Building Large Scale Text Corpus for Tibetan Natural Language Processing by Extracting Text from Web Pages

Huidan LIU[1,2] *Minghua NUO*[1,2] *Jian WU*[1] *Yeping HE*[1]
(1) Institute of Software, Chinese Academy of Sciences, Beijing, China, 100190
(2) University of Chinese Academy of Sciences, Beijing, China, 100190
{huidan,minghua,wujian,yeping}@iscas.ac.cn

Abstract
In this paper, we propose an approach to build a large scale text corpus for Tibetan natural language processing. We find the distribution of Tibetan web pages on the internet with a crawler which can identify whether or not a web page contains Tibetan text. Three biggest web sites are selected, and topic pages are selected with a rule based method by checking the url. The layout structures of selected pages are analysed, and topic related information are extracted based on web site specific rules. Consequently, we get a corpus including more than 65 thousands documents, nearly 1.59 million sentences or 35 million syllables in total.

Title and Abstract in Chinese

抽取网页文本为藏文自然语言处理构建大规模文本语料库

在本文中，我们提出了一种为藏文自然语言处理构建大规模文本语料库的方法。我们首先使用网络爬虫技术，并结合藏文网页的编码识别技术判断一个网页中是否包含藏文文本，并据此考察互联网上藏文网页的分布情况。然后，我们选择了三个最大的藏文网站，根据网页的 URL，利用预先定义的规则，判断网页是 Hub 页面还是 Topic 页面。之后，我们分析了每个网站的 Topic 页面的布局结构特点，并利用正则表达式编制了相应的 Topic 相关文本的抽取规则。采用上述方法，我们构建了一个包含 6.5 万文档，共计 159 万句、3500 万藏文音节字的文本语料库。

Keywords: Tibetan, text corpus, web page, crawler, information extraction .

Keywords in Chinese: 藏文, 文本语料, 网页, 爬虫, 信息抽取.

Proceedings of the 10th Workshop on Asian Language Resources, pages 11–20,
COLING 2012, Mumbai, December 2012.

1 Introduction

Text corpora are the basis of natural language processing. But text corpora for Tibetan are seldom reported. It's an urgent task to build text corpora for Tibetan. As the web is a large data source, we have been seeking methods to get text from the web to build a large scale Tibetan text corpus. This paper reports the work.

The paper is organized as follows: In Section 2 we recall related work on Tibetan corpora and web as corpus for other language. In Section 3, we describe our research on the distribution of Tibetan web pages, then propose the strategy and methods to select web sites, get web pages and extract text from them. We introduce the corpus in Section 4 and then concludes the paper.

2 Related work

We review the work related to Tibetan corpora in this section. As we are reporting our work on getting corpus from the web, we also review the work on "web as corpus".

2.1 Reported Tibetan text corpora

Currently, the reported Tibetan corpora are all task-oriented, mainly for word segmentation and POS tagging. Chen et al. (2003a,b) built a corpus including 500 sentences (5890 words) as the test set. Caizhijie (2009a,b) also built a corpus including about 800 Kb text. Sun et al. (2009, 2010) used a corpus including 435 sentences (4067 words) as the test set in their research. These corpora are all for word segmentation. Norbu et al. (2010) described the initial effort in segmenting the Dzongkha (Tibetan) scripts. Their experiments are made on 8 corpora in different domains, which include only 714 words in total. Chungku et al. (2010) described the Dzongkha corpus for part-of-speech tagging and proposed a tag set containing 66 tags which is applied to annotate their corpus. The corpus contains 570247 tokens in 7 domains.

From the merely reports, we find that not only the number of corpora but also the scales of them are both very small, which shows that Tibetan text corpora are far from enough.

2.2 Web as corpus

In recent years, as the internet grows rapidly, it's already an over large data source and has been increasingly used as a source of linguistic data (Kilgarriff and Grefenstette, 2003). Many researchers has begin to building corpora with web text. Boleda et al. (2006); Zuraw (2006); Guevara (2010); Dickinson et al. (2010) presented the monolingual corpora for Catalan, Tagalog, Norwegian and Korean respectively which are built by crawling the web. Resnik (1998, 1999) developed "STRAND" while Chen and Nie (2000) also developed "PTMiner" to mining parallel bilingual text from the web. We are inspired by those work to build a large scale text corpus for Tibetan natural language processing.

But web pages are semi-structured data, it's a problem how to extract only topic related text. Cai et al. (2003) presents an automatic top-down, tag-tree independent approach to detect web content structure. In this paper, we will adopt the idea to analyse the layout structure of the web page, but use more simple rules to extract text.

