., Cancedda, N., Dymetman, M., and Venkatapathy, S. (2012). Prediction of learning curves in machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 22–30, Stroudsburg, PA, USA. Association for Computational Linguistics.

Leong, C., Wong, D. F., and Chao, L. S. (2018). Um-paligner : Neural network based parallel sentence identification model. In *Proceedings of the 11th Workshop on Building and Using Comparable Corpora*. European Language Resources Association (ELRA).

Lison, P. and Tiedemann, J. (2016). Opensubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016.*

Lobachev, S. (2008). Top languages in global information production. *Partnership: The Canadian Journal of Library and Information Practice and Research*, 3(2):1.

Luong, T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1412–1421.

McEnery, A. M. and Xiao, R. Z. (2007). Parallel and comparable corpora: What are they up to? *Translating Europe. Incorporating Corpora: The Linguist and the Translator*, pages 18–31.

Munteanu, D. S. and Marcu, D. (2005). Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *ACL 2002*.

Patry, A. and Langlais, P. (2011). Identifying parallel documents from a large bilingual collection of texts: Application to parallel article extraction in wikipedia. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web, BUCC@ACL 2011, Portland, OR, USA, June 24, 2011*, pages 87–95.

Post, M., Callison-Burch, C., and Osborne, M. (2012). Constructing parallel corpora for six indian languages via crowdsourcing. In *Proceedings of the Seventh Work-*

[13] https://creativecommons.org/licenses/by-nc-nd/4.0/

shop on Statistical Machine Translation, pages 401–409. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Smith, J. R., Saint-Amand, H., Plamada, M., Koehn, P., Callison-Burch, C., and Lopez, A. (2013). Dirt cheap web-scale parallel text from the common crawl. In *ACL (1)*, pages 1374–1383.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., and Varga, D. (2006). The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006, Genoa, Italy, May 22-28, 2006.*, pages 2142–2147.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112.

Tian, L., Wong, D. F., Chao, L. S., Quaresma, P., Oliveira, F., and Yi, L. (2014). Um-corpus: A large english-chinese parallel corpus for statistical machine translation. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Wang, L., Wong, D. F., Chao, L. S., Lu, Y., and Xing, J. (2014). A systematic comparison of data selection criteria for smt domain adaptation. *The Scientific World Journal*, 2014(Article ID 745485):1–10.

Weber, G. (1999). Top languages: The world's 10 most influential languages. *AATF National Bulletin*, 24(3):22–28.

Werbos, P. J. (1990). Bachpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560.

Wong, F., Chao, S., Hao, C. C., and Leong, K. S. (2009). A maximum entropy (me) based translation model for chinese characters conversion. *Advances in Computational Linguistics, Research in Computer Science*, 41:267–276. 10th Conference on Intelligent Text Processing and Computational Linguistics - CICLing, Mexico City. http://www2.dc.ufscar.br/ helenacaseli/.

Wong, D. F., Chao, L. S., and Zeng, X. (2014). isentenizer-μ: Multilingual sentence boundary detection model. *The Scientific World Journal*, 2014:1–10. http://www.hindawi.com/journals/tswj/2014/196574/.

Wong, D. F., Lu, Y., and Chao, L. S. (2016). Bilingual recursive neural network based data selection for statistical machine translation. *Knowledge-Based Systems*, 108:15 – 24. New Avenues in Knowledge Bases for Natural Language Processing.

Xiao, T., Zhu, J., Zhang, H., and Li, Q. (2012). Niutrans: an open source toolkit for phrase-based and syntax-based machine translation. In *ACL 2012*.

Xing, J., Wong, D. F., Chao, L. S., Leal, A. L. V., Schmaltz, M., and Lu, C. (2016). Syntaxtree aligner: A web-based parallel tree alignment toolkit. In *Proceedings of the International Conference on Wavelet Analysis and Pattern Recognition*, pages 37–42, Jeju, South Korea. IEEE.

Xu, M., Li, Q., Ao, C. H., Li, Y., Chao, L. S., and Wong, D. F. (2017). The um-nlp2ct neural machine translation system for cwmt2017 translation task. In *Proceedings of the 13th China Workshop on Machine Translation (CWMT 2017)*, Dalian, China, Sept. 27-29, 2017. CIPS.

Yang, B., Wong, D. F., Xiao, T., Chao, L. S., and Zhu, J. (2017). Towards bidirectional hierarchical representations for attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1443–1452. Association for Computational Linguistics.

Ye, T., Wang, T., McGuinness, K., Guo, Y., and Gurrin, C. (2016). Learning multiple views with orthogonal denoising autoencoders. In *MultiMedia Modeling - 22nd International Conference, MMM 2016, Miami, FL, USA, January 4-6, 2016, Proceedings, Part I*, pages 313–324.

Zamani, H., Faili, H., and Shakery, A. (2016). Sentence alignment using local and global information. *Computer Speech & Language*, 39:88–107.

Zaremba, W., Sutskever, I., and Vinyals, O. (2015). Recurrent neural network regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Zeng, X., Wong, D. F., Chao, L. S., and Trancoso, I. (2013a). Co-regularizing character-based and word-based models for semi-supervised chinese word segmentation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, pages 171–176, Sofia, Bulgaria, August. Association for Computational Linguistics.

Zeng, X., Wong, D. F., Chao, L. S., and Trancoso, I. (2013b). Graph-based semi-supervised model for joint chinese word segmentation and part-of-speech tagging. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pages 770–779.

Ziemski, M., Junczys-Dowmunt, M., and Pouliquen, B. (2016). The united nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016.*

A Study on Machine Translation-oriented Parallel Corpus Construction Techniques for Tibetan, Chinese and English

Cizhen Jiacuo, Sangjie Duanzhu, Zhoumao Xian

Qinghai Normal University

E-mail: 543819011@qq.com, sangjeedondrub@live.com, 513576745@qq.com

Abstract

The scarcity of parallel resources between Tibetan and other languages lays a great difficulty for application of current researches in the fields of neural networks and deep learning. The construction of a large-scale parallel Tibetan corpus for Chinese, English and other languages also serves as a great importance for Tibetan NLP in general. More importantly, machine translation between Tibetan and other languages also poses many challenges compared to some current mature machine translation systems like English and other languages. The availability of large-scale multi-lingual parallel language resources is essential to enable minority language machine translation services to better serve the "Belt and Road"'. In this work, through the research of the chapter-level, paragraph-level, sentence-level and word-level automatic acquisition techniques of Tibetan to other language texts, we proposed methods to acquire the knowledge needed for machine translation from the depth and breadth of knowledge mining. This first task in the work is to research on web-oriented automatic discriminant and extraction algorithms for acquiring the comparable corpus, at the same time, by maximizing local matching, to expand the size of the word alignment, phrase alignment library (block aligned library), in order to enrich the Tibetan related parallel language resources. The second is to study on individual paragraph representations based on the large-scale Chinese, Tibetan and English monolingual corpus. And by comparing the similarity of representations and optimizing the threshold to evaluate bilingual comparability both in horizontal and vertical directions. And third is to study the methods to improve the alignment of language resources using monolingual and trilingual word representations as well as the paragraph representations.

Keywords: Tibetan, Language resources, comparable corpus, alignment

1. Introduction

Availability of large-scale parallel corpus is the most indispensable resource for machine translation system, especially in Neural Machine Translation(NMT)(Bahdanau, Cho, & Bengio, 2014; Sutskever, Vinyals, & Le, 2014) Lacking large-scale language corpus poses a major practical problem for many language pairs (Artetxe, Labaka, Agirre, & Cho, 2017). Tibetan related studies in Natural Language Processing fields bloom in recent years, however, scarcity of parallel language resources is largely restraining the further researches. In this work, we present an automatic approach for constructing Tibetan-Chinese-English trilingual parallel corpus by exploiting sentence similarity which is computed via comparing the continuous representation of sentences(Le & Mikolov, 2014) extracted from comparable corpus we collected. Comparable corpus is a set of monolingual corpora which usually in same or similar topics in different languages. This kind of corpora paralleling on document level can be obtained from the web and other sources due to sentences in it are not necessarily translations of each other language pair, therefore building and using comparable corpora is often a more feasible option in multilingual information processing (Liu & Zhang, 2013).

Many Chinese governmental sites such as http://www.people.com.cn/ have multiple language locale settings including Chinese, Tibetan, English etc., and the content in those sites can serve as a good source to construct comparable corpus. Furthermore, we have large amount in-house parallel documents (shown in Table 1) in Tibetan, Chinese and English languages.

2. Related Works

Building comparable corpus attracted attention in the fields due to its flexibility and cost to construct relatively large parallel corpus which is usually hand-crafted with a huge amount of time and efforts. (Resnik, Philip, Smith, & Noah, 2003) presented an approach to mining parallel using STRAND software on Internet through supervised modeling based on structural features. (Talvensaari, Laurikkala, Juhola, & Keskustalo, 2007) presented a method to create comparable corpus by using relative term frequency (RAFT) value from collections in very different in origin. (Shang et al., 2017) proposed a framework, AutoPhrase, which extract high-quality phrases from the public knowledge base in an effective and automatic manner without the availability of POS tagger and rules designed by human experts. (Hashemi, Shakery, & Faili, 2010) align documents crawled from BBC news in different languages by comparing the similarity of document topics and corresponding publishing dates. In (Yasuda & Sumita, 2008) the authors reported an approach to translate original article and translate it into another language, then calculate the evaluation scores upon the translated and original articles. The evaluation is utilized to predict the similarity between two original Wikipedia articles.

