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

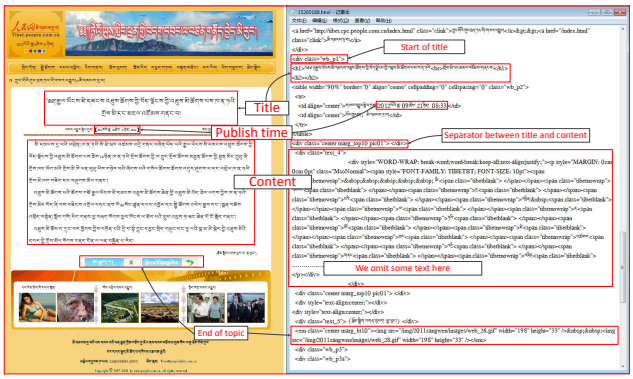

Figure 1: The structure of a web page from "China Tibet Online".

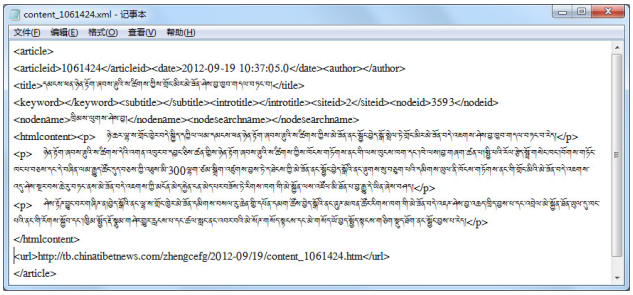

Figure 2: The text extracted from a page from "China Tibet news", in XML format.

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.

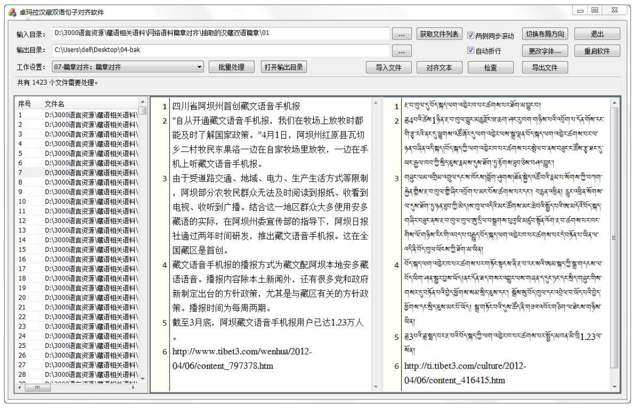

Figure 3: Chinese Tibetan sentence alignment tool.

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་བཤེ་ས་ཁྲི་122ཨོང་དཔྱ་དང་། རྒྱ་མཚོའི་རིས་ལས་མཐོ་ཚད་ཆསྟེམས་སུ་སྨྲི་4000ཡན་ཆེད་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་བཤེ་ས་ཁྲི་122ཨོང་དཔྱ་དང་།
	Chinese	平均海拔在4000米以上,
	Tibetan	རྒྱ་མཚོའི་རིས་ལས་མཐོ་ཚད་ཆསྟེམས་སུ་སྨྲི་4000ཡན་ཆེད་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.

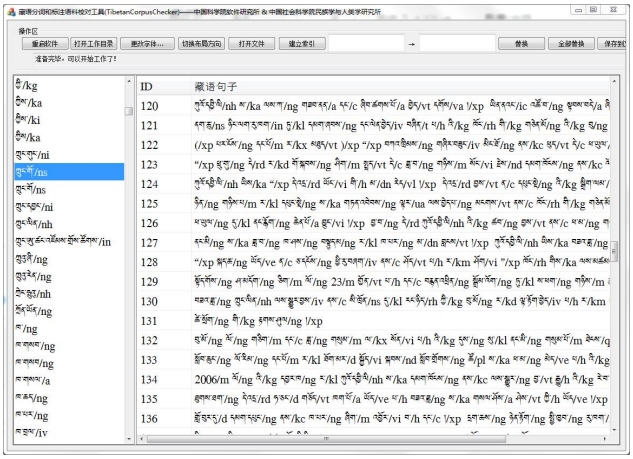

Figure 4: Tibetan annotation inconsistence checking tool.

