Dzongkha Text Corpus

Chungku Chungku, Jurmey Rabgay, Pema Choejey
Department of Information Technology & Telecom
{chungku, jrabgay, pchoejey}@dit.gov.bt

Abstract

In this paper, we present the methodology for collecting and compiling text corpus in Dzongkha. The corpus resources is essential for developing applications such as Automatic Part of Speech Detection, Word Segmentation, Text to Speech and Morphological Analysis. This study resulted in building a corpus database containing at least 5 million words collected from relevant websites, print media and manually typed from printed documents. The corpus is tagged with automatic part of speech tagger to enrich the data and to make it more resourceful. In addition, the text corpus is transduced from their original format to an XML format compliant with the XML Corpus Encoding Standard (XCES).

1 Introduction

This is the first attempt ever made to build Dzongkha Text corpus. The objective of this research study is to develop a balanced Dzongkha text corpus which is maximally representative of rich linguistic diversity. This will provide us with huge language resources to be used for developing several natural language processing (henceforth NLP) tools.

A corpus from linguistic point of view is defined as a collection of transcribed speech or written text which have been selected and brought together as a source of data for linguistic research studies. Creation of text corpus from available resources was necessitate due to lack of electronic text in Dzongkha. The corpus was collected from wide range of sources such as the print media, electronic documents and text from websites.

At present the Dzongkha text corpus contains at least 5 million words which are divided into 8 domains and 13 sub domains. First the raw text undergoes preprocessing, that is the cleaning of data by maintaining its standard format in Unicode. Then it is tokenized into one word per line format using Dzongkha word segmentation[1] tool developed based on lexicon and the longest string matching method.

In order to increase the utility of the corpus, it is annotated using an Automatic Dzongkha part of speech (henceforth called POS) tagger tool (Chungku, et al., 2010) based on Tree tagger (Schmid, 1994). The corpus is automatically annotated with a grammatical tag set containing 66 tags. The corpus is currently being used for other linguistic research studies such as word segmentation, text to speech, morphological analysis and automatic annotation.

Furthermore, to make corpus readily usable by the computers, it is encoded using markup language which marks important language structure including the boundary and POS of each word, sentence, phrase, sections, headings and the meta-textual information. Hence, the text corpus is transduced from the original formats to an XML format with XML corpus encoding standard XCES (Ide, et al., 2000).

Section 2 describes the literature review of the language, section 3 presents the methodology of compilation. Section 4 describes the future work and section 5 concludes the paper.

2 Literature Review

Dzongkha Language

Dzongkha is the national and official language of Bhutan spoken by about 130,000 people in Bhutan. It is a Sino-Tibetan language and is closely related to Tibetan, which was introduced by Thonmi Sambhota. It is an alphabetic

[1] This tool was develop at NECTEC (National Electronics and Computer Technology Center), Thailand.

language with consonants, vowels, phonemes and phonetics characteristics. Like many of the alphabets of India and South East Asia,

Dzongkha Script is syllabic[2]. A syllable can contain one character or as many as six characters. Linguistic words may contain one or more syllables which is separated by superscripted dot called “tsheg” that serves as syllable and phrase delimiters. The sentence is terminated with vertical stroke called “shed”.

Related Works

Initial text corpus, however in lesser size, were used in Dzongkha text-to-speech (Sherpa, et al., 2008), Word Segmentation (Norbu, et al., 2010) and POS Tagger. It was also used for analysis of POS tagset based on Penn Tree Bank. But the corpus was raw in nature, unformatted and unstructured, and had no linguistic annotation.

This study, therefore, expands the initial text corpus base into larger corpus database which is well formated, structured and linguistically meaningful.

3 Methodology

The process of building text corpus involves two important steps: the planning (design stage) and the execution (creation stage) as described below:

3.1 Design stage

The process of collecting a corpus to its broadest and widest range should be based on its purpose. Our goal was to build corpus that is to be used for multipurpose application both in language technology and general linguistic research. Thus from the initial planning stage certain design principles, selection criteria and classification features was drawn up.

a) Selection Criteria

Three main independent selection criteria, namely domain, medium and time were considered.

Domain

The domain of a text indicates the kind of writing it contains. The data was collected from broad range of domains. The table (1) shows the portable domain percentage:

Domain	Share %
Text Books	25%
Mass Media	50%
World wide Web	10%
Others	15%
Total	100%

Table (1): Portable domain percentage

The corpus contains the written texts from informative writings in the fields of science and medicine, world affairs and political, and leisure. It also contains descriptive writing in the fields of arts, entertainment and finance. More details on domain percentage are found in, cf. table (2).

Medium

The medium of text indicates the kind of publication in which it occurs. The broad classification was made as followed.

- 60% contains periodicals (newspaper etc.)
- 25% contains written text from books
- 10% includes other kinds of miscellaneous published material (brochures, advertising leaflets, manual, etc.)
- 5% includes unpublished written material such as poems, songs, sayings, personal letters, essays and memorandum, etc.