3 Distribution of Tibetan web pages on the internet

We use a crawler with a seed url list including a certain Tibetan web sites. Then, with a Tibetan web page examiner, the crawler checks each fetched web page whether or not there is some Tibetan text in it. If Tibetan text is found in the page, then urls of all pages it links to will be append to the fetching list. This procedure continues until the fetching list is empty, which means that there is no new Tibetan web pages are found. The procedure is also described in Figure 1.

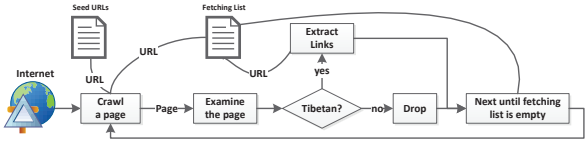

Figure 1: The procedure of finding Tibetan web pages.

Implementing this method, we build a Tibetan web text mining system. It starts to run on January 12 in 2011, and keeps running until now. Data are collected on April 13 in 2012. It's shown that the mining system find 150 Tibetan web sites and more than 130000 web pages after deduplication. Table 1 shows the web page numbers of the biggest 10 web sites.

				accumulative	
Order	URL	♯page	(%)	♯page	(%)
1	http://tb.chinatibetnews.com	18,160	13.79%	18,160	13.79%
2	http://tibet.people.com.cn	12,343	9.37%	30,503	23.16%
3	http://ti.tibet3.com	11,923	9.05%	42,426	32.21%
4	http://tb.tibet.cn	9,177	6.97%	51,603	39.17%
5	http://tibet.cpc.people.com.cn	6,251	4.75%	57,854	43.92%
6	http://blog.nbyzwhzx.com	4,203	3.19%	62,057	47.11%
7	http://blog.himalayabon.com	3,786	2.87%	65,843	49.98%
8	http://www.qhtb.cn	3,574	2.71%	69,417	52.70%
9	http://www.tibetcm.com	3,462	2.63%	72,879	55.32%
10	http://ti.gzznews.com	3,358	2.55%	76,237	57.87%

Table 1: Page numbers of the 10 biggest Tibetan web sites.

From Table 1 we see that nearly half (49.98%) of the web pages are intensively distributed in the 7 web sites, which is a plus factor for us to extract Tibetan text from web pages, because we can focus on only some biggest web sites.

4 Get Tibetan text from the web

In this section, we report the methods to select web sites, web pages and to extract text from web pages.

4.1 Selection of web sites

Because there are not so many Tibetan web sites and the web pages are intensively distributed, it's practical for us to use manually generated site specific rules to extract text one site by one site. With this idea, three biggest websites are selected. Table 2 shows in-

formation about them. As the sites are hold by newspaper offices, which have high quality standards for publishing, the quality of the corpus is guaranteed.

Order	Host URL	Site Name	Holder
1	http://tb.chinatibetnews.com	China Tibet News	Tibet Daily
2	http://tibet.people.com.cn	China Tibet Online	People's Daily
3	http://ti.tibet3.com	Tibetan's Web of China	Qinghai Daily

Table 2: Information about the selected web sites.

4.2 Selection of web pages

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 text from the web pages. We only care about the topic pages rather than the hub pages. But in the URL list, which one is the URL of a topic page?

Site	Example URLs
China Tibet News	http://tb.chinatibetnews.com/news/2012-02/16/content_884280.htm
	http://tb.chinatibetnews.com/xzmeishi/2011-12/05/content_831210.htm
	http://tb.chinatibetnews.com/xzzongjiao/2011-10/21/content_798694.htm
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 3: Example URLs of topic pages in the three sites.

Site	Example URLs
China Tibet News	http://tb.chinatibetnews.com/xzpinglun/node_698.htm
	http://tb.chinatibetnews.com/shehuiminsheng/index.html
	http://tb.chinatibetnews.com/xzcaijing/index.html
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 4: Example URLs of hub pages in the three sites.

Table 3 and Table 4 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 "China Tibet News" and "Tibetan Web of China" have the pattern of "{host}/{column}/{year}-{month}/{date}/content_{articleid}.htm". Everyone of them contains the string "content_".
- The hub URLs of "China Tibet News" and "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, while URL extraction are made on both the hub pages and the topic pages.

4.3 Text extraction

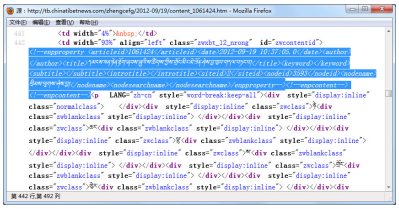

Figure 2: Commented text in a web page from "China Tibet News".