In many recent researches, neural networks models are also proposed on the subject of extraction parallel sentences from comparable corpus, in (Chenhui Chu Raj Dabre, 2016) the author proposed a new method to first train a filter using

a seed parallel corpus and then use this filter to classify parallel sentence candidates.

3. Methods

3.1 Comparable Corpus

A *noisy parallel corpus* contains bilingual sentences that are not perfectly aligned or have poor quality translations. Nevertheless, most of its contents are bilingual translations of a specific document.

A *comparable corpus* is built from non-sentence-aligned and untranslated bilingual documents, but the documents are topic-aligned.

This article mainly uses the Internet to collect corpora, which will eventually be used in the construction of comparable Chinese-English-Tibetan corpora. The collected linguistic data should be of high research value while guaranteeing its stability and extensiveness. The obtained network text corpora are basically from Translation portal for Tibetan, Chinese, and English. In order to ensure the practicality of the trilingual data, we have collected a large amount of news corpora because of the high accuracy of news linguistic materials, clear logical logic, and certain representation. The corpora of this article are partly from Xinhua.net and people.com. Some of the nets come from their own data. The scale of the collected corpus is shown in Table 1.

Language	Xinhua	People.com.cn	In-house	Total
Chinese	3200	2500	5600	11300
English	2900	2450	3600	8950
Tibetan	2000	2400	5600	10000

Table 1: Comparable Corpus size in document counts.

In this paper, we use the similarity descending order to set the threshold value for selecting the comparable expectation. The experimental steps for putting the similarity greater than the threshold value into the comparable prediction library are as follows:

1. Filter the sample data
2. Segment the sample corpora, remove the stop words, and add the word bag model
3. Calculate the feature vectors of each word in the word bag and use these feature vectors to represent the document vector
4. Train the model and calculate the index with the calculated feature vector values
5. Mapping the trained sample corpora and candidate corpora to a two-dimensional space to calculate the Euclidean distance
6. Tag the text part-of-speech and filter the text not within the score range
7. Threshold Similarity Filtering, Compliant Documents Added in Comparable Languages, Deletes Not Threshold

Delete the corpus with accuracy less than 80%, and add the accuracy higher than 80% to the corpus3.2 3.2 Training Word Vector Words vector training for Tibetan, Chinese, and English trilingual words is performed using word2vec. Training samples are collected news corpus and own corpus.

3.2 Training word embedding

Word embedding is the collective name for a set of language modeling and feature learning techniques in natural language processing (NLP) where words or phrases from the vocabulary are mapped to vectors of real numbers. Conceptually it involves a mathematical embedding from a space with one dimension per word to a continuous vector space with much lower dimension. Methods to generate this mapping include neural networks (Bengio, Ducharme, Vincent, & Janvin, 2003),dimensionality reduction on the word co-occurrence matrix, and explicit representation in terms of the context in which words appear. Word and phrase embeddings, when used as the underlying input representation, have been shown to boost the performance in NLP tasks such as syntactic parsing and sentiment analysis.

Words vector training for Tibetan, Chinese, and English trilingual words is performed using word2vec. Training samples are collected news corpus and own corpus.

In this work, each word is export to a vector(Le & Mikolov, 2014) . The concatenation or sum of the vectors is then used as features for prediction of the next word in a sentence. More formally, given a sequence of training words $\{w_1, w_2, \dots, w_T\}$, the objective of the word vector model is to maximize the average log probability:

$$\frac{1}{T}\sum_{n=1}^{T} \log \mathrm{p}(w_t | w_{t-k}, \dots, w_{t+k})$$

The prediction task is typically done via a multiclass classifier, such as softmax. There, we have:

$$\mathrm{p}(w_t | w_{t-k}, \dots, w_{t+k}) = \frac{e_{y_{w_t}}}{\sum_i e^{y_i}}$$

3.3 Sentence Alignment Technique

Large corpora used as training sets for machine translation algorithms are usually extracted from large bodies of similar sources, such as databases of news articles written in the first and second languages describing similar events.

However, extracted fragments may be noisy, with extra elements inserted in each corpus. Extraction techniques can differentiate between bilingual elements represented in both corpora and monolingual elements represented in only one corpus in order to extract cleaner parallel fragments of bilingual elements. Comparable corpora are used to directly obtain knowledge for translation purposes. High-quality parallel data is difficult to obtain, however, especially for under-resourced languages.

Comparability is an important indicator of internal evaluation of comparable corpus. The current academic community does not clearly define its concept, but it is usually closely related to the concept of similarity. In most cases, the comparable degree of corpus can be considered as its

similar degree. We believe that comparability can be understood as the degree of similarity of the comparable corpus in terms of authorship, time, space, genre, source, domain, etc., or body and body information (grammar morphology, semantic content, pragmatic characteristics, etc.). This article mainly investigates the comparability of the comparable content of the Chinese-Tibetan-English news corpus, i.e. the similarity. Sentence similarity is calculated as follows: Suppose the sentence T_1 consists of n words, each word's word frequency weight is set to q_n, then there are:$T_1 = \{q_1, q_2, \dots, q_n\}$, T_1 is a multidimensional vector; Let sentence T_2 be composed of n words. The word frequency weight of each word is set to w_n. Then there are: $T_2 = \{w_1, w_2, \dots, w_n\}$, , T_2 is a multidimensional vector; then T_1 and T_2 similarity $\mathrm{Sim}(T_1, T_2)$ is calculated as:

$$\mathrm{Sim}(T_1, T_2) = \cos\alpha = \frac{\sum_{i=1}^{n}(q_i \times w_i)}{\sqrt{\sum_{i=1}^{n} {q_i}^2} \times \sqrt{\sum_{i=1}^{n} {w_i}^2}}$$

The corpus constructed can improve the use of corpus and its application scope by improving the alignment methods, corpus segmentation processing, similarity calculation, and the establishment of comparable relationships.

4. Experiments

In our experiments, we firstly train word embedding for Chinese, Tibetan and English with collected comparable corpus using fasttext[1]. And then two separate machine translation (MT) models were trained bi-directionally for two language pairs, namely Chinese↔English and Chinese↔Tibetan. For the former language pairs we used Google's NMT (Neural Machine Translation) system[2] due to WMT officially provides large amount training datasets. For the latter ones, given we have no access to enough Chinese-Tibetan parallel corpus, and as reported in (Chen, Liu, Cheng, & Li, 2017), training NMT models on low-resource language pairs usually perform poorly than SMT (Statistical Machine Translation) systems, we turn to Moses SMT toolkit [3]to train the model.

After training the MT models, the whole trilingual comparable corpus is processed with language dependent sentence segmenter. And then translate these sentences in to its corresponding languages.

We evaluate sentence similarity in two metrics: BLEU (Papineni, Roukos, Ward, & Zhu, 2002) score of original sentence and translated sentence (Yasuda & Sumita, 2008) and Cosine similarity (Luo, Zhan, Wang, & Yang, 2017) between the embeddings of the two sentences aforementioned.

The aligned sentences is pipelined into retraining process of both NMT and STM systems to gradually improve system's performance in a repeated iterative manner.

5. Conclusion

This paragraph collects and collates comparable corpus through the Internet, and gives a detailed account of the construction techniques and scale, uses word2vec to vectorize the collected corpora, and proposes a method for constructing a Chinese-English-Chinese-English comparative corpus. a corpus with a total of one million words, and evaluates the similarity calculation methods of sentences and the comparability of corpus.

6. Acknowledgements

This work is supported by grant (project ID: 2015-SF-520) from Provincial Science and Technology Department, Qinghai, PR China.

7. References

Artetxe, M., Labaka, G., Agirre, E., & Cho, K. (2017). UNSUPERVISED NEURAL MACHINE TRANSLATION. Retrieved from https://arxiv.org/pdf/1710.11041.pdf

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate, 1–15. https://doi.org/10.1146/annurev.neuro.26.041002.131047

Bengio, Y., Ducharme, R., Vincent, P., & Janvin, C. (2003). A Neural Probabilistic Language Model. The Journal of Machine Learning Research, 3, 1137–1155. https://doi.org/10.1162/153244303322533223

Chen, Y., Liu, Y., Cheng, Y., & Li, V. O. K. (2017). A Teacher-Student Framework for Zero-Resource Neural Machine Translation. Retrieved from http://arxiv.org/abs/1705.00753

Chenhui Chu Raj Dabre, S. K. (2016). Parallel Sentence Extraction from Comparable Corpora with Neural Network Features.

Hashemi, H. B., Shakery, A., & Faili, H. (2010). Creating a Persian-English Comparable Corpus. dx.doi.org.

Le, Q. V, & Mikolov, T. (2014). Distributed Representations of Sentences and Documents, 4, II-1188.

Liu, S., & Zhang, C. (2013). Termhood-based Comparability Metrics of Comparable Corpus in Special Domain. ArXiv E-Prints.

Luo, C., Zhan, J., Wang, L., & Yang, Q. (2017). Cosine Normalization: Using Cosine Similarity Instead of Dot Product in Neural Networks. Retrieved from http://arxiv.org/abs/1702.05870

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. Retrieved from http://www.aclweb.org/anthology/P02-1040.pdf

Resnik, Philip, Smith, & Noah, A. (2003). The Web as a parallel corpus. Computational Linguistics, 29(3),

[1] https://github.com/facebookresearch/fastText

[2] https://github.com/tensorflow/nmt

[3] http://www.statmt.org/moses/

349–380.