Time

The time criterion refers to the date of publication of the text. Since this is the first attempt made to build Dzongkha text corpus, we collected electronic newspapers published since 2009. This condition for books was relaxed because of its rich linguistic diversity and also due to shortage of electronic books.

b) Classification features

Apart from selection criteria, a few number of classification features were identified for the text in the corpus. Our intention was to make an appropriate level of variation within each criterion, so classification criteria includes following features:

- Sample size or number of words
- Topic or subject of the text

[2] Omniglot.com. “Tibetan” http://omniglot.com/writing/tibetan.htm

- Authors name
- Text genre and text type
- Source of text

3.2 Creation stage

Creation of Dzongkha text corpus includes following steps:

a) **Copyright Permission**

Request for copyright permission or clearance was formally sought from the concerned agencies before creation of the text corpus. Prior approval was obtained from the right owners on conditions listed below:

- to allow their materials to be used for creation of text corpus,
- that the text corpus be used for the academic research and not for commercial purpose,

b) Text Collection

Corpus acquisition

Text corpus from different sources and belonging to different domains, as described in table (2), were collected. The acquired corpus contains 5 million tokens (words) approximately.

Domain	Share %	Text Type
1) World Affairs/Political	8 %	Informative
2) Science/Medicine	2 %	Informative
3) Arts/Entertainment	15 %	Descriptive
4) Literature	35 %	Expository
5) Sports/Games	3 %	Procedural
6) Culture/History	7 %	Narrative
7) Finance	10 %	Descriptive
8) Leisure	20 %	Informative

Table (2): Textual domains contained in corpus

Apparently, as seen in the Table (2), the corpus is not well balanced due to lack of electronic text as most of the text are available in printed forms. The highest share (35%) of created corpus belongs to expository text and the lowest share (2%) belongs to informative text in the domain of Science and Medicine.

Corpus Extraction

Extraction of text from the websites was made by studying the richness in the linguistic structure of the text and considering the variety of domain required. Though there are tools for extraction of texts from websites such as web crawler, we stuck to the process of copying manually from websites (only few websites in Dzongkha is available).

Pre-processing

Cleaning Process: Text cleaning process is divided into two major steps. Firstly, the data gathered was standardized into Unicode character encoding scheme of UTF-8 given its flexibility and portability across platforms. Secondly, correction of spelling mistakes, removal of repetitive and foreign words and deletion of white spaces were performed.

Tokenization

Segmentation of text into words is one of the necessary preprocessing steps for linguistic annotation and also for the frequency analysis of words. POS tagging usually requires a one-token-per line format. This is achieved by using the process of word segmentation. As mention earlier (cf. section 2) Dzongkha belongs to the alphabetic Sino-Tibetan language and is written in continuous form. Therefore, there is no mark to identify word boundary between words.

Word Segmentation: The training data consisting of 40247 token of words was created from existing corpus by segmenting manually to achieve higher accuracy of word boundary. Using this training data and lexicon (dictionary), Dzongkha word segmentation tool (cf. section 1) was developed based on longest matching technique which is a dictionary based method. Then the whole text corpus was tokenized into one word per line based on characters like white space, punctuation marks, symbols etc. The segmentation accuracy of 85.69% is achieved.

c) Encoding of Texts

It is known that for text corpus to be usable by computers, it should be marked-up with its important textual information, in case of Dzongkha text corpus such as:

- The boundary and POS of each word
- Phrase, sections, headings, paragraphs and similar features in written texts
- Meta-textual information about the source or encoding of individual texts

This textual information and others are all encoded by standard mark-up language to help ensure that the corpus will be usable no matter what the local computational set-up may be.

In addition, texts is further transduced from their original formats to an XML format complaint with XML Corpus Encoding Standard (henceforth XCES) (Ide, et al., 2000). Each corpus file pertaining to different domains is stored in XML structure. The marking is done at word level. This format makes easier for developing web access application for corpus. The design of XCES was strongly influenced by the development of the Guidelines for Encoding of Electronic Text of the international Text Encoding Initiative (TEI).

We found that XML based format is more convenient for corpus since it:

- Supports Unicode
- Programming interface adaptable
- Simplicity, generality and usability mark-up language over internet

The following, cf. (1) shows the example of the how Dzongkha text corpus is encoded using XCES (sentence level).

(1) <p id=p1>

<s id=”p1s1”>ཀ་ལི་ཕྲུག་ལུ་འགྱོཝ་ད།།</s>

<s id=”p1s2”>ཁ་ལུ་ཟ་ནི་མིན་འདུག།</s>

<s id=”p1s3”>ག་ཏེ་བཤལ་ཏེ་འགྱོ་རུང་།།</s>

</p>

d) Linguistic Annotation of Text (“POS tagging”)

POS tagging means annotating each words with their respective POS label according to its definition and context. It provides significant information for linguistic researcher with morphological, syntactic or semantic kind. Such enriched data is useful, especially when designing higher level NLP tools.