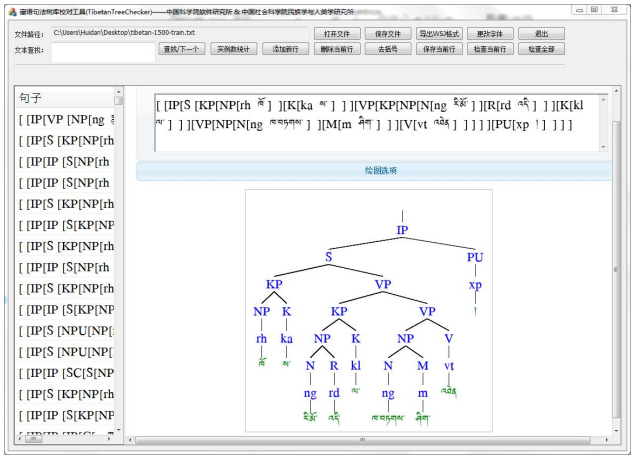

Figure 5: Tool for building Tibetan tree bank.

- སྒྲིན་དཀར་/ng རི་མོ་/kl ཐིས་/vt ཥ་/h ནང་བཞིན་/ua མཇོས་/a །/xp
- བར་སྣང་/ng མཐོ་རས་/ng ཤུ་/kl ཁྲུང་ཁྲུང་/ng ཁྱུ་/q གཅིག་/m བྱལ་བསྒྲིགས་/iv ནས་/c སྤྲོ་སྒྲོགས་/ng ཤུ་/kx འཕྲུར་/vi ཕྱིན་/t འདུག་/ve
- ས་གཞི་/ng ར་/kl གསེར་ཁུ་/ng ཐུགས་/vt ཥ་/h ནང་བཞིན་/ua ཟེར་ལམ་ལས་/ia ཅིར་/vt །/xp
- ཞིང་ཤ་/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་སྒོ་/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 རིད་))) (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 ར་)) (VP(TP(ng ཉིན་ཆུན་)) (VP(KP(NP(ng སྨོད་ལྡ་)) (K(kx ར་))) (V(vi འགྲོ་))))) (I(T(h ཀི་)) (E(ve ཡོད་)))) (PU(xp ।)))
- (IP(IP(S(NP(rh ར་)) (VP(KP(NP(N(ng ཆ་བདང་)) (NP(N(ng སོད་ཡིག་)) (N(ng སློབ་འབྲིང་)))) (K(kx ཏུ་)) (vi འགྲོ་)))) (I(T(h ཀི་)) (E(ve ཡོད་)))) (PU(xp ।)))
- (IP(S(KP(NP(rh ཁོང་)) (K(ka ཀྱིས་))) (VP(NP(NNP(KP(NP(ng ཆུ་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 ཁོང) (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.

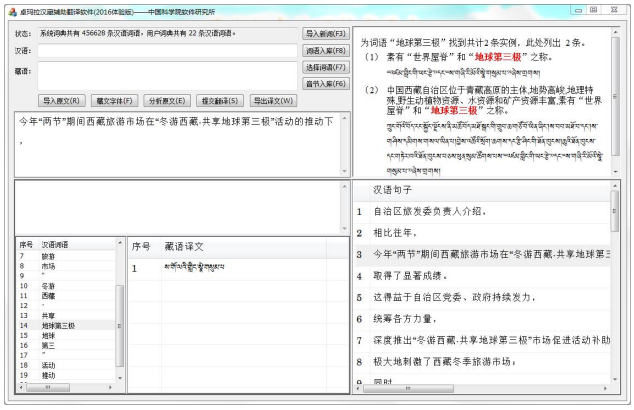

Figure 7: Chinese-Tibetan assistant translation tool.

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

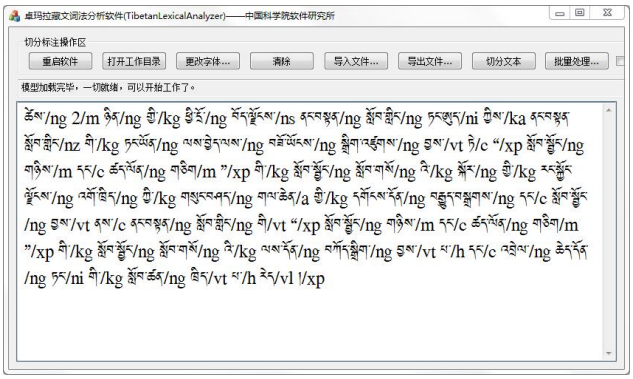

Figure 8: Tibetan lexical analyser.

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.