TurkLang.2018

VI International Conference on Computer Processing of Turkic Languages

Tashkent, Uzbekistan
18-20 October, 2018

ТАШКЕНТСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
УЗБЕКСКОГО ЯЗЫКА И ЛИТЕРАТУРЫ ИМЕНИ АЛИШЕРА НАВОИ

АКАДЕМИЯ НАУК РЕСПУБЛИКИ ТАТАРСТАН

ИНСТИТУТ ПРИКЛАДНОЙ СЕМИОТИКИ

ЕВРАЗИЙСКИЙ НАЦИОНАЛЬНЫЙ УНИВЕРСИТЕТ
ИМЕНИ Л. Н. ГУМИЛЁВА

МИНИСТЕРСТВА ОБРАЗОВАНИЯ И НАУКИ
РЕСПУБЛИКИ КАЗАХСТАН

НИИ «ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ»

VI МЕЖДУНАРОДНАЯ КОНФЕРЕНЦИЯ ПО КОМПЬЮТЕРНОЙ ОБРАБОТКЕ ТЮРКСКИХ ЯЗЫКОВ

«TURKLANG-2018»

(труды конференции)

ТАШКЕНТ
ИЗДАТЕЛЬСКО-ПОЛИГРАФИЧЕСКИЙ
ДОМ «NAVOIY UNIVERSITETI»
2018

С 23
УДК 811.512.1 (063)
ББК 81.2 ТЮРК (я43)

Шестая Международная конференция по компьютерной обработке тюркских языков «TurkLang-2018». (Труды конференции) –Ташкент: Издательско-полиграфический дом «NAVOIY UNIVERSITETI», 2018. – 390 с.

Научные редакторы: **PhD Н.З. Абдурахмонова;**
к.т.н. А.Р. Гатиатуллин

Сборник содержит материалы Шестой Международной конференции по компьютерной обработке тюркских языков «TurkLang-2018» (Ташкент, Узбекистан, 18–20 октября 2018 г.)

Данная публикация предназначена для научных работников, преподавателей, аспирантов и студентов, специализирующихся в области компьютерной лингвистики и ее приложений.

(Издается в авторской редакции)

ISBN 978-9943-5635-1-3

Издание рекомендовано к публикации Постановлением №8 Ученого совета Ташкентского государственного университета узбекского языка и литературы имени Алишера Навои от 28 декабря 2018 года.

FORMATION OF THE SYNTHETIC CORPORA FOR KAZAKH ON THE BASE OF ENDINGS COMPLETE SYSTEM

A. Karibayeva, B.Abduali, D. Amirova, Al-Farabi Kazakh National University,
Institute of Information and Computational Technologies, Almaty, Kazakhstan,
a.s.karibayeva@gmail.com , balzhanabdualy@gmail.com, amirovatdina@gmail.com

The problem of absence of parallel corpora are actual for a large number of language pairs and can severely detriment the quality of neural machine translation systems. The lack of parallel corpora for Kazakh is actual in machine translation system. The creation and collection of corpus limits the creation of a neural machine translation with a good quality of translation. Since, the mentioned kind of machine translation needs large data for system training. For this reason was created synthetic corpora to extend the number of sentences to training Kazakh neural machine translation system(NMT). As, synthetic corpora is mentioned the sentences that automatically created from program that generated construction by part of speech and their description of changing by person, case and in number. The method is a language-dependent to enable machine translation between a low-resource language and a high-resource language, e.g. English and Russian. Kazakh language has 8 types of changing by person, 2 types of changing in number, 7 types of connection dependency and 6 cases. 'Мен [Men] (I)', 'Сен [sen] (you)', 'Сіз [Siz] (you)', 'Ол [ol] (he)', 'Біз [biz] (we)', 'Сендер [sender] (you)', 'Сіздер [sizder] (you)', 'Олар [olar] (they)' — the types of person. The singular and plural is type of number. The connection dependency based on the possessive form of nouns. Belonging in the Kazakh language is expressed with the help of the endings of belonging — 'тәуелдік жалғау'. Such a construction "noun" + "end of belonging" is also called the possessive form of nouns. The word in the Kazakh language is based on adding an ending to the stem. Taking into account the number of types of change, a complete system of word endings was created, which consists of 3550 possible combinations of endings structures. Synthetic corpora is created from the longest construction of the offer to the shortest. The novelty of approach in generation synthetic corpora by using sentence structure pattern and complete set of endings. In this paper will be shown the creating process of synthetic corpora of Kazakh language by sentence construction. The results are shown in number of created sentences for Kazakh-English, Kazakh-Russian language pairs.