In this research study, the morpho-syntactic annotation is automatically added to the entire corpus using a probabilistic tagger developed (Chungku, et al., 2010). Annotated texts produced by this automatic tagger uses tag set[3] containing 66 tags, its design is based on Penn Guidelines[4] (though there is some changes made to fit the language structure). The highest accuracy achieved by this automatic tagger is about 93.1%. The accuracy is further enhanced by performing manual post-edit thereby resulting in better annotated texts. Table (3) shows an example of how the process of automatic tagger takes place.

Input text	Output text	
Word	Word	POS tag
འབྲུག་	འབྲུག་	NNP
གི་	གི་	CG
རང་ལུགས་	རང་ལུགས་	NNP
འཆམ	འཆམ	NN
།	།	PUN

Table (3) Example of automatic annotation

e) Storage, Documentation and Web Interface

In the last stage, we manually added detailed descriptive information to each text in the form of header. The header information contains specific information such as author’s name, source of the text, etc. which is useful for computer programming.

A web based interface to provide easy access to the corpus is being developed. The corpus database thus built can be made available on CD ROM for research purposes.

4 Future Work

This is the first attempt ever made to build the corpus database. Therefore, there are enough rooms for improvements in terms of quality and usefulness.

Increasing the text corpus size may lead to further improvement in tagging and segmentation accuracies thereby leading to better quality annotated text corpus.

Tools for collection of text corpus may be explored. Optical character recognition (OCR) system being currently developed by the department may ease the collection process.

In addition, balancing corpus from broad ranges of domains and annotating text using other annotation techniques may improve the quality.

[3] The original Dzongkha tag set is described at http://www.panl10.net

[4] The Penn Guidelines can be downloaded from: http://www.cis.upenn.edu

5 Conclusion

Corpus is very important and is the basic resource for language processing and analysis. This document demonstrates the collection and compilation methodology of Dzongkha text corpus. It also demonstrates how the corpus is automatically annotated with automatic POS tagger. The corpus database contains 5 million tokens of words. The corpus is being used for Dzongkha word segmentation, automatic corpus tagger and advanced text-to-speech system.

Furthermore, it is expected that the corpus will become extremely useful for developing other language processing tools such as lexicon, machine translation and frequency analysis.

Acknowledgments

This research work was carried out as part of the PAN Localization Project (www.PANL10n.net) with the aid of grant from International Development Research Center (IDRC), Ottawa, Canada, administered through the Center of Research in Urdu Language Processing (CRU LP), National University of Computer and Emerging Sciences (NUCES), Pakistan. The research team would also like to thank media agencies (Bhutan Observer, Bhutan Broad Casting Service, Druk Nyeltshul, Bhutan Today) in Bhutan, Bhutan National library and Dzongkha Development commission for their valuable contribution that made this research successful.

References

British National Corpus. 2009. *Reference Guide for the British National Corpus(XML Edition).* Retrieved October 30, 2009, from http://www.natcorp.ox.ac.uk/

Chungku, Chungku, Gertrud Faaß and Jurmey Rabgay. 2010. *Building NLP resources for Dzongkha: A Tag set and a tagged Corpus.* Proceedings of the Eighth Workshop of Asian Language Resources (WS1), 23rd International Conference on Computational Linguistics (COLING 2010), 103-110, Beijing, China,

Nancy Ide, Laurent Romary, Patrice Bonhomme. 2000. *XCES: An XML-based standard for Linguistic Corpora.* In proceedings of the Second Annual conference on language Resources and Evaluation, , 825-30. Athens.

Nancy Ide, Randi Reppen, Keith Suerman. 2002. *The American National Corpus: More Than the Web can provide*. Retrieved November 1, 2010, from http://www.cs.vassar.edu.

Sithar Norbu, Pema Choejey, Tenzin Dendup, Sarmad Hussain, Ahmed Mauz. 2010. *Dzongkha Word Segmentation.* Proceedings of the Eighth Workshop of Asian Language Resources (WS1), 23rd International Conference on Computational Linguistics (COLING 2010), 95-102, Beijing, China.

Asif Iqbal Sarkar, Shahriar Hossain Pavel, and Mumit Khan. 2007. *Automatic Bangla Corpus Creation.* PAN Localization Working Papers, pages 22-26, 2004-2007.

Helmut Schmid. 1994. *Probabilistic Part-of-Speech Tagging Using Decision Tree.* Proceedings of the International Conference on New Methods in Language Processing, pages 44-49, Manchester, UK.

Eden Sherpa, Dawa Pemo, Dechen Choden, Anocha Rugchatjaroen, Ausdang Thangthai, Chai Wutiwatchai. 2008. *Pioneering Dzongkha text-to-speech Synthesis.* Proceedings of Oriental COCOSDA, pages 150-154, Kyoto, Japan.

Martin Wynne (Ed.). 2005. *Developing Linguistic Corpora: a Guide to Good Practice.* Oxford: Oxbow books.

XML Corpus Encoding Standard Document XCES 1.0.4. 2008. *XCES Corpus Encoding Standard for XML.* Retrieved August 27, 2009, from http://www.xces.org/