Building Kazakh Language Open Source Corpora using Wikipedia Resources

Turapbekov Bekzat[1]
[1] Suleyman Demirel University, bekzat.turapbekov@ce.sdu.edu.kz

Abstract. *The lack of free public accessible Kazakh corpus is one of the difficulties that Kazakh linguistics researches face. The aim of this paper is a step towards supporting Kazakh linguistics with the open source corpus built on Wikipedia dumps and one of its applications a Kazakh spellchecker.*

1. Introduction

Current research aims collection and development of resources to solve the need for a databank of Kazakh language words and to serve as training data for data-centric computing. It can be used as a data source for spell checkers, word processing documents, and in language detection problems and etc.

Kazakh language is one of Turkic languages, so it has inflectional and derivational agglutinative morphology, which causes that these languages have a lot of various word forms.

Since there is only paper of research provided by Nazarbayev University team, we decided to create open-source corpus of n-grams for kazakh language.

Corpuses are used as a data-source in statistical linguistics to detect unigram, bigrams, and n-grams. This data helps to analyze language structure and to find most used words and etc.

1.1 Previous work

Most popular corpuses nowadays are Brown corpus [1], Oxford [2] which were collected by people. New trend is gathering the corpus using Internet and Crowdsourcing.

The first attempts to build a large scale corpus of Kazakh language were made by Kazakh Language Corpus [3] and Almaty Corpus of Kazakh [4] projects.

2. Building Wikipedia Corpora

Wikipedia is a very valuable resource. It is online encyclopedia that is collaboratively edited by volunteers. The idea of building a Corpus using Wikipedia is not new. It was used in most of previous works, because of its availability and popularity. Entire Wikipedia is publicly available through XML dumps and edition in Kazakh languages exist [5].

A Wiki dump is a single large XML file containing all the articles of the Wikipedia. Generally the size of the dump is not very large. The compressed version of all articles in Kazakh language takes only 92.2 megabytes. After extracting the size of XML file becomes 1.1 gigabytes.

Despite its benefits, the information extracted from Wikipedia dumps cannot be easily used as a corpus, because they consists of heavy and complex wiki markup (Table 1.).

Wikipedia content is not primarily written in a standard XML-based markup language such as HTML, but in a specific markup language for wiki called wiki markup. (Table 2.)

```
[[Санат:Химия]]
[[Санат:Физика]]</text>
      <sha1>nd1xsx9xrvtf0v60pyl0vgqno4n3boz</sha1>
    </revision>
  </page>
  <page>
    <title>Шұңқыркөл (Ақмола облысы)</title>
    <ns>0</ns>
    <id>42584</id>
    <revision>
      <id>1884857</id>
      <parentid>1814276</parentid>
      <timestamp>2013-03-24T07:03:06Z</timestamp>
      <contributor>
        <username>Sibom</username>
        <id>4616</id>
      </contributor>
      <comment>/* Сілтемелер */</comment>
      <model>wikitext</model>
      <format>text/x-wiki</format>
      <text xml:space="preserve">{{Елді мекен-Қазақстан
|статусы                = Ауыл
|атауы                  =Шұңқыркөл
|сурет                  =
|әкімшілік күйі         =
 |lat_deg = 51 |lat_min = 22 |lat_sec = 6.6
 |lon_deg = 68 |lon_min = 14 |lon_sec = 30.85
```

Table 1. Wikipedia dump structure

Wiki markup	Function
`== heading level 2 ==`	heading
`----`	horizontal rule
`:indentation level 1`	indentation
`* item`	unordered list
`# item`	ordered list
`: definition 1`	definition list
`''italic text''`	italic
`'''bold text'''`	bold
`'''''bold italic text'''''`	bold italics
`<small>small text</small>`	small font-size
`O<sub>2</sub>`	subscripts
`[[target page name|link label]]`	internal link to another wiki page
`[[http://www.wikipedia.org Wikipedia]]`	external link
`[[File:Image.png]]`	image

Table 1. Wiki markup

In order to get the desired plain text from wikitext we used a modified version of extractor from Evan Jones [6]. Wiki text frequently contains incorrectly formed wiki markup such as a missing of a closing tag for a table or wrong line breaks, and these elements may cause a problem. Fortunately, for Kazakh language uses Cyrillic symbols, and this problem was be solved just by removing all characters except Kazakh letters and selected delimiters

That frequency table show that most repeated words are connected with Geography field.

After extraction of plain text corpus, the python library nltk allowed generating bigrams and trigrams. [5]

3 Applications: Corpus for spell checking

An important application of corpus is spell checking. To give a first impression of the use and effectiveness of corpus, we made some preliminary experiments with a spellchecker algorithm based on example of trie taken from Steve Hanov [7]. The example code can be found in repository of project.

4 Future work

We would like to extend corpus with adding more resources to make it competitive with commercial versions. The logical extension of this goal is through obtaining textual data from the web Most popular strategy which was used on building web articles based corpora is to use the Wikipedia corpus as a seeds for bootstrapping the a lot more information from World Wide Web[10] .

5 Conclusion

We have built corpus for Kazakh language, a total of 20 million words were collected. With almost 600 thousand words with different derivations. We believe that availability of this data will be helpful for other people interested in further research.

References

1. https://en.wikipedia.org/wiki/Brown_Corpus
2. http://www.oxforddictionaries.com/words/the-oxford-english-corpus
3. Assembling the Kazakh Language CorpusMakhambetov, Olzhas et al. "Assembling the Kazakh Language Corpus." EMNLP Oct. 2013: 1022-1031.
4. http://web-corpora.net/KazakhCorpus/search/?interface_language=en
5. https://dumps.wikimedia.org/kkwiki/
6. http://www.evanjones.ca/software/wikipedia2text.html
7. http://stevehanov.ca/blog/index.php?id=114
8. https://github.com/bekzattt/Open-Source-Kazakh-Corpus
10. ApplicatGhani, Rayid, Rosie Jones, and Dunja Mladenic. "Building minority language corpora by learning to generate web search queries." Knowledge and Information Systems 7.1 (2005): 56-83.ions: Corpus for spell checking