Shang, J., Liu, J., Jiang, M., Ren, X., Voss, C. R., & Han, J. (2017). Automated Phrase Mining from Massive Text Corpora.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. Retrieved from http://arxiv.org/abs/1409.3215

Talvensaari, T., Laurikkala, J., Juhola, M., & Keskustalo, H. (2007). Creating and exploiting a comparable corpus in cross-language information retrieval. Acm Transactions on Information Systems, 25(1), 4.

Yasuda, K., & Sumita, E. (2008). Method for Building Sentence-Aligned Corpus from Wikipedia.

NLP for Chinese L2 Writing: Evaluation of Chinese Grammatical Error Diagnosis

Gaoqi Rao[1], Lung-hao Lee[2]
1.Beijing Language and Culture University, 2.National Taiwan Normal University
15. Xueyuan Rd., Beijing, China; 162, Section 1, Heping E. Rd., Taipei City 106, Taiwan
E-mail: raogaoqi@blcu.edu.cn, lhlee@ntnu.edu.tw

Abstract

This paper presents the shared task of Chinese grammatical error diagnosis (CGED) which seeks to identify grammatical error types and their range of occurrence within sentences written by L2 learners of Chinese. We describe the task definition of CGED, and overview the past 4 CGED shared tasks, especially CGED2016 and CGED2017 containing simplified character track of HSK, in data preparation, performance metrics, and evaluation results. Until now, none of the participants has developed an over performed system, showing potential of solving the task, although approaches were significant since the first CGED in 2014. We expected this evaluation campaign could lead to the development of more advanced NLP techniques for educational applications, especially for Chinese error detection and automatic correction. All data sets with gold standards and scoring scripts are made publicly available to researchers.

Keywords: CGED, error detection, L2 Chinese learning

1. Introduction

In recent years, automated grammar checking for learners of English as a foreign language has attracted more attention. For example, Helping Our Own (HOO) is a series of shared tasks in correcting textual errors (Dale and Kilgarriff, 2011; Dale et al., 2012). The shared tasks at CoNLL 2013 and CoNLL 2014 focused on grammatical error correction, increasing the visibility of educational application research in the NLP community (Ng et al., 2013; 2014).

Many of these learning technologies focus on learners of English as a Foreign Language (EFL), while relatively few grammar checking applications have been developed to support Chinese as a Foreign Language(CFL) learners. Those applications which do exist rely on a range of techniques, such as statistical learning (Chang et al, 2012; Wu et al, 2010; Yu and Chen, 2012), rule-based analysis (Lee et al., 2013) and hybrid methods (Lee et al., 2014). In response to the limited availability of CFL learner data for machine learning and linguistic analysis, the ICCE-2014 workshop on Natural Language Processing Techniques for Educational Applications (NLP-TEA) organized a shared task on diagnosing grammatical errors for CFL (Yu et al., 2014). A second version of this shared task in NLP-TEA was collocated with the ACL-IJCNLP-2015 (Lee et al., 2015), COLING-2016 (Lee et al., 2016) and IJCNLP 2017 (Rao et al., 2017). In 2018, the shared task for Chinese grammatical error diagnosis is organized again at NLP-TEA workshop in conjunction with ACL2018.

The main purpose of these shared tasks is to provide a common setting so that researchers who approach the tasks using different linguistic factors and computational techniques can compare their results. Such technical evaluations allow researchers to exchange their experiences to advance the field and eventually develop optimal solutions to this shared task.

2. Task Description

The goal of this shared task is to develop NLP techniques to automatically diagnose grammatical errors in Chinese sentences written by L2 learners. Such errors are defined as redundant words (denoted as a capital "R"), missing words ("M"), word selection errors ("S"), and word ordering errors ("W"). The input sentence may contain one or more such errors. The developed system should indicate which error types are embedded in the given unit (containing 1 to 5 sentences) and the position at which they occur. Each input unit is given a unique number "sid". If the inputs contain no grammatical errors, the system should return: "sid, correct". If an input unit contains the grammatical errors, the output format should include four items "sid, start_off, end_off, error_type", where start_off and end_off respectively denote the positions of starting and ending character at which the grammatical error occurs, and error_type should be one of the defined errors: "R", "M", "S", and "W". Each character or punctuation mark occupies 1 space for counting positions. Example sentences, corresponding notes and data in SGML format are shown as Table 1 and Figure 1 show. In 2014 and 2015, we organized one track of TOCFL (Test Of Chinese as a Foreign Language) (Lee et al., 2016). In 2016, two tracks of TOCFL and HSK (Hanyu Shuiping Kaoshi)(Cui et al, 2011; Zhang et al, 2013) were organized, while in 2017 and 2018, only HSK track was and will be organized. We welcome the affiliations constructing data set of traditional characters to join the shared task in organization.

3. Datasets

Native Chinese speakers were trained to manually annotate grammatical errors and provide corrections corresponding to each error. The data were then split into Training Set and Test Set. Each unit (contain at least 1 sentence) with annotated grammatical errors and their corresponding corrections is represented in SGML format. The scale and error type distribution of the Training Set

in CGED2016 and CGED2017 are reported in Table2. In test set, correct sentences are contained, in order to test the false positive rate of the systems. The distributions of error types (shown in Table 3) are similar with that of the training set.

4. Performance Metrics

Table 4 shows the confusion matrix used for evaluating system performance. In this matrix, TP (True Positive) is the number of sentences with grammatical errors are correctly identified by the developed system; FP (False Positive) is the number of sentences in which non-existent grammatical errors are identified as errors; TN (True Negative) is the number of sentences without grammatical errors that are correctly identified as such; FN (False Negative) is the number of sentences with grammatical errors which the system incorrectly identifies as being correct.

The criteria for judging correctness are determined at three levels as follows.

(1) Detection-level: Binary classification of a given sentence, that is, correct or incorrect, should be completely identical with the gold standard.

(2) Identification-level: This level could be considered as a multi-class categorization problem. All error types should be clearly identified. A correct case should be completely identical with the gold standard of the given error type.

(3) Position-level: In addition to identifying the error types, this level also judges the occurrence range of the grammatical error. That is to say, the system results should be perfectly identical with the quadruples of the gold standard.

(4) Correction-level: In the coming CGED2018 in conjunction with ACL2018 in July 2018, the participant systems are required to offer 0 to 3 recommended corrections to error types of missing and selection. The amount of the correction to recommend depends on the trust computation at each error. More recommendation would increase the recall, but somehow reduce precision, since the gold standard only offers one correction to each error.

The following metrics are measured at all levels with the help of the confusion matrix.

- False Positive Rate = FP / (FP+TN)
- Accuracy = (TP+TN) / (TP+FP+TN+FN)
- Precision = TP / (TP+FP)
- Recall = TP / (TP+FN)
- F1 = 2*Precision*Recall / (Precision + Recall)

5. Evaluation Results and Analysis

Table 5 and Table 6 summarize the submission statistics and best F1 of position-level for the participants in CGED2016 and CGED2017. In summary, none of the submitted systems provided superior performance using different metrics, indicating the difficulty of developing systems for effective grammatical error diagnosis, especially in L2 contexts, although approaches were significant since the first CGED in 2014.

From the proceedings of the 2 shared tasks, we observed the transformation in methods: from traditional statistical modeling to deep neuro networks. About one third of the participants in CGED2016 conduct the system based on Ngram or fined turned CRF, while none of the teams continued to carry out the experiments in these ways. LSTM+CRF has been nearly standard solution to task by each team, similar to other NLP tasks.

Also like what happened in other NLP tasks, deep learning modeling as resource intensive required methods, approached better performance easier in big dataset with high quality. Unfortunately, writing data of L2 Chinese learner are quite limited in both size and quality. Track of HSK as an example, organizers from BLCU digitalized the scored writing section from the exam. Teachers in exam scoring were not required the high consistency, like other annotation task like word segmentation or sentiment analysis. On the other hand, the NLP for Chinese as L2 learning does not have a long history and impact among academia, leading to the relative low resource construction, comparing with other newly appeared task like SQuAD.

These problems in resource aspect partially lead to the limited performance of deep learning modeling. However, this task can be viewed as a low resource NLP task to challenge.

6. Conclusions

This study describes the shared task for Chinese grammatical error diagnosis, including task design, data preparation, performance metrics, and evaluation results. Regardless of actual performance, all submissions contribute to the common effort to develop Chinese grammatical error diagnosis system, and the individual reports in the proceedings provide useful insights into computer-assisted language learning for CFL learners.

We hope the data sets collected and annotated for this shared task can facilitate and expedite future development in this research area. Therefore, all data sets with gold standards and scoring scripts are publicly available online at www.cged.science.

7. Acknowledgments

We thank all the participants for taking part in our shared task. We would like to thank Kuei-Ching Lee for implementing the evaluation program and the usage feedbacks from Bo Zheng (in CGED2016). Gong Qi, Tang Peilan, Luo Ping and Chang Jie contributed in the proofreading of the data in CGED2017/2018.

This study was supported by the projects from P.R.C: High-Tech Center of Language Resource(KYD17004), BLCU Innovation Platform(17PT05), Institute Project of BLCU(16YBB16) Social Science Funding China (11BYY054, 12&ZD173, 16AYY007), Social Science Funding Beijing (15WYA017), National Language

Committee Project (YB125-42, ZDI135-3), MOE Project of Key Research Institutes in Univ(16JJD740004).