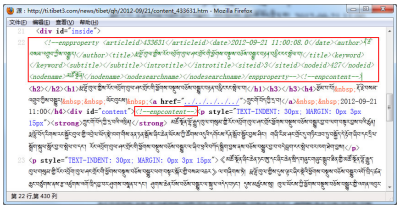

Figure 3: Commented text in a web page from "Tibetan's Web of China".

Then, we analyse each topic page to find whether there is a rule to extract topic related text. It's a surprise that we find some commented text from the html file of topic pages of "China Tibet News". Figure 2 shows the commented text in a web page[1] from "China Tibet News". The shadowed text in the figure shows many information about the page, including title, subtitle, publish date, author, and so on. Although some of the field values are kept empty, it provides us a simple method to extract those information. Unsurprisingly, the text following the shadowed text are the content of the topic, which is followed by another segment of commented text: "`<!--/enpcontent--><!--/enpcontent-->` ".

Then, we get the following rules.

- Tags and text between '<!–enpproperty>' and '/enpproperty–>' are useful information about the topic. which can be directly used as XML format text.
- HTML tags and text between the inner pair of '`<!--enpcontent-->`' and '`<!--/enpcontent-->`' are the content of the topic in HTML format.

What a big surprise! We find almost the same commented text in the topic pages from the third web site "Tibetan's Web of China". Figure 3 shows the HTML file of a web page[2] from it. But is it a coincidence? We get the information that both of the two web sites are using the same computer management system of news gathering and editing, which is a product of Beijing Founder Electronics company, to manage their articles and web pages.

We have no luck in processing pages from "China Tibet Online". But we still get a clue after analysing the structure of some web pages from this site. Figure 4 shows the structure of a web page[3] from this site. From the figure, we see that there are some HTML tags giving the boundaries of different text blocks, including:

- String ' ' indicates the start of the title, and the title is surrounded by HTML tags "<h1>" and "</h1>". The text between the following "<h2>" and "</h2>" may be the subtitle.

[1] `http://tb.chinatibetnews.com/zhengcefg/2012-09/19/content_1061424.htm`
[2] `http://ti.tibet3.com/news/tibet/qh/2012-09/21/content_433631.htm`
[3] `http://tibet.people.com.cn/15260188.html`

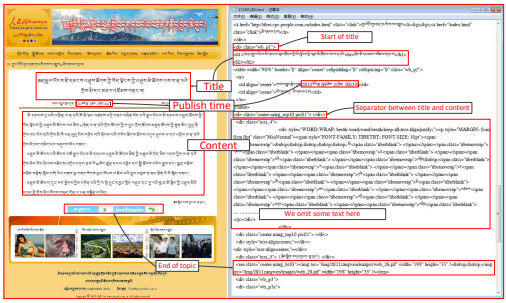

Figure 4: The structure of a web page from "China Tibet Online".

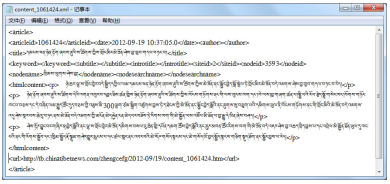

Figure 5: The text extracted from a page from "China Tibet news".

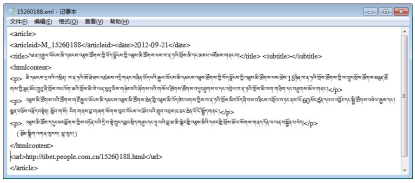

Figure 6: The text extracted from a page from "China Tibet Online".

- String ' ' is the separator between the title block and the content block.
- String ' ' indicates the start of the content block.
- String '...' indicates the end of the content block.

Taking advantage of the HTML tag information, we also can extract the topic related information from the web pages in the web site "China Tibet Online".

Then, we use another crawler, which is deployed with a larger link depth than the former crawler, to download web pages from the three web sites, and implement the rules mentioned above by regular expressions. Topic related text are extracted from the web pages. we filtered out most HTML tags, and convert them into XML format. Figure 5 and Figure 6 show the results.

5 The corpus

In this section, we have to introduce the counting units because Tibetan is different from many other languages. The corpus is introduced after that.