Key words: *synthetic corpora; parallel corpora; neural machine translation system; set of endings, Kazakh language.*

ФОРМИРОВАНИЕ СИНТЕТИЧЕСКОЙ КОРПОРА ДЛЯ КАЗАХА НА ОСНОВЕ ЗАВЕРШЕНИЯ СИСТЕМЫ

А. Карибаева, Б. Абдуали, Д. Амирова,
Казахский национальный университет им. Аль-Фараби, факультет информационных систем, Пр-т Аль-Фараби, 71, 050040, Алматы, Казахстан;
Институт Информационных и Вычислительных Технологий, ул. Пушкина, 125, 050010, Алматы, Казахстан, a.s.karibayeva@gmail.com , balzhanabdualy@gmail.com, amirovatdina@gmail.com

Проблема отсутствия параллельных корпусов актуальна для большого числа языковых пар и может серьезно ухудшить качество систем нейронного машинного перевода. Отсутствие параллельных корпусов для казахского языка актуально в системе машинного перевода. Создание и сбор корпусов ограничивают создание нейронного машинного перевода с хорошим качеством перевода. Поскольку упомянутый вид машинного перевода требует больших данных для обучения системы. По этой причине были созданы синтетические корпуса для расширения количества предложений для обучения казахской системе нейронного машинного перевода (НМП). В качестве синтетических корпусов упоминаются предложения, которые автоматически создаются из программы, которая генерировала конструкцию по части речи, и их описанию, изменяющееся по лицу, падежу и количеству.

Способ зависит от языка, чтобы обеспечить возможность машинного перевода между языком с низким уровнем ресурсов и языком с высоким ресурсом, например английский и русский. Казахский язык имеет 8 типов изменения по лицу, 2 типа изменения по количеству, 7 типов притяжания и 6 падежей. «Мен (Я)», «Сен (Ты)», «Сіз (вы)», «Ол (он, она)», «Біз (мы)», «Сендер (вы)», «Сіздер (вы)», «Олар (они)» — типы по лицу. Единственное и множественное число является типом числа. Зависимость соединения основана на притяжательной форме существительных. Принадлежность в казахском языке выражается с помощью окончаний принадлежности — ‘тәуелдік жалғау’. Такая конструкция «существительное» + «конец принадлежности» также называется притяжательной формой существительных. Слово в казахском языке образуется при добавлении окончания к основанию. С учетом количества типов изменений была создана полная система окончаний слов, состоящая из 3550 возможных комбинаций структур

окончаний. Синтетические корпуса создаются от самой длинной конструкции предложения до самой короткой. Новизна подхода в генерации синтетических корпусов с использованием шаблона структуры предложений и полного набора окончаний. В данной статье будет показан процесс создания синтетических корпусов казахского языка по конструкции предложения. Результаты показаны в количестве созданных предложений для казахско-английских, казахско-русских языковых пар. Этот метод зависит от языка, обеспечивающий машинный перевод между языком низкого ресурса и языком высокого ресурса, например, английский и русский.

***Ключевые слова**: синтетические корпуса; параллельные корпуса; система нейронного машинного перевода; система окончаний, казахский язык.*

Introduction

The creating the neural machine translation system required a big number of data. For low-resources languages, like a Kazakh needs to qualitative parallel corpora.

