IRSTI 20.19.27

D. Chapaev[1], B. Turapbekov[1]
[1]Suleyman Demirel University
Kaskelen, Kazakhstan

BUILDING KAZAKH LANGUAGE OPEN SOURCE CORPORA USING WIKIPEDIA RESOURCES

Abstract. The lack of free public accessible Kazakh language corpus is one of the difficulties that Kazakh linguistics researchers face. Corpuses are used as a data source in statistical linguistics for the detection of unigrams, bigrams and n-grams. These data help analyze the structure of the language and find the most used words, etc. 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 spell checker. Now, corpus contains over 21 million words. It is also open source and waiting for any contributors and suggestions.

Key words: N-gram, corpus, spell checker, tokenizer.

***

Аңдатпа. Қазақ тілі корпусына қоғамдық қолжетімділіктің жетіспеушілігі - қазақ лингвистерінің қиындықтарының бірі. Корпустар негізінде униграмм, биграмм және n-граммдарды табу үшін статистикалық лингвистикада деректердің қайнар көзі ретінде пайдаланылады. Бұл деректер тіл құрылымын талдауға және ең жиі қолданылатын сөздерді табуға көмектеседі. Осы мақаланың басты мақсаты - қазақ тіл білімін Википедия деректеріне негізделген ашық бастапқы кодымен қамтамасыз етудегі қадам, ал оның бір қосымшасы ретінде қазақ тілінің орфографиясын тексеру. Қазіргі уақытта корпуста 21 миллионнан астам сөздер бар. Ол сондай-ақ бәріне де ашық, кез-келген серіктестер мен ұсыныстарды күтуде.

Кілт сөздер: N-грамм, корпус, орфографияны тексеру, токенизатор.

***

Аннотация. Отсутствие свободного общественного доступа к казахскому языковому корпусу является одной из трудностей, с которыми сталкиваются казахстанские лингвисты. Корпусы используются в качестве источника данных в статистической лингвистике для обнаружения униграмм, биграмм и n-граммов. Эти данные помогают проанализировать структуру языка и найти наиболее часто используемые слова и т.д. Цель данной статьи – шаг к поддержке казахской лингвистики с открытым исходным корпусом, построенным на данных Википедии, и одним из его приложений является проверка орфографии казахского языка. На данный момент корпус содержит более 21 миллиона слов. Он также открыт для всех и ждет любых соучастников и предложений.

Ключевые слова: N-грамм, корпус, проверка орфографии, токенизатор.

Introduction

Current research aims collection and development of resources to solve the need for a data bank 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. It is the official state language of Kazakhstan and the mother tongue for more than 10 million people around the world. But before the 90s of the last century because of historical reasons, the Russian language was used in colloquial and written speech. This fact, in turn, caused the problem of small amounts of work in the Kazakh language in such areas as entertainment, science, official documentation, etc.

Related Work

The Brown corpus [1] was the first text corpus of American English. The original corpus was published in 1963-1964 by W. Nelson Francis and Henry Kučera at Department of Linguistics, Brown University Providence, Rhode Island, USA. The corpus consists of 1 million words (500 samples of 2000+ words each) of running text of edited English prose printed in the United States during the year 1961 and it was revised and amplified in 1979.

The British National Corpus (BNC) [2] is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of British English from the later part of the 20th century, both spoken and written. BNC is used as a model for many modern corpora.

The Russian National Corpus (RNC) [3] is over 300 million words. The corpus of Russian is a reference system based on a collection of Russian texts in electronic form. The Russian National Corpus covers primarily the period from the middle of the 18th to the early 21st centuries. This period represents the Russian language of both the past and the present in a wide range of sociolinguistic variants: literary, colloquial, vernacular, in part dialectal. The Corpus includes original (non-translated) works of fiction (prose, drama and poetry) of cultural importance which are interesting from a linguistic point of view. Apart from fiction, the Corpus includes a large volume of other sources of written (and, for the later period, spoken) language: memoirs, essays, journalistic works, scientific and popular scientific literature, public speeches, letters, diaries, documents, etc.

Unfortunately, to date, very little work has been done to create a publicly available Kazakh language corpus, but we found some good corpora. The first attempts to build a large-scale corpus of Kazakh language were made by Kazakh Language Corpus [4] and Almaty Corpus of Kazakh [5] projects.

Kazakh Language Corpus (KLC) [4] which is provided by Nazarbayev University research team containing over 135 million words and more than 400 000 documents classified into five major genres (domains): literary, publicistic, official, scientific and informal. Second, Almaty Corpus of Kazakh language (NCKL) [5] which is created by efforts of the department of general linguistics and foreign philology of faculty of philology, literary study and world languages of Al-Farabi Kazakh National University under the leadership of the chief of the department G.B. Madiyeva with the participation of the staff of the faculty of philology of Higher School of Economics National Research University (Moscow). At the moment the size of the corpus is more than 40 million word tokens. Finally, The Kazakh Web Corpus [6] created by Sketch Engine. This corpus is collected from the web using SpiderLing crawler. The corpus was prepared according to standards described in the document A Corpus Factory for Many Languages (Kilgarriff et al. at LREC 2010). Data were downloaded in January 2012 with the total size 139 million words. Texts were cleaned and deduplicated.

All these corpora are probably good, but there is one minus, we can not use them in our studies since they are private. Because of the fact that we do not have a widely available open source corpus, many studies are being ended at this stage. We want to improve it with our small contribution to Kazakh linguistics by creating this corpus.

