TS Wikipedia Corpus

What is TS Wikipedia Corpus?

TS Wikipedia Corpus Data Set is a collection of processed “Turkish Wikipedia pages”. The source of the data is Turkish wiki-dumps[1].
The set is a collection of eight *(8)* separate files which are named as[2]:

- TS_Wikipedia_ Raw_Data.xml (266 Mb)
- TS_Wikipedia_ Text_ID.xml (15 Mb)
- TS_Wikipedia_ Tokenized.xml (271 Mb)
- TS_Wikipedia_ POSTagged_Corpus.xml (1.4 Gb)
- TS_Wikipedia_ bi_gram.xml
  - Case-Sensitive (92 Mb)
  - Case-Insensitive[3] (91 Mb)
- TS_Wikipedia_ tri_gram.xml
  - Case-Sensitive (76 Mb)
  - Case-Insensitive[4] (79 Mb)

1- General Background

As known, Wikipedia has its own format. The first step of processing the dump file is cleaning “*Wikipedia Tags*” such as 'gallery', 'timeline', 'noinclude', 'pre', 'table', 'tr', 'td', 'th', 'caption', 'form', 'input', 'select', 'option', 'textarea', 'ul', 'li', 'ol', 'dl', 'dt', 'dd', 'menu', 'dir', 'ref', 'references', 'img', 'imagemap', 'source' , 'TR' , 'TD' , 'TABLE' etc.
The cleaning is performed by a modified version of Wikipedia Extractor[5].

2- Data Selection

Wikipedia is a valuable source due to it's lexical diversity. The entries vary from medicine to astronomy, mathematics to politics, arts to sports etc. Therefore Wikipedia is a useful source for linguistic studies.
However, some entries are not suitable so these entries had to be omitted. The obtained “clean” text file forms the “***Raw_Data***” and “***Text_ID***” files. All the other files were generated by using these two files. Some entries were omitted as they are not suitable for linguistic processing. The omitted entries are:
a- Wikipedia entries are written by volunteer writers. Some of these writers do not obey Wikipedia format while writing.
b- Some entries are derived from other Wikipedia pages of different languages such as English, French, etc. These entries have been deleted[6].
c- Some Wikipedia entries are based on a template. This causes repetitions. Therefore these entries have been deleted.
d- Wikipedia entries with mathematical formulas have been deleted.

[1] http://dumps.wikimedia.org/trwiki/20131011/trwiki-20131011-pages-articles.xml.bz2
[2] Uncompressed file sizes. Total size is 3.6 Gb.
[3] Source text is in lowercase.
[4] Source text is in lowercase.
[5] Wikipedia Extractor is a python software by Giuseppe Attardi and Antonio Fuschetto from Pisa University. Software is licenced under GNU General Public License.
[6] A list of deleted entries with samples given under “What has Cleaned” title below.

3- The Data Set

3.1 TS Wikipedia Raw Data: This file includes “*cleaned*” form of Wiki-Dumps. File contains every Wikipedia entry tagged with an XML header including “text id”, “url” and “title”. A sample entry is given below:

```
<text id="53" url="http://tr.wikipedia.org/wiki?curid=53" title="Bilgisayar">
Bilgisayar Kendisine verdiğimiz bilgileri istediğimizde saklayabilen, istediğimizde geri verebilen cihaza denir.
İlk bilgisayar ENIAC'tır.
....
</text>
```

3.2 TS Wikipedia Text ID: This file includes Wikipedia entry id's matched to other files in the data set. Each line is as an XML tag that includes “text id”, “url” and “title”.

```
<text id="661" url="http://tr.wikipedia.org/wiki?curid=661" title="Basketbol">
<text id="662" url="http://tr.wikipedia.org/wiki?curid=662" title="Tarım">
```

3.3 TS Wikipedia Tokenized: This is the tokenized form of TS Wikipedia Raw Data. Tokenization is processed by a script modified to adapt Turkish[7].

```
<text id="10" url="http://tr.wikipedia.org/wiki?curid=10" title="Cengiz Han">
Cengiz
Han
Moğol
Börçigin
ailesinden
siyasetçi
.......
</text>
```

3.4 TS Wikipedia POS Tagged Corpus: This is the tagged[8] Turkish Wikipedia Corpus[9]. File consists of five tab-separated columns, each respectively represents, the token itself, POSTag, morphological analysis, possible lemma and the correct form of the word[10].

Word	PoSTag	Morpohological Parse	Lemma	Correct Form
Tarih	Noun	Noun+A3sg+Pnon+Nom	tarih	Tarih
boyunca	Postp	Postp+PCNom	boyunca	boyunca
ortama	Noun	Noun+A3sg+Pnon+Dat	ortam	ortama
ve	Conj	Conj	ve	ve
uygulamalara	Noun	Noun+A3pl+Pnon+Dat	uygulama	uygulamalara
göre	Postp	Postp+PCDat	göre	göre
değişik	Adj	Adj	değişik	değişik
egitim	YY	NoMorph	NoLemma	eğitim
tanımları	Noun	Noun+A3pl+P3sg+Nom	tanım	tanımları
yapılmıştır	Verb	Verb+Verb+Pass+Pos+Narr+A3sg+Cop+A3sg	yap	yapılmıştır

[7] An online interface for the tokenizer is available at http://gui.tscorpus.com/tokenizer/

[8] See Appendix for used tag set.

[9] An online interface for the PosTagger is available at http://gui.tscorpus.com/parser/

[10] If there is a spelling mistake that can be analyzed the correct form will be written in this column, else the word itself. If the input word is misspelled and analyzed succesfully it will be tagged by YY tag. The morphological analysis and lemma columns will be shown as “NoMorph” and “NoLemma”.