Kazakh language is agglutinative language with rich morphology with various combinations of suffixes. This language doesn’t have enough resources like linguistic resources and parallel corpora. So, Kazakh Language is low-resource language. Lack of data is the main problem for creating a neural machine translation with high quality for the Kazakh language. For having good translation with NMT it should to train significant number of data.

For that reason we present method to create synthetic corpora for Kazakh language on the base of complete set of endings[1]. Each part of speech has its own characteristics and its kinds of endings, which it can have. There are about 3550 combinations of endings. Based on this complete set of endings, tables were created for all parts of the speech of the Kazakh language.This paper is structured as follows. Related works are described in section 2. In section 3 we present method of generating synthetic corpora. Results are discussed in section 4. Finally, conclusion is given in section 5.

Related works

The most of work were considered with researchers. The absences of parallel data were inspired researchers to creating and investigation the low-resources languages. The Kazakh language related to low resources languages too. Under synthetic corpora the most considers the automatically-generated corpora, translated texts from different translation systems, and etc. Anna Currey and et al. used monolingual data with mixing main corpora in target language, namely to Romanian and Turkish languages to train the NMT system for low-resource. This method improved the BLEU to 1.2 for latter languages[2].

One of the methods of generating synthetic parallel corpora is using back-translation. That means NMT system is trained in the reverse translation direction

(target-to-source), and is then used to translate target-side monolingual data back into the source language (in the backward direction, hence the name back translation)[3]. The received sentences can be added to the existing training data and increase a volume of synthetic parallel corpus. In [3] authors for training NMT systems use iterative back-translation for generating synthetic parallel corpora from monolingual data. They used method to both high (German-English) and low (English-French, English-Farsi) resourced scenarios.

In [4] presented dual learning method on English-French language pairs.They develop a dual-learning mechanism, which can enable an NMT system to automatically learn from unlabeled data through a dual-learning game[4].

Generation process of synthetic corpora for Kazakh language

The process of generation depend on language direction. The proposed method of synthetic generation based on part of speech and Kazakh endings. As all Turkic languages Kazakh is agglutinative language. The word forms constituted from adding suffixes to the base of word.

The complete set of endings used for create synthetic corpora. Based on complete set of Kazakh endings was created morphological language. It consists about 3550 various combinations of endings.

The structure of sentences changed by person, case and etc. For example one of the part of speech presented in table 1.

Table 1.

The tense of Kazakh language, structural forms, examples and their communication with English language.

The tense of Kazakh language and transliteration	Grammar structure for Kazakh	Example for Kazakh and transliteration	English names of the tense	Grammar structure for English	English translation
Нақосы шақ (Nakosy shak)	V+A(PresSm)+(Sg,Pl)+(P1, P2,P2в,P3)	Мен істеп жүрмін (Men istep zhurmin)	Present Simple	V	I work

Нақ осы шақтың күрделі түрі (Nakosy shaktyng kurdeli turi)	V+A(PresComp(PresSm))+(Sg,Pl)+(P1,P2,P2в,P3)	Мен істеп жатырмын (Men istep zhatyrmyn)	Present Continuous	to be + V + ing	I am working
Ауыспалы осы шақ (Auyspaly osy shak)	V+A(PresNow)+(Sg,Pl)+(P1,P2,P2в,P3)	Мен істедім (Men istedim)	Present Perfect	to be + V + ed	I have worked
Жедел өткен шақ (Zhedel otken shak)	V+A(PastOper)+(Sg,Pl)+(P1,P2,P2в,P3)	Мен істедім (Men istedim)	Past Simple	V + ed	I worked
Бұрынғы өткен шақ (burynggy otken shak)	V+A(PastOld)+(Sg,Pl)+(P1,P2,P2в,P3)	Мен істегенмін (Men istegenmin)	Past Continuous	to be + V + ing	I was working