TOCFL (Traditional Chinese)	**HSK (Standard Chinese)**
• Example 1 Input: (sid=A2-0007-2)　聽說妳打算開一個慶祝會。可惜我不能參加。因為那個時候我有別的事。當然我也要參加給你慶祝慶祝。 Output: A2-0007-2, 38, 39, R (Notes: “參加”is a redundant word) • Example 2 Input: (sid=A2-0011-1)　我聽到你找到工作。恭喜恭喜！ Output: A2-0011-1, 2, 3, S A2-0011-1, 9, 9, M (Notes: “聽到”should be “聽說”. Besides, a word “了”is missing. The correct sentence should be “我聽說你找到工作了”. • Example 3 Input: (sid=A2-0011-3)　我覺得對你很抱歉。我也很想去，可是沒有辦法。 Output: A2-0011-3, correct	• Example 1 Input: (sid=00038800481)　我根本不能了解这妇女辞职回家的现象。在这个时代，为什么放弃自己的工作，就回家当家庭主妇? Output: 00038800481, 6, 7, S 00038800481, 8, 8, R (Notes: “了解”should be “理解”. In addition, “这” is a redundant word.) • Example 2 Input: (sid=00038800464)我真不明白。她们可能是追求一些前代的浪漫。 Output: 00038800464, correct • Example 3 Input: (sid=00038801261)人战胜了饥饿，才努力为了下一代作更好的、更健康的东西。 Output: 00038801261, 9, 9, M 00038801261, 16, 16, S (Notes: “能” is missing. The word “作”should be “做”. The correct sentence is “才能努力为了下一代做更好的”)

Table 1: Example sentences and corresponding notes.

```
<DOC>
<TEXT id="A2-0005-1">
我聽說你打算開一個慶祝會。對不起，我要參加，可是沒有空。你開一個慶祝會的時候我不能會參加，是因為我在外國做工作。
</TEXT>
<CORRECTION>
我聽說你打算開一個慶祝會。對不起，我要參加，可是沒有空。你開慶祝會的時候我不能參加，是因為我在外國工作。
</CORRECTION>
<ERROR start_off=" 31" end_off=" 32"  type="R"></ERROR>
<ERROR start_off=" 42" end_off=" 42"  type="R"></ERROR>
<ERROR start_off=" 53" end_off=" 53"  type="R"></ERROR>
</DOC>
<DOC>
<TEXT id="200210543634250003_2_1x3">
对于“安乐死”的看法，向来都是一个极具争议性的题目，因为毕竟每个人对于死亡的观念都不一样，怎样的情况下去判断，也自然产生出很多主观和客观的理论。每个人都有着生存的权利，也代表着每个人都能去决定如何结束自己的生命的权利。在我的个人观点中，如果一个长期受着病魔折磨的人，会是十分痛苦的事，不仅是病人本身，以致病者的家人和朋友，都是一件难受的事。
</TEXT>
<CORRECTION>
对于“安乐死”的看法，向来都是一个极具争议性的题目，因为毕竟每个人对于死亡的观念都不一样，无论在怎样的情况下去判断，都自然产生出很多主观和客观的理论。每个人都有着生存的权利，也代表着每个人都能去决定如何结束自己的生命。在我的个人观点中，如果一个长期受着病魔折磨的人活着，会是十分痛苦的事，不仅是病人本身，对于病者的家人和朋友，都是一件难受的事。
</CORRECTION>
```

```
<ERROR start_off="46" end_off="46" type="M"></ERROR>
<ERROR start_off="56" end_off="56" type="S"></ERROR>
<ERROR start_off="106" end_off="108" type="R"></ERROR>
<ERROR start_off="133" end_off="133" type="M"></ERROR>
<ERROR start_off="151" end_off="152" type="S"></ERROR>
</DOC>
```

Figure 1: Example units in SGML format (in traditional and standard character).

Evaluation	Track	#Units	#Error	#R	#M	#S	#W
CGED2016	TOCFL	10,693	24,492 (100%)	4,472 (18.3%)	8,739 (35.7%)	9,897 (40.4%)	1,384 (5.7%)
	HSK	10,071	24,797 (100%)	5,538 (22.3%)	6,623 (26.7%)	10949 (44.2%)	1,687 (6.8%)
CGED2017	HSK	10,449	26,448 (100%)	5,852 (22.1%)	7,010 (26.5%)	11,591 (43.8%)	1,995 (7.5%)

Table 2: The statistics of training set.

Evaluation	Track	#Units	#Correct	#Erroneous	#Error	#R	#M	#S	#W
CGED2016	TOCFL	3,528	1,703 (48.3%)	1,825 (51.7%)	4,103 (100%)	782 (19.06%)	1,482 (36.12%)	1,613 (39.31%)	226 (5.51%)
	HSK	3,011	1,539 (51.1%)	1,472 (48.9%)	3,695 (100%)	802 (21.71%)	991 (26.82%)	1620 (43.84%)	282 (7.63%)
CGED2017	HSK	3,154	1,173 (48.4%)	1,628 (51.6%)	4,876 (100%)	1,062 (21.78%)	1,274 (26.13%)	2,155 (44.20%)	385 (7.90%)

Table 3: The statistics of testing set.

Confusion Matrix		System Results	
		Positive (Erroneous)	Negative(Correct)
Gold Standard	Positive	TP (True Positive)	FN (False Negative)
	Negative	FP (False Positive)	TN (True Negative)

Table 4: Confusion matrix for evaluation.

Participant (Ordered by abbreviations of names)	#TRuns	F1	#HRuns	F1
NLP Lab, Zhengzhou University (**ANO**)	0	-	2	0.2666
Central China Normal University (**CCNU**)	0	-	1	0.0121
Chaoyang University of Technology (**CYUT**)	3	**0.1248**	3	0.2125
Harbin Institute of Technology (**HIT**)	0	-	3	**0.3855**
Institute of Computational Linguistics, Peking University (**PKU**)	3		3	0.0724
National Chiao Tung University & National Taipei University of Technology (**NCTU+NTUT**)	3	0.0745	0	-
National Chiayi University (**NCYU**)	3	0.0155	3	0.0183
NLP Lab, Zhengzhou University (**SKY**)	0	-	3	0.3627
School of Information Science and Engineering, Yunnan University (**YUN-HPCC**)	3	0.0007	3	0.0035

Table 5: Submission statistics for all participants in CGED2016.

Participant (Ordered by abbreviations of names)	#Runs	F1
ALI_NLP	3	**0.2693**
BNU_ICIP	3	0.1152
CVTER	2	0.0653
NTOUA	2	0.0348
YNU-HPCC	3	0.1255

Table 6: Submission statistics for all participants in CGED2017.

8. References

Ru-Yng Chang, Chung-Hsien Wu, and Philips Kokoh Prasetyo. 2012. Error diagnosis of Chinese sentences usign inductive learning algorithm and decomposition-based testing mechanism. ACM Transactions on Asian Language Information Processing, 11(1), article 3.

Xiliang Cui, Bao-lin Zhang. 2011. The Principles for Building the "International Corpus of Learner Chinese". Applied Linguistics, 2011(2), pages 100-108.

Robert Dale and Adam Kilgarriff. 2011. Helping our own: The HOO 2011 pilot shared task. In Proceedings of the 13th European Workshop on Natural Language Generation(ENLG'11), pages 1-8, Nancy, France.

Reobert Dale, Ilya Anisimoff, and George Narroway. 2012. HOO 2012: A report on the preposiiton and determiner error correction shared task. In Proceedings of the 7th Workshop on the Innovative Use of NLP for Building Educational Applications(BEA'12), pages 54-62, Montreal, Canada.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 shared task on grammatical error correction. In Proceedings of the 18th Conference on Computational Natural Language Learning (CoNLL'14): Shared Task, pages 1-12, Baltimore, Maryland, USA.

Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. 2013. The CoNLL-2013 shared task on grammatical error correction. In Proceedings of the 17th Conference on Computational Natural Language Learning(CoNLL'13): Shared Task, pages 1-14, Sofia, Bulgaria.

Lung-Hao Lee, Li-Ping Chang, and Yuen-Hsien Tseng. 2016. Developing learner corpus annotation for Chinese grammatical errors. In Proceedings of the 20th International Conference on Asian Language Processing (IALP'16), Tainan, Taiwan.

Lung-Hao Lee, Li-Ping Chang, Kuei-Ching Lee, Yuen-Hsien Tseng, and Hsin-Hsi Chen. 2013. Linguistic rules based Chinese error detection for second language learning. In Proceedings of the 21st International Conference on Computers in Education(ICCE'13), pages 27-29, Denpasar Bali, Indonesia.

Lung-Hao Lee, Liang-Chih Yu, and Li-Ping Chang. 2015. Overview of the NLP-TEA 2015 shared task for Chinese grammatical error diagnosis. In Proceedings of the 2nd Workshop on Natural Language Processing Techniques for Educational Applications (NLP-TEA'15), pages 1-6, Beijing, China.

Lung-Hao Lee, Liang-Chih Yu, Kuei-Ching Lee, Yuen-Hsien Tseng, Li-Ping Chang, and Hsin-Hsi Chen. 2014. A sentence judgment system for grammatical error detection. In Proceedings of the 25th International Conference on Computational Linguistics (COLING'14): Demos, pages 67-70, Dublin, Ireland.

Lung-Hao Lee, Rao Gaoqi, Liang-Chih Yu, Xun, Eendong, Zhang Baolin, and Chang Li-Ping. 2016. Overview of the NLP-TEA 2016 Shared Task for Chinese Grammatical Error Diagnosis. The Workshop on Natural Language Processing Techniques for Educational Applications (NLP-TEA' 16), pages 1-6, Osaka, Japan.

Chung-Hsien Wu, Chao-Hong Liu, Matthew Harris, and Liang-Chih Yu. 2010. Sentence correction incorporating relative position and parse template language models. IEEE Transactions on Audio, Speech, and Language Processing, 18(6), pages 1170-1181.