3.5 TS Wikipedia bi_gram bi-gram caluculations: Punctuations have been removed from the input data but diachronic characters (eg. â, î) are protected.

> **3.5.1 TS Wikipedia bi_gram Case Sensitive:** The file is in a three column tab-separated, case sensitive format. Each column respectively represents, the first word, the second word and the occurrence of bi-gram in the set.
>
> **3.5.2 TS Wikipedia bi_gram Case Insensitive:** The file is in a three column tab-separated, case insensitive format. Each column respectively represents, the first word, the second word and the occurrence of bi-gram in the set.

3.6 TS Wikipedia tri_gram: bi_gram calculations: Punctuations have removed from the input data but diachronic characters (eg. â, î) are protected.

> **3.6.1 TS Wikipedia tri_gram Case Sensitive:** The file is in a four column tab-separated, case sensitive format. Each column respectively represents, the first word, the second word, the third word and the occurrence of tri-gram in the set.
>
> **3.6.2 TS Wikipedia bi_gram Case Insensitive:** The file is in a four column tab-separated, case insensitive format. Each column respectively represents, the first word, the second word, the third word and the occurrence of tri-gram in the set.

4- What has been Cleaned?

As mentioned above, some entries of Wikipedia dump has been cleaned. The key term during this cleaning was "usefulness".

4.1 Data typed with non-Turkish Characters – ASCIIfied characters

Some entries include data typed by using non-Turkish characters. These entries have been deleted as much as it is possible. These entries are different from entries that include misspelled words.

> " Osmanli Imparatorlugu Bursa dolayarinda devletlesme..."

4.2 Mathematical Formulas

Entries including many mathematical formulas have been deleted.

> [x+y.z(abc)]

4.3 Unuseful Lists

The entries that have unuseful lists have been deleted, such as irregular English verbs

```
<BR>Rived </BR>
<BR> Ran </BR>
<BR> Sawed </BR>
<BR> Said </BR>
<BR> Saw </BR>
<BR>Sought </BR>
```

4.4 Entries copied from another language of the same entry.

There are noticeable number of entries in Turkish Wikipedia dump, copied from another language and partially translated or not translated into Turkish at all. These entries have been deleted.

4.5 Entries that has no content or entries that has no data left after automatic cleaning.

Some entries only has “template” of the entry or a list. For example the entry “List of cities in Okinawa” has been deleted as no data left after automatic cleaning. These entries have been deleted.

4.6 Entries about cities with a little information and almost identical.

Some entries, especially the entries based on a template has almost identical content. These entries have been deleted.

4.7 Entries about repetitive calendar dates.

The entries which do not contain any information but specific calendar dates have been deleted.

4.8 Other repetitive entries.

Turkish Wikipedia has a great number of entries about villages in Turkey. These entries have enormous number of structural errors by means of Wikipedia format. They are based on a template that causes repetitive usage of same schema as well. These entries have been deleted.

5.1 Tokenization

After Wiki dump is cleaned, the data tokenized by a script in order to prepare word per line files to use with tagger. The tokenizer script is based on “uft8-tokenize[11]”, by Serge Sharoff and Helmut Schmid. The script is modified in order to fit the needs and some special cases of Turkish[12].

5.2 Tagging

The part of speech tagging and morphological analysis is processed by “An averaged perceptron-based morphological disambiguator[13] ” by Sak et al[14]. The original software outputs every possible morphological analysis of the input word. Therefore, a new script is used in order to produce the desired result. Also the software has been enhanced in order to process seven new PosTags[15].

5.3 n-gram calculations

All n-gram calculations run by Ngram Statistics Package[16] (Text-NSP). Unfortunately, the original software was not capable of processing Turkish character set. Therefore, some basic modifications are applied to software[17] in order to support UTF-8.

[11] http://corpus.leeds.ac.uk/tools/

[12] An online interface for the tokenizer is available at http://gui.tscorpus.com/tokenizer/

[13] http://www.cmpe.boun.edu.tr/~hasim/

[14] Haşim Sak, Tunga Güngör, and Murat Saraçlar. Turkish Language Resources: Morphological Parser, Morphological Disambiguator and Web Corpus. In GoTAL 2008, volume 5221 of LNCS, 2008, pages 417-427. Springer.

[15] An online interface for the PosTagger is available at http://gui.tscorpus.com/parser/

[16] http://www.d.umn.edu/~tpederse/nsp.html

[17] An online demo will soon be available at http://gui.tscorpus.com/ngram/

TS Wikipedia Data Set PosTag List			
#	POSTag	TAG	Tag Used in Data Set
1	Verb	Verb	_Verb
2	Noun	Noun	_Noun
3	Adj	Adjective	_Adj
4	Adv	Adverb	_Adverb
5	Det	Determiner	_Det
6	Conj	Conjunction	_Conj
7	Postp	Postposition	_Postp
8	Interj	Interjection	_Interj
9	Pron	Pronoun	_Pron
10	Dup	Duplication	_Dup
11	Num	Number	_Num
12	Punc	Punctuation	_Punc
13	UnDef	Undefinite	_UnDef
14	Ques	Question	_Ques
15	YY	Misspell	_YY
16	Abbr	Abbreviation	_Abbr
17	intEmphasis	Internet Emphasis	_intEmphasis
18	intAbbrEng	Internet English Abbreviation	_intAbbrEnglish
19	tinglish	Tinglish	_tinglish
20	bor	Borrowed	_bor
21	intSlang	Internet Slang	_intSlang

The tags “YY, Abbr, intEmphasis, intAbbrEng, tinglish, bor and intSlang” are processed by enhanced PosTagger.