Corpus Purpose

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. The main goal of the corpus is to facilitate academic studies of the vocabulary and grammar of the language, as well as subtle, but constant processes of changing the language in a relatively short

period of time: from one to two centuries. Another purpose of the corpus is to serve as a reference point for lexical, grammatical and accentological questions and the history of language. Modern IT-technologies greatly simplify and accelerate the processing of large amounts of text, which creates an opportunity for mass statistical analysis of texts. As a result, the study of the language now yields results that could only be guessed earlier. At present, truly scientific descriptions of grammars and academic dictionaries should be based on the corpus of their respective languages. The use of the body data is desirable (if not always strictly necessary) in other, more specialized language studies.

Therefore, the main users of national corpora are linguists of various profiles. Nevertheless, the corpus is also useful for non-linguists. Reliable statistical information on the use of language in a certain period or a certain author may be of interest to researchers of literature, history and other humanitarian subjects. National corporations are also useful for language teachers, both local and foreign; language textbooks and curricula are increasingly oriented toward the corpus. The case can be used to find out the options for using unknown words by foreigners, students, teachers, journalists, writers. Thus, the corpus is designed for people who are interested in the structure and use of the language, whether it is their professional interest or not.

Building Wikipedia Corpora

Most popular Brown [1] and Oxford [7] corpuses are collected by people. A new trend is gathering the corpus from the web using spiders and crawlers.

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 and SQL dumps and edition in Kazakh languages exist [8].

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 about 100 megabytes. After extracting the size of XML file becomes 1.2 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.). In order to get the desired plain text from wikitext we used a WikiExtractor [9] and wiki2text [10] libraries. Wikipedia contains about 225647 articles in Kazakh language. 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.

Table 1

Wikipedia dump structure

```
[[Санат:Химия]]
[[Санат:Физика]]</text>
      <sha1>nd1xsx9xrvtf0v60pyl0vgqno4n3boz</sha1>
    </revision>
  </page>
  <page>
    <title>Шұқыркөл (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                  =Шұқыркөл
|сурет                  =
|әкімшілік күйі         =
 |lat_deg = 51 |lat_min = 22 |lat_sec = 6.6
 |lon_deg = 68 |lon_min = 14 |lon_sec = 30.85
```

That frequency table show that most repeated words are connected with Geography field. After extraction of plain text corpus, the python nltk [11] library allowed generating bigrams and trigrams [12]. Totally our corpus containing over 21 million words and 1,3 unique words.

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 spell checker algorithm based on example of taken from Peter Norvig [13]. The example code can be found in repository of project [12].

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[14].

Table : Wiki markup

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

Conclusion

Totally we have build corpus for Kazakh language, which contains over 21 million words and 1,3 unique words. 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 Francis, W., Kucera, H. Brown Corpus Manual. [online] Clu.uni.no. Available at: http://clu.uni.no/icame/manuals/BROWN/INDEX.HTM [Accessed 21 May 2018].

2 Burnard, L. *[bnc]* About the British National Corpus. [online] Natcorp.ox.ac.uk. Available at: http://www.natcorp.ox.ac.uk/corpus/index.xml [Accessed 21 May 2018].

3 Ruscorpora.ru. Russian National Corpus. [online] Available at: http://ruscorpora.ru/en/index.html [Accessed 21 May 2018].

4 Makhambetov, O., Makazhanov, A., Yessenbayev, Zh., Matkarimov, B., Sabyrgaliyev, I., Sharafudinov, A. Assembling the Kazakh Language Corpus // In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, Washington, USA, October. Association for Computational Linguistics. – P. 1022–1031

5 Madieva, G., Daniel, M., Umatova, Z. (n.d.). *Almaty Corpus of Kazakh*. [online] Available at: http://web-

corpora.net/KazakhCorpus/search/?interface_language=ru [Accessed 21 May 2018].

6 Sketchengine.eu. (n.d.). kkWaC – Kazakh corpus from the web | Sketch Engine. [online] Available at: https://www.sketchengine.eu/kkwac-kazakh-corpus/ [Accessed 21 May 2018].

7 Oxford Dictionaries | English. (n.d.). The Oxford English Corpus | Oxford Dictionaries. [online] Available at: http://www.oxforddictionaries.com/words/the-oxford-english-corpus [Accessed 21 May 2018].

8 Dumps.wikimedia.org. (n.d.). Index of /kkwiki/. [online] Available at: https://dumps.wikimedia.org/kkwiki/ [Accessed 21 May 2018].

9 Attardi, G. attardi/wikiextractor. [online] GitHub. Available at: https://github.com/attardi/wikiextractor [Accessed 21 May 2018].

10 Speer, R. rspeer/wiki2text. [online] GitHub. Available at: https://github.com/rspeer/wiki2text [Accessed 21 May 2018].

11 Nltk.org. (n.d.). Natural Language Toolkit — NLTK 3.3 documentation. [online] Available at: https://www.nltk.org/ [Accessed 21 May 2018].

12 Chapayev, D. chapayevdauren/kazakh-language-corpus. [online] GitHub. Available at: https://github.com/chapayevdauren/kazakh-language-corpus [Accessed 21 May 2018].

13 Norvig, P. How to Write a Spelling Corrector. [online] Norvig.com. Available at: http://norvig.com/spell-correct.html [Accessed 21 May 2018].

14 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

15 Altenbek, G., Wang Xiao-long. Kazakh segmentation system of inflectional affixes. // In Joint Conference on Chinese Language Processing, Cips-Sighan. – pages 183-190.