Chi-Hsin Yu and Hsin-Hsi Chen. 2012. Detecting word ordering errors in Chinese sentences for learning Chinese as a foreign language. In Proceedings of the 24th International Conference on Computational Linguistics (COLING'12), pages 3003-3017, Bombay, India.

Liang-Chih Yu, Lung-Hao Lee, and Li-Ping Chang. 2014. Overview of grammatical error diagnosis for learning Chinese as foreign language. In Proceedings of the 1stWorkshop on Natural Language Processing Techniques for Educational Applications (NLP-TEA'14), pages 42-47, Nara, Japan.

Bao-lin Zhang, Xiliang Cui. 2013. Design Concepts of "the Construction and Research of the Inter-language Corpus of Chinese from Global Learners". Language Teaching and Linguistic Study, 2013(5), pages 27-34.

The Characteristics of Southeast Asian Languages and Their Influence on Translation

Lianfang LIU[1], Zixian DENG[1], Jiakai WEN[2], Liangchun LU[2], Yuanyuan PAN[3], Lixiang ZHAO[3]

1. Guangxi Computing Center, Nanning, Guangxi 530022, P. R. China;

2.Guangxi Daring E-Commerce Services Co., Ltd., Nanning, Guangxi 530007, P. R. China

Abstract: The exchange and cooperation between China and Southeast Asian countries serves as an important part of the Maritime Silk Road under the "Belt and Road" Initiative. Linguistic communication is a prerequisite for the exchange and cooperation. There are 9 major official languages, including English, Vietnamese, Thai, Malay, Indonesian, Lao, Burmese, Cambodian and Filipino in 10 countries of Southeast Asian. The latter 8 Southeast Asian languages are characterized by different grammars and are grammatically subsidiary to the Austroasiatic Language Family, Austronesian Language Family and Sino-Tibetan Language Family. The premise for works on the construction of Southeast Asian language resources and machine translation is to learn about these languages. As part of learning the basic knowledge of Southeast Asian languages, this paper introduces the characteristics of Southeast Asian languages and their influence on translation.

Keywords**:** Southeast Asian languages; Austroasiatic language family; Austronesian language family; Sino-Tibetan language family; translation

I. Terms and definitions

Analytic language：also known as “isolating language” or “radical language”, refers to a linguistic form characterized by: expressing grammatical relations by different word orders and function words when organizing sentences; using many compound words and few derivative words; using nouns without changes on gender, quantity and case.

Agglutinative language：refers to a linguistic form with rich morphological changes and expresses various grammatical relations via changes on forms of the words themselves. It is characterized by the combination of word roots and additional components and the overlap (or partial overlap) of word roots as the main means of word-formation and morpho-formation. The additional components are divided into prepositive, central and postpositive ones.

Loanwords：also known as "foreign words", refers to those words which are loaned from foreign languages phonetically and semantically. Every language has a certain number of loanwords. For example, the Chinese words like “ 葡萄”, “石榴”, “狮子” and “玻璃” were loaned from the Western Region in the Han Dynasty of ancient China; the Chinese Buddhist terms like “佛”,“菩萨”, “罗汉”,“尼”, “和尚” were borrowed from the ancient India after the Han Dynasty; and words like “胡同” and “站” were loaned from the ancient Mongolia during the Chinese Yuan Dynasty.

Tonal language: refers to languages in which meanings of single words are distinguished with tones, and is characterized that different meanings are formed by tones of different lengths and levels without changes in pronunciation.

II. Linguistic Characteristics

The 8 Southeast Asian languages are subsidiary to the Austronesian Language Family, Austroasiatic Language Family and Sino-Tibetan Language Family respectively, present a large number of common characteristics yet have their own features.

2.1 In Morphology

The Austroasiatic Vietnamese and Cambodian and the Sino-Tibetan Thai, Burmese and Lao are analytic languages just like Chinese. In these languages, the grammatical relations are expressed mainly by word orders and function words without changes on gender, quantity and case of nouns.

The Austronesian Indonesian, Malay and Filipino are agglutinative languages with rich morphological variations which can express various grammatical relations. The combination of word roots and additional components (i.e. prefix, infix and suffix) and the overlap (or partial overlap) of word roots is the main means of word-formation and morpho-formation.

2.2 In Phonetics and Tones

The Austroasiatic Vietnamese and the Sino-Tibetan Thai, Burmese and Lao are tonal languages like Chinese. In these languages, different meanings are formed by tones of different lengths and levels without changes in pronunciation. Vietnamese and Lao thereof have the richest tones of up to six.

The Austroasiatic Cambodian and the Austronesian Indonesian, Malay and Filipino belong to non-tonal languages (namely "intonation languages") like English, where the phonetic tones in different lengths represent only tones rather than meanings.

2.3 In Writing System

The Austroasiatic Vietnamese and the Austronesian Indonesian, Malay and Filipino adopt Latin alphabet. The spelling of Vietnamese is even more complicated, including 7 Latin letter variants (letters for Vietnamese only) and 6 tone symbols, where careless misspelling would lead to wrong meanings.

The Austroasiatic Cambodian and the Sino-Tibetan Thai, Burmese and Lao adopt their own unique but similar-sourcing alphabet letters.
These languages are quite different from the pictographic and ideographic Chinese in writing.

2.4 In Syntax

Vietnamese, Cambodian, Indonesian, Malay, Thai and Lao in the three major language families all use the same basic order of "Subject-Predicate-Object" as Chinese, yet what is different from Chinese is that the modifiers are placed after the central words modified.

Moreover, the Austronesian Filipino uses quite unusual word orders like “Object - Predicate - Subject” or “Predicate - Subject - Object”; and the Sino- Tibetan Burmese adopts a word order of“Subject - Predicate - Object” but the modifiers are placed before the central words like Chinese.

2.5 In Vocabulary

All of these languages contain a certain amount of loanwords (i.e. "foreign words"), and have been affected to varying degrees by Sanskrit and Pali since the introduction of Buddhism.

Cambodian, Malay, Thai, Lao and Burmese are most affected by Sanskrit. Most loanwords from Sanskrit and Pali languages are polysyllabic words and still retain the original gender and quantity characteristics. Vietnamese is deeply influenced by Chinese, French and English, specifically, loanwords from Chinese account for about 60% of all the loanwords in Vietnamese and even up to 70-80% in political, economic and legal fields while technical words are mainly loaned from French and English. Indonesian is heavily influenced by the Dutch system and contains a large number of Javanese and Dutch loanwords in its vocabulary. Filipino vocabulary is deeply affected by Spanish.

In addition, Malay and Indonesian differ slightly in the pronunciation and vocabulary of the writing system, and people who use these two languages are basically able to communicate with each other. The Javanese and Dutch loanwords in Indonesian are the main reason for this difference.

III. Difficulties in human translation

It is generally known that the development and maturity of an industry requires the accumulation of time and practice. However, the translation industry for Southeast Asian languages does not have the two conditions in nature comparing with generally used languages like English.

3.1 Transfer between syntactical meanings

This is one of the major difficulties for beginners in translation for Southeast Asian languages. Especially when translating complex long sentences, if not handle in a proper way, they may produce "translationese” (foreign-styled Chinese or Chinese-styled foreign language) that would be neither precise nor fluent and even result in serious mistranslations that are completely contrary to the original meanings.

Taking Chinese to Thai translation for instance, there are often many gorgeous attributes appearing before central words in Chinese sentences which make it difficult for translator to identify these attributes, to decide how to put the translation after central words in Thai without producing literal faults and to find enough gorgeous Thai words.

Taking Indonesian to Chinese translation as another example, the "Yang" structure serves as an important linkage in long sentences of Indonesian. It is often difficult for translators to read and translate such a long sentence which describes many things without using any punctuation and with multi-level clauses. Any mistake would lead to literal faults of the entire sentence or even the paragraph.

3.2 Translation of loanwords

Vocabularies of Cambodian, Thai, Lao and Burmese contain a large number of Sanskrit and Pali loanwords used in religious and aristocratic life. It is very difficult to memorize and use these words. For one thing, Sanskrit is considered as the world's most difficult language to learn for its polysyllable words and changes on gender and quantity. For another thing, due to limitation in cultural and developing level of Southeast Asian countries, English loanwords are commonly used and written in the countries’ own special writing system to express modern vocabulary. Thus it is difficult for translators who are not familiar with English to translate Southeast Asian languages into Chinese quickly.

The Chinese loanwords in Vietnamese can be divided into 3 categories: phonemical & semantical, phonemical only, and Vietnamized. Translators should distinguish them and make good understanding to translate them correctly. Indonesian vocabulary contains a number of Dutch and Javanese loanwords that cannot be found in dictionaries. In addition, English loanwords spelled in Indonesian pronunciation are commonly used probably because they use the same Latin alphabet writing system. For example, some project contracts and documents for bidding and tendering (accounting for about 70% of Indonesian documents needed to be translated in the market) often contain Indonesian- styled English words that cannot be found in dictionaries and are difficult to be translated. Conversely, translators would find it hard to complete their work on translating Chinese engineering and mechanical documents into Indonesian if they are not familiar with English and unable to produce "Indonesian-styled" English loan words.

3.3 Conversion of Punctuations

In countries colonized by French or Dutch in the history like Vietnam and Indonesia, a comma (",") is often used as a decimal point and a full stop (".") is used as a thousand separator. However, the Chinese usage is quite the contrary. For example, "12,345" in Vietnamese should be translated into "12.345" in Chinese and vice versa.