5.1 About the counting units

In Tibetan, syllables are separated by a delimiter known as "tsheg", which is simply a superscripted dot. Thus the syllable in Tibetan is a unit smaller than or equal to a word,

and meanwhile it's larger than or equal to a character, just like a "hanzi" in Chinese. Figure 7 shows the structure of a Tibetan word which is made up of two syllables and means "show" or "exhibition". Another delimiter known as "shed" indicates the sentence boundary, which looks like a vertical pipe. Figure 8 shows a Tibetan sentence. In this paper, we use "sentence" and "syllable" as the units to give a more sensitive description of the scale of the corpus.

Figure 7: A Tibetan word.

ཁ་ལ་མི་སྡུག་པོ་འདི་ཁང་པ་གོང་ཆེན་པོ་ཞིག་གིས་བསོད།									
ཁ་ལ་	མི་	སྡུག་པོ་	འདི་	ཁང་པ	གོང་ཆེན་པོ་	ཞིག	གིས་	བསོད་	།
Yesterday	man	rich	this	house	expensive	an	bought	did	.
Yesterday this rich man bought an expensive house.									

Figure 8: A Tibetan sentence.

5.2 The scale of the corpus

We get more than 65 thousands documents from the three web sites, including nearly 1.59 million sentences or 35 million syllables in total. Table 5 shows the data about the corpus. The corpus may be used to make many kinds of statistics, such as character frequencies, syllable frequencies, to train the language models. It can also be used as a basis to build corpus for other natural language processing tasks.

Order	Site Name	♯document	(%)	♯sentence	♯syllable
1	China Tibet News	31,304	47.88%	815,000	20,264,896
2	China Tibet Online	18,558	28.39%	339,096	7,595,860
3	Tibetan's Web of China	15,515	23.73%	435,623	7,182,505
Total		65,377	100.00%	1,589,719	35,043,261

Table 5: The scale of the corpus.

Note that the document numbers in Table 5 are great than those in Table 1. It results from two factors:(1) New pages are created after the data in Table 1 are collected. (2) The crawler used to extract Tibetan text corpus is deployed with a bigger link depth than the one in section 3.

5.3 Domains in the corpus

Table 3 shows some URLs from the three web sites. We find that the URL of a page from "China Tibet News" or "Tibetan's Web of China" gives the evidence that which column or sub column, and consequently which domain it belongs to. The following examples shows it.

- 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 topic page about Tibetan foods, because "xz" is the abbreviated form of Chinese word "xizang" written in Pinyin[4], which means the Tibetan Autonomous Region, while "meishi" means "delicious food".
- Another url "http://ti.tibet3.com/medicine/2009-10/27/content_99171.htm" from "Tibetan's web of China" shows it belongs to "Tibetan Medicine" domain.

[4]A Latin transliteration method for Chinese.

With this method, we classify the documents into different domains. Table 6 and Table 7 show the numbers of documents from the two web sites , in different domains. Comparing the two tables, we find that "News" shares a large part of the documents, especially for those from "Tibetan's web of China", which is up to 86.88%. Documents from "China Tibet News" are more balanced. These parts of corpus can be used for text classification.

Order	Domain	♯document	(%)	♯sentence	♯syllable
1	Art	1,277	4.08%	49,250	614,269
2	Finance & Economy	503	1.61%	9,785	268,098
3	History & Geometry	443	1.42%	8,546	151,663
4	News	10,395	33.21%	272,745	7,446,822
5	Picture	2,548	8.14%	16,935	346,175
6	Politics & Law	5,329	17.02%	181,545	4,659,379
7	Rural Life	1,238	3.95%	23,891	646,246
8	Social Life	473	1.51%	6,385	173,766
9	Special Issues	6,100	19.49%	175,173	4,561,724
10	Technology & Education	943	3.01%	24,716	600,806
11	Tibetan Buddhism	792	2.53%	22,318	352,642
12	Tibetan Food	92	0.29%	1,682	16,640
13	Tibetan Medicine	508	1.62%	10,436	155,372
14	Tour	663	2.12%	11,593	271,294
Total		31,304	100.00%	815,000	20,264,896

Table 6: Domains in the documents from "China Tibet News".

Order	Domain	♯document	(%)	♯sentence	♯syllable
1	Art	77	0.50%	2,987	47,558
2	Culture	710	4.58%	86,155	860,747
3	Economy	73	0.47%	7,143	121,440
4	Education	11	0.07%	683	14,542
5	Music	78	0.50%	2,296	31,806
6	News	13,480	86.88%	284,337	5,218,266
7	Photo	63	0.41%	2,493	38,090
8	Policy	116	0.75%	7,062	128,992
9	Politics	124	0.80%	7,668	145,206
10	Special Issues	523	3.37%	17,537	309,100
11	Tibetan Medicine	107	0.69%	11,417	173,974
12	Tour	131	0.84%	5,489	86,773
13	Video	19	0.12%	314	5,493
14	Other	3	0.02%	42	518
Total		15,515	100.00%	435,623	7,182,505

Table 7: Domains in the documents from "Tibetan's web of China".