In Thai language, there are no comma, full stop, question mark, exclamation and other punctuations. The text takes the form of continuous writing without any punctuation and space between words, and uses an interval of two letters or a small pause in sentences to indicate the ending of a sentence. In that case, the translators should analyze the context carefully to judge the boundaries of words and sentences and decide how to do the translation. It is easy to make mistakes if translators have poor vocabulary or contextual thinking.

IV. Difficulties in Machine Translation

4. 1 Text Encoding

Comparing with languages based on Latin alphabet writing system such as Indonesian, Malaysian, and Filipino, it is more difficult for other Southeast Asian languages in text recording, recognizing and handling. On one hand,the data entry can only be completed by a person who knows how to type the language of the country with specified fonts, or else the typing results cannot be recognized by the system. However, as for Indonesian, anyone who can type in 26 letters can enter data and check the texts. On the other hand, there are few or even no available tools to support text encoding of these

languages, which requires specific research and development work. Besides, it is easy to generate messy codes during switching and processing of different procedures and tools.

4.2 Syntactical Translation

The major difficulties for Rule-based Machine Translation (RBMT) lie in the differences of syntactic rules between two languages and the analysis of sentence constituents.

Despite the current popular Neural Machine Translation (NMT) may not have to face the difficulties mentioned above, its features were limited greatly by lacking of corpora.

4.3 Segmentation Rules

In Thai, there is no any ending punctuation and clear segmentation rules, which impacts the results of Machine Translation to a certain extent. In Burmese, full stops are written in special Burmese character, which should be considered in Machine Translation.

4.4 Translation of Loanwords

Most loanwords are unlogged words without corresponding meanings in corpora, and they require human translation.

4.5 Conversion of Punctuation

As same as human translation, the usage of decimal points and digital symbols in Vietnamese and Indonesian is quite contrary to that in Chinese. How to judge and convert punctuations correctly and automatically to avoid serious quality defects remains an issue that needs to be solved.

References

WANG Hui. The Language Situation and Language Policy in the “Belt and Road” Countries, Volume 1. SOCIAL SCIENCES ACADEMIC PRESS (CHINA).

LIN Minghua. A Brief Discussion on written Vietnamese [J]. Modern Foreign Languges, 1983(3):55-59.

TAN Zhici. A Primary Analysis on Reasons for which Written Vietnamese is Profoundly Influenced by Chinese Characters [J]. SOUTHEAST ASIAN AND SOUTH ASIAN STUDIES, 1998(2):47-50.

CHEN Hui. The Situation and Developing Trend of Language Department of South-east Asia in China[J]. 2007(3):72-75.

AWord and Its Rules

Inner Mongolia University

Nasun-urt

As"deep learning"and"neural network" has become the mainstream technology of natural language processing today, language resources of many small languages in the world are relatively deficient, and they can not meet the processing pattern based on the "big data".In this case, it is necessary to consider the injection of language knowledge to expect the realization of the understanding and processing of language at different levels through "deep learning".Therefore, it is decisive to pay attention to the grammatical and semantic characteristics of different languages and sum up the rules for the natural language processing.

Mongolian language is a typical agglutinate language, and its word formation and configuration are all realized by attaching various supplementary components to the stem.The grammatical meaning of a Mongolian word can only be expressed with phrases or sentences in most western languages and oriental languages, and this kind of changes of the words in real Mongolian text constitute about 82% of all words. If these changes are not taken into account and each word is only understood by its stem meaning, it would be impossible to correctly handle the entire text. As a special case of the language knowledge description, we have summarized the grammar rules of a Mongolian word  . It is worth mentioning that the rules are only at the lexical change level and its related parts of one semantic item of the word stem , and the other more semantic items and the relevant rules are not exhaustively described.

The Mongolian word "go " and its relevant rules

 (They) have let (all people) go.

---YABV/Ve2+GVL/Fe11+JAGA/Fe5+CIHA/Fi21+JAI/Fs11

---goYABV/Ve2(stem-imperative form-second person)+GVL/Fe11（causative voice）+JAGA/Fe5（multiple voice）+CIHA/Fi21（perfective aspect）+JAI/Fs11（past tense-statement）

 YABV/Ve2

 YABV/Ve2+JAI/Fs11

 YABV/Ve2+GVL/Fe11

 YABV/Ve2+GVL/Fe11+JAI/Fs11

YABV/Ve2+JAG_A/ Fe5

 YABV/Ve2+JAGA/Fe5+JAI/Fs11

 YABV/Ve2+CIH_A/Fi21

 ?YABV/Ve2+CIHA/Fi21+JAI/Fs11

 YABV/Ve2+GVL/Fe11+JAG_A/Fe5

 YABV/Ve2+GVL/Fe11+JAGA/Fe5+JAI/Fs11

 YABV/Ve2+GVL/Fe11+CIH_A/Fi21

 YABV/Ve2+GVL/Fe11+CIHA/Fi21+JAI/Fs11

 YABV/Ve2+JAGA/Fe5+CIH_A/Fi21

 YABV/Ve2+JAGA/Fe5+CIHA/Fi21+JAI/Fs11

 YABV/Ve2+GVL/Fe11+JAGA/Fe5+CIH_A/Fi21

 YABV/Ve2+GVL/Fe11+JAGA/Fe5+CIHA/Fi21+JAI/Fs11

The pronouns in co-occurrence with the verb

BI/Rb11NAM_A/Rb12 BIDE/Rb13

CI/Rb21CIM_A/Rb22TANAR/Rb23

TERE/Rb31TEGUN/Rb32 TEDENER/Rb33

The following is B0 Rule Set (the subject or the agent is the second person singular or plural, or singular plural).

CI/Rb21|| TANAR/Rb23→YABV/Ve2

CI/Rb21 || TANAR/Rb23(NAM_A/Rb12 || BIDE/Rb13 || TEGUN/Rb32 || TEDENER/Rb33) →YABV/Ve2+GVL/Fe11

TANAR/Rb23→YABV/Ve2+JAG_A/Fi5

CI/Rb21 TANAR/Rb23→YABV/Ve2+CIH_A/Fi21

TANAR/Rb23(BIDE/Rb13TEDENER/Rb33) →YABV/Ve2+GVL/Fe11+JAG_A/Fe5

CI/Rb21TANAR/Rb23 (NAM_A/Rb12BIDE/Rb13TEGUN/Rb32TEDENER/Rb33) →YABV/Ve2+GVL/Fe11+CIH_A/Fi21

TANAR/Rb23 YABV/Ve2+JAGA/Fe5+CIH_A/Fi21

TANAR/Rb23(BIDE/Rb13TEDENER/Rb33)→YABV/Ve2+GVL/Fe11+JAGA/Fe5+CIH_A/Fi21

The following is A0 Rule Set

BI/Rb11BIDE/Rb13CI/Rb21TANAR/Rb23TERE/Rb31 TEDENER/Rb33→YABV/Ve2+JAI/Fs11

BI/Rb11BIDE/Rb13CI/Rb21TANAR/Rb23TERE/Rb31 TEDENER/Rb33 (NAM_A/Rb12BIDE/Rb13CIM_A/Rb22TANAR/Rb23TEGUN/Rb32TEDENER/Rb33) →YABV/Ve2+GVL/Fe11+JAI/Fs11

BIDE/Rb13TANAR/Rb23TEDENER/Rb33→YABV/Ve2+JAGA/Fe5+JAI/Fs11

BI/Rb11BIDE/Rb13CI/Rb21TANAR/Rb23TERE/Rb31 TEDENER/Rb33→YABV/Ve2+CIHA/Fi21+JAI/Fs11

BIDE/Rb13TANAR/Rb23TEDENER/Rb33(BIDE/Rb13TANAR/Rb23TEDENER/Rb33)→YAB V/Ve2+GVL/Fe11+JAGA/Fe5+JAI/Fs11

BIDE/Rb13TANAR/Rb23TEDENER/Rb33→YABV/Ve2+JAGA/Fe5+CIH_A/Fi21+JAI/Fs11

( )

BIDE/Rb13TANAR/Rb23TEDENER/Rb33(BIDE/Rb13TANAR/Rb23TEDENER/Rb33)

→YABV/Ve2+GVL/Fe11+JAGA/Fe5+CIH_A/Fi21+JAI/Fs11

……

Although these rules appear very complicated, there are certain laws and large coverage.We provide these rules to the computer through the training set of machine learning and other various channels so as to make up the deficiencies brought about by the "sparse data"of a small language, to improve the accuracy of machine learning , and to make the "learning""deeper".

A Semi-manual Annotation Approach for Large CAPT Speech Corpus

Yanlu Xie, Xin Wei, Wei Wang, Jinsong Zhang
Beijing Advanced Innovation Center for Language Resources
Beijing Language and Culture University, Beijing 100083, China

xieyanlu@blcu.edu.cn blcuweixing@163.com vickyyzq@126.com jinsong.zhang@blcu.edu.cn

Abstract

Annotation plays an important roles in speech database. However annotation is time and annotators consuming. This paper proposes to provide phoneme-level labeling candidates with the state-of-the-art ASR models. The annotators could manually choose the appropriate labels and make final decision. Also a posterior probability evaluation method is applied to measure the annotation results. BLCU-SAIT speech corpus, a corpus aimed at computer aided pronunciation training (CAPT) is labeled with the annotation approach. Experimental results show that the mean consistency rate of manual labels is 87.2%. The posterior F1 score is 0.857. The annotation problems are converted from the open-ended questions to multiple-choice questions with the method. And the annotation results meet the requirements of CAPT systems.

Keywords: annotation, manual label, F1 score

1. Introduction

Annotation plays an paramount roles in speech database (Bird, S. 2001), especially in the database for language learning. For instance, the annotation is rewarding in studying the language phenomenon and in developing technology in assisting language learning. More and more interlanguage speech database is developed for the second language learning task recently. The scale of the database becomes larger. For example, the iCALL consists of 90,841 utterances from 305 speakers (Chen, N. F. 2016), the ERJ consists of 68,000 utterances from 200 speakers(Minematsu, N. 2002). It is a difficult task to annotate so much speech data manually. And the accuracy of the annotation results is questionable.

Some researchers have proposed Computer-Aided Annotation methods. CHAT (Codes for the Human Analysis of Transcripts) provided the instrument for producing and analyzing data (MacWhinney, 2000). DARCLE Annotation Scheme (DAS) proposed a workflow for annotating long natural language recordings. The transcripts and speech boundaries could be labeled automatically by some tools(Marisa Casillas, 2017). SLAM and Speech Analyzer POSCAT even could automatic give phone level labels (Kim, B., 2000)(Godwinjones, R. 2009) to the annotators.

These methods help to relief the human burden in annotation of some fields. However in some specific task, transcripts and speech boundaries are insufficient. For example, some CAPT (Computer-Aided Pronunciation Training) systems could detect mispronunciation and provide multi-level feedback (e.g., pronunciation score and phone substitution) to guide L2 learners to practice their pronunciation(Yingming Gao, 2015) (Yanlu Xie, 2016) (Leyuan Qu, 2016).

The performance of these systems is highly dependent on the quality of phone-level labeling of the non-native corpus. In fact, labeling non-native speech data is much more challenging than labeling native speech data, especially facing non-native mispronunciations. Moreover, as phonetic annotation is a subjective task, the familiarity of annotation conventions and psychological factors will also greatly affect the annotation consistency rate. Therefore, it is necessary to develop an automatic speech annotation system to assist human annotation.

Our recently proposed CAPT framework requires a large amount of phoneme labels, so this paper mainly focus on labeling phoneme-level mispronunciation patterns. Therefore, in this paper we attempted to use state-of-the-art ASR models based annotation system to automatically label a Chinese L2 speech corpus, then annotators were asked to check the detection results and make a final annotation.

Due to the difficulty of labeling non-native mispronunciations, the percentage of consistency between annotators is not always so high. Also the ground truth of the phone-level labeling is controversial. The performance of the annotation could not be indicated by the consistency merely. In order to measure the labeling precisely, a posterior probability annotation evaluation method is proposed.

The rest of this paper is organized as follow: Section II presents annotation framework, including automatic labels, manual labels and annotation evaluation criterion. Section III gives a brief description of the annotation corpus. Section IV shows experiments and results. Conclusions are given in Section V.

2. Annotation Methods

The annotation procedure could be divided into two parts: automatic label and manual label. In the automatic label part, the automatic speech recognition system will identify the possible erroneous (segmental and tonal) and label them. In the manual labeling part, human will decide which erroneous will be labeled finally.

2.1 Automatic Label

The Automatic label procedure will automatically label the boundaries and the possible erroneous phones.

Firstly an automatic speech recognizer is used to force-align the speech data into phonetic segments of Initials and Finals, and different levels of phonetic boundaries are assigned properly(Cao W 2010). After automatic mispronunciation detection is done, the erroneous phones which speech recognizer identified are transcribed in

Pinyin. Thus the mispronunciation types are labeled to assist annotators in making the final decision.

The illustration of the detection system is provided in Figure 1. In the acoustic module, we compare Long-Short Term Memory (LSTM) and Chain model which are both the state-of-the-art methods used in the ASR system.

The chain model we used in this study was introducted firstly by Povey et al. (D.Povey, Vijayaditya Peddinti, 2016) named as 'lattice-free maximum mutual information' (LF-MMI). The chain model has several differences, compared to traditional DNN-HMM model. This model use a three time smaller frame rate at the output of the neural network, which can significantly reduce a quantities of computation required in the test time and make real-time decoding much faster. Because of reducing the frame rate, unlike convential HMM topology, this model use a topology that can be traversed in one frame. During the training procedure of chain model, a forward-backward algorithm is run to estimate the sequence corresponding to the transcript (Juho Leinonen et al, 2018).

In the decoding module, we substitute the original grammar with an expanded grammar according to the length of input speech. For example, the input speech contains two syllables, then the corresponding expanded grammar will be limited to give a detection result with two syllables. After decoding, we will obtain Top-2 results based on the likelihoods of the output layer. The top two recognized results of the two models are given to annotators as the reference. Because of lacking of inter-Chinese databases, we used some Chinese databases to train the acoustic model.

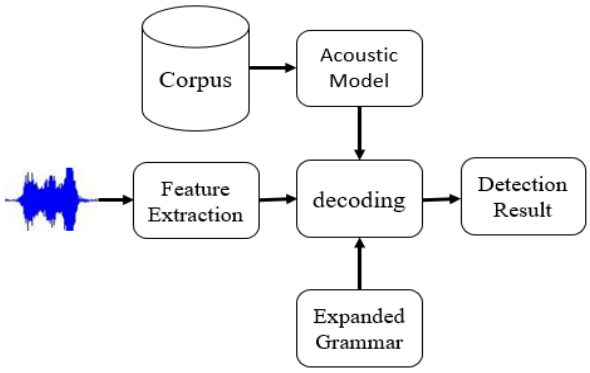

Fig.1 Flow chart of the detection framework

2.2 Manual Label

With the automatic labels, annotators will further check and label the data. The annotation problems are converted from the open-ended questions to multiple-choice questions with the method

Firstly annotators will judge if Initials or Finals in a bi-syllable word is similar with native's pronunciation or not. Then annotators will label the bi-syllable using Pinyin. If Pinyin could not remark the erroneous, PET diacritics will be used to describe the erroneous tendencies(Cao W 2010). For instance, as to the word 'ba ba'(father), if the pronunciation of 'a' is not native-like enough, and is more like 'e' in Chinese. Thus 'e' is used to indicate that the error sound is between 'a' and 'e'. If 'a' sounds like a native-like 'e', '/e' is used to reveal that 'a' is replaced by a standard Chinese 'e' sound. If the place of articulation of 'a' is too far behind, '{-}' is used to show backing of 'a' according to PET annotation conventions(Cao W 2010). All the work is deal with the software Praat.

A annotation example is shown in Fig.2. The first four tiers are corresponding to the orthographic given by speech recognizers. The fifth and the sixth tiers are respectively automatic results and manual annotation tier. In the sixth tier, there are two kinds of annotation symbols. If an error is described in Pinyin, it is annotated outside curly braces '{}'. On the contrary, PET diacritics are entered into '{}'.

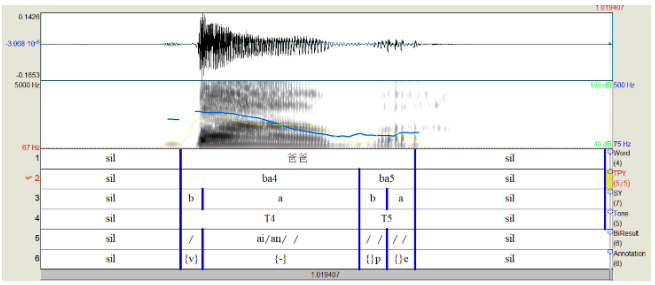

Figure 2. An annotation example

2.3 Posterior Probability Annotation Evaluation

The Mean consistency rate (MCR) with respect to each pair of annotators is widely applied in measuring the annotations results and the agreements. However the consistency rate is not comprehensive in measuring the binary classification. As an extreme case, if the erroneous is very little and one annotator is lazy and labels zero erroneous. The consistency rate will also be high.

In statistical analysis, the F1 score is a measure of a test's accuracy. It considers both the precision p and the recall r of the test to compute the score.

$$F_1 = \frac{2\,Precision*Recall}{Precision+Recall} \quad (1)$$

The F1 score assumes that false negatives, true negatives, false positives and true positives are certain. But in terms of annotation, especially for non-native mispronunciations, the four values are not certain. Thus we proposed Posterior F1(F1p). F1p could measure the F1 score and the consistency rate together. So the uncertainty could be considered in the formula.

$$F_1p = \frac{2\,Precision*Recall}{Precision+Recall} * MCR \quad (2)$$

3. Annotation Corpus

The annotation corpus used here is BLCU-SAIT corpus. BLCU-SAIT is an interlanguage speech corpus aiming at Chinese learning. This corpus is composed of four sections which cover most of the Chinese phoneme types and tri-tone types bounded by prosodic boundary using a 103 sentence set.

The corpus is divided into four parts.

(1)103 declarative sentences.
(2)237 bi-syllable words. These words cover 97% of segmental phonemes and all the 20 kinds of bi-tone types in Mandarin.
(3) 1520 tonal syllables .
(4)A discourse (The North Wind and the Sun).

302 non-native speakers have been recruited to record the corpus. The mean age of all the speakers is 23 years with a standard deviation of 4.1 years. In that, 66% are female speakers, 34% are male. All the speakers stay in China and have studied Chinese for a few years.
The corpus was recorded in studio using USB M-audio sound card, a SHURE microphones, a software of recorder PC3.0. The data was recorded into 16 bits pulse-code modulation (PCM), sampled at 16 kHz. In order to minimize the effect of growing familiarity with the order of difficulty affecting the quality of the recording, the subpart of the data was presented in a randomized order.