Conclusion and perspectives

In this paper, we proposed an approach to build a large scale text corpus for Tibetan natural language processing by extracting text from three biggest web sites. Consequently, we get a corpus including more than 65 thousands documents, nearly 1.59 million sentences or 35 million syllables in total. Parts of the corpus are classified into different domains. In the following work, we will build corpora for other tasks based on the present corpus.

Acknowledgements

The research is partially supported by National Science and Technology Major Project (No.2010ZX01036-001-002, No.2010ZX01037-001-002), National Science Foundation (No.61003117, No.61202219, No.61202220), Major Science and Technology Projects in Press and Publishing (No.0610-1041BJNF2328/23, No.0610-1041BJNF2328/26), and CAS Action Plan for the Development of Western China (No.KGCX2-YW-512).

References

Boleda, G., Bott, S., Meza, R., Castillo, C., Badia, T., and López, V. (2006). Cucweb: a catalan corpus built from the web. In *Proceedings of the 2nd International Workshop on Web as Corpus*, pages 19–26. Association for Computational Linguistics.

Cai, D., Yu, S., Wen, J., and Ma, W. (2003). Vips: a visionbased page segmentation algorithm. Technical report, Microsoft Technical Report, MSR-TR-2003-79.

Caizhijie (2009a). The design of banzhida tibetan word segmentation system. In *the 12th Symposium on Chinese Minority Information Processing*.

Caizhijie (2009b). Identification of abbreviated word in tibetan word segmentation. *Journal of Chinese Information Processing*, 23(01):35–37.

Chen, J. and Nie, J.-Y. (2000). Automatic construction of parallel english-chinese corpus for cross-language information retrieval. In *Proceedings of the sixth conference on Applied natural language processing*, ANLC '00, pages 21–28, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chen, Y., Li, B., and Yu, S. (2003a). The design and implementation of a tibetan word segmentation system. *Journal of Chinese Information Processing*, 17(3):15–20.

Chen, Y., Li, B., Yu, S., and Lancuoji (2003b). An automatic tibetan segmentation scheme based on case auxiliary words and continuous features. *Applied Linguistics*, 2003(01):75–82.

Chungku, C., Rabgay, J., and Faaß, G. (2010). Building nlp resources for dzongkha: A tagset and a tagged corpus. In *Proceedings of the Eighth Workshop on Asian Language Resouces*, pages 103–110, Beijing, China. Coling 2010 Organizing Committee.

Dickinson, M., Israel, R., and Lee, S.-H. (2010). Building a korean web corpus for analyzing learner language. In *Proceedings of the NAACL HLT 2010 Sixth Web as Corpus Workshop*, pages 8–16, NAACL-HLT, Los Angeles. Association for Computational Linguistics.

Guevara, E. R. (2010). Nowac: a large web-based corpus for norwegian. In *Proceedings of the NAACL HLT 2010 Sixth Web as Corpus Workshop*, pages 1–7, NAACL-HLT, Los Angeles. Association for Computational Linguistics.

Kilgarriff, A. and Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. *Computational linguistics*, 29(3):333–347.

Norbu, S., Choejey, P., Dendup, T., Hussain, S., and Muaz, A. (2010). Dzongkha word segmentation. In *Proceedings of the Eighth Workshop on Asian Language Resouces*, pages 95–102, Beijing, China. Coling 2010 Organizing Committee.

Resnik, P. (1998). Parallel strands: A preliminary investigation into mining the web for bilingual text. *Machine Translation and the Information Soup*, pages 72–82.

Resnik, P. (1999). Mining the web for bilingual text. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 527–534. Association for Computational Linguistics.

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

Sun, Y., Luosangqiangba, Yang, R., and Zhao, X. (2009). Design of a tibetan automatic segmentation scheme. In *the 12th Symposium on Chinese Minority Information Processing*.

Sun, Y., Yan, X., Zhao, X., and Yang, G. (2010). A resolution of overlapping ambiguity in tibetan word segmentation. In *Proceedings of the 3rd International Conference on Computer Science and Information Technology*, pages 222–225.

Zuraw, K. (2006). Using the web as a phonological corpus: a case study from tagalog. In *Proceedings of the 2nd International Workshop on Web as Corpus*, pages 59–66. Association for Computational Linguistics.