4. Annotation Results

18 native speakers from north China are selected to annotate the corpus. The annotation is divided into two phases. In the first phase, 237 bi-syllable words spoken by 156 speakers are annotated. There are totally 44,304 words had been annotated. The speakers were from four countries shown as table 1.

Table 1: Speaker numbers of annotated data

		Speaker number
Country	Korea	19
	Russia	44
	Japan	45
	Kazakhstan	48
Totally number		156

In the automatic label phase, the bi-syllable words are decoded by the Long-Short Term Memory (LSTM) and Chain models. The acoustic feature used in this study is Mel-Frequency Cepstral Coefficient (MFCC). The input feature is a 39-dimension MFCC+Δ+ΔΔ vector. After forced-alignment, context-dependent (CD) feature labels are used to train corresponding neural network, which containing 6 hidden layers and 625 nodes. The Long-Short Term Memory (LSTM) and Chain models are trained with non-native speech corpus such as BLCU inter-Chinese corpus (Cao W 2010). Because of lacking of inter-Chinese databases, we also used several native speech corpus to train the acoustic model, including the Chinese National Hi-Tech Project 863, which containing 94000 utterances spoken by 160 speakers at about 100 hours (Sheng Gao 2010), and THCHS-30 (Dong Wang 2015), etc.
In the manual label phase, two annotators of each speaker are randomly assigned to avoid pairing effects. Each annotator will judge the automatic labels and label the bi-syllable using Pinyin. The third annotator will check and verify two annotators' results. It took about five months to finish the phase.
The final annotation results are shown in table 3, figure 3 and figure 4.

Table 2: Phoneme annotation results

	MCR (Mean consistency rate)	F1-a1	F1p-a1	F1-a2	F1p-a2
Japan	85.9%	0.981	0.842	0.981	0.843
Korea	87.1%	0.995	0.867	0.995	0.866
Kaza-khstan	87.6%	0.981	0.860	0.981	0.860
Russia	88.2%	0.972	0.858	0.972	0.857
means	87.2%	0.982	0.857	0.982	0.857

The consistency rate of phoneme annotations with respect to each pair of annotators was evaluated in percentage agreements. The ratios range from 72% to 97% and average as 87.2%. As shown in table 3 and figure 2. The results can be regarded as good for the nature of phonetic labels. Compared with the previous manual annotation results, the consistency rate of the two annotators in this study raised from 80.7% to 87.2%, the consistency rate is improved remarkably(Cao W 2010). The main reason of the improvement maybe that annotation problems are converted from the open-ended questions to multiple-choice questions. The annotators could choose the right answers from the automatic labels.

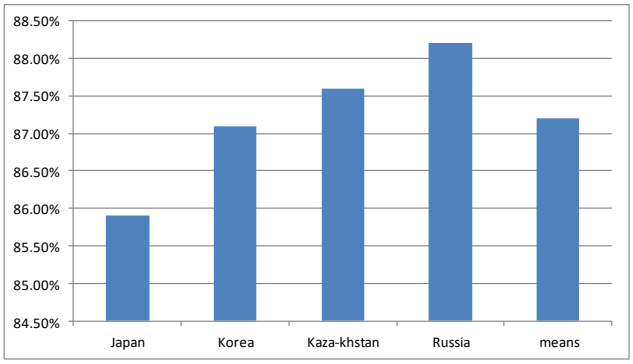

Fig.3 Mean consistency rate of each annotators

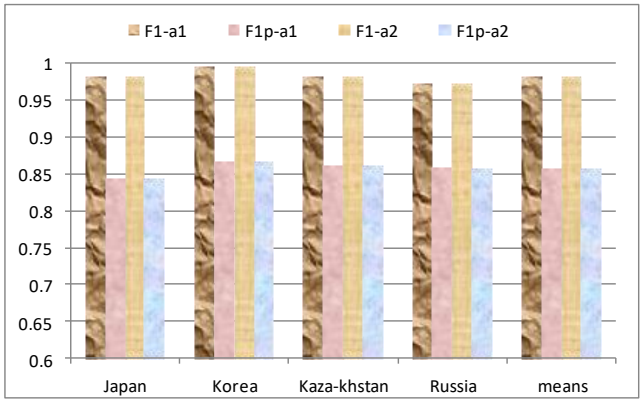

Fig.4 two F1 scores of the two annotators

Furthermore, F1 score is calculated to evaluate the performance. Since the standard answers of mispronunciation are unknown. Granted that the third annotator's label result is the ground truth. Thus we can get two F1 score. F1-a1 and F1-a2 are the F1 score of the first annotator and the second annotator respectively. As shown in table 2 and figure 4, F1-a1 and F1-a2 are extremely high. It means the difference between the two annotators is slight. In fact the third annotator's label result is still unreliable. F1-a1 and F1-a2 is not the reliable scores.

In order to measure the results more reliable, Posterior F1 is proposed. From formula (2), F1p-a1 and F1p-a2 are calculated. The results show that the new F1 drops a little as to the original F1. The ground truth is unknown as to the label problem. So the original F1 is not so precisely. The real F1 is unable to be calculated and shall be smaller than the original F1. Thus the Posterior F1 will be more reasonable. It considers the variance between the third annotator's label result and the ground truth. Even the mean consistency rate(MCR) is not equal to the actual variance. It is proportional to the actual variance. Thus the F1p-a1 and F1p-a2 could reflect the true F1. F1p-a1 and F1p-a2 is similar. It shows that the two annotators' performance is similar. The method could reduce the label ability between the annotators and make the results more objective.

5. Conclusion

This paper mainly focus on labeling phoneme-level mispronunciation patterns. In order to lighten the workload of human annotators, we attempt to use state-of-the-art ASR models based annotation system to automatically label a Chinese L2 speech corpus, then annotators could check the detection results and make a final annotation.

156 speakers' bi-syllable from 4 language backgrounds as pilot data were manually labeled, using both Pinyin, and the PET labeling system. The results show that mean consistency rate of manual labels is 87.2%. The posterior F1 score is 0.857. The results consistency rate is higher than previous report. So the annotation database could applied in the CAPT system. The posterior F1 score shows that the performance of the annotation could be improved further. Also the alternative labels annotated by the ASR models could help human annotators making final decision. They could choose one from four answers. Without the alternative labels, they will choose one from all the initials and finals. And the automatic labels also provide the possible mispronunciation for the human annotators. The annotation problems are converted from the open-ended questions to multiple-choice questions with the method.

In the near future, further efforts will be made to improve the system and more data will be used to develop CAPT systems. Also, the other part of BLCU-SAIT corpus, such as 103 declarative sentences will be labeled.

6. Acknowledgements

This work is supported by Wutong Innovation Platform of Beijing Language and Culture University (16PT05), Advanced Innovation Center for Language Resource and Intelligence(KYR17005), Research Funds of State Language Commission(ZDI135-51). The first author is corresponding author.

7. Bibliographical References

B. MacWhinney,(2000) The CHILDES Project: Tools for Analyzing Talk. 3rd Edition. Mahwah, NJ: Lawrence Erlbaum Associates,

Bird, S., & Liberman, M. (2001). A formal framework for linguistic annotation. Speech Communication,, 23-60.

Cao W, Wang D, Zhang J, et al.(2010) "Developing a Chinese L2 speech database of Japanese learners with narrow-phonetic labels for computer assisted pronunciation training". Eleventh Annual Conference of the International Speech Communication Association, .

Chen, N. F., Wee, D., Tong, R., Ma, B., & Li, H. (2016). Large-scale characterization of non-native Mandarin Chinese spoken by speakers of European origin. Speech Communication,, 46-56.

Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur (2016), "Purely sequence-trained neural networks for ASR based on lattice-free MMI". INTERSPEECH.

Godwinjones, R. (2009). Emerging Technologies Speech Tools and Technologies.. Language Learning & Technology, 13(3), 4-11.

Juho Leinonen, Peter Smit, Sami Virpioja, Mikko Kurimo (2018), "New Baseline in Automatic Speech Recognition for Northern Sámi". Proceedings of the 4th International Workshop for Computational Linguistics for Uralic Languages.

Kim, B., Lee, J., Cha, J., & Lee, G. (2000). POSCAT: A Morpheme-based Speech Corpus Annotation Tool.. language resources and evaluation.

Leyuan Qu, Yanlu Xie, Jinsong Zhang (2016) , "Senone log-likelihood ratios based articulatory features in pronunciation erroneous tendency detecting" in Proc.ISCSLP

Marisa Casillas etc(2017)A New Workflow for Semi-automatized Annotations: Tests with Long-Form Naturalistic Recordings of Childrens Language Environments INTERSPEECH , Stockholm, Sweden

Minematsu, N., Tomiyama, Y., Yoshimoto, K., Shimizu, K., Nakagawa, S., Dantsuji, M., & Makino, S. (2002). English Speech Database Read by Japanese Learners for CALL System Development.. language resources and evaluation.

S. Sheng Gao, Bo Xu, Hong Zhang, et al (2010), "Update of Progress of Sinohear: Advanced Mandarin LVCSR System At NLPR," in Proc. ICSLP.

Wang, D., & Zhang, X. (2015). Thchs-30 : a free chinese speech corpus. Computer Science.

Yanlu Xie, Mark Hasegawa-Johnson, Leyuan Qu, Jinsong Zhang (2016) "Landmark of Mandarin nasal codas and its application in pronunciation error detection" IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

Yingming Gao, Yanlu Xie, Wen Cao, Jinsong Zhang, (2015)"A Study on Robust Detection of Pronunciation Erroneous Tendency Based on Deep Neural Network", INTERSPEECH .