Ауыс палы өткен шақ (Auys paly otken shak)	V+A(PastMay)+(Sg,Pl)+(P 1,P2,P2в,P3)	Мен істеп отырд ым (Men istep otyrdy m)	Past Perfe ct	to be + V + ed	I had wor ked
Болжа лды келер шақ (Bolzh aldy keler shak)	V+A(FutCast)+(Sg,Pl)+(P1, P2,P2в,P3)	Мен істей мін (Men isteimi n)	Futur e Simp le	to be + V	I shall wor k
Мақса тты келер шақ (Maks atty keler Shak)	V+A(FutObj)+(Sg,Pl)+(P1, P2,P2в,P3)	Мен істеп отырм ын (Men istep otyrm yn)	Futur e Cont inuo us	to be + be + V + ing	I shall be wor king
Ауыс палы келер шақ (Auys paly keler shak)	V+A(FutSub)+(Sg,Pl)+(P1, P2,P2в,P3)	Мен істей мін (Men isteimi n)	Futur e Perfe ct	to be + hav e + V+ ed	I shall have wor ked

Similarly, we fill out the table of all parts of speech and get many options for ending. Then with helping this complete set of endings created files, and prepared sentence structure. Each part of speech are in different files and connect to the software part. For example one structure of sentences «Сіз мектепке бүгін ерте келдіңіз», for this sentence created 6 files:

- pronoun (мен, сен, сіз, ол, біз, сендер, сіздер, олар);
- nouns (мектепке, жұмысқа, сабаққа, бақшаға, үйге, паркке);
- adverb1 (бүгін, кеше, арғыкүні, таңертең);
- adverb2 (ерте, кеш, асығып, жүгіріп, баяу);

- verb (кел, келме);
- endings (дім, дің, діңіз, ді, дік, діңдер, діңіздер, ді).

And similarly, create exactly the same files in English, but the sixth files must be empty, because in English does not has a suffixes. There is first and fifth files does not has many types of variants, but other nouns, adverbs can be filled more, it is help to create lots of options of sentences.

Then through the automatic generated was get following structure of sentences with changing context of words in the following table 2:

Table 2.

Automatic generated sentences.

Sentences in Kazakh	Sentences in English
Сіз мектепке бүгін ерте келдіңіз	You come to school early today
Сіз мектепке бүгін ерте келмедіңіз	You did not come to school early today
Сіз мектепке бүгін келдіңіз.	You come to school today
Сіз мектепке бүгін келмедіңіз.	You did not come to school today
Сіз мектепке ерте келдіңіз.	You come to school early
Сіз мектепке ерте келмедіңіз.	You did not come to school early
Сіз бүгін ерте келдіңіз.	You come early today
Сіз бүгін ерте келмедіңіз.	You did not come early today
Сіз мектепке келдіңіз.	You come to school
Сіз мектепке келмедіңіз.	You did not come to school
Сіз ерте келдіңіз	You come early
Сіз ерте келмедіңіз.	You did not come early
Сіз бүгін келдіңіз.	You come today
Сіз бүгін келмедіңіз.	You did not come today
Сіз келдіңіз.	You come
Сіз келмедіңіз.	You did not come

It was for one case, and exactly the same for other structural proposals, the files are created and by using automatic generation, was created the parallel synthetic corpora.

Results

The automatic generation of the sentence helps to increase the volume of corpora, and in the future, use a variety of options and structured proposal to get more sentences. Thus, the volume of the corpora increases.

The results of translation are shown in the next table 3 below.

Table 3.

Automatic generated synthetic corpora.

Corpora	Number of generated sentences
Kazakh-English	600K
Kazakh-Russian	700K

Conclusion

In this paper was considered the generation synthetic corpora of Kazakh language by using complete set endings and logical structure of sentences. The reason of using it is understand with absences of resources. By using this method we will try to extend the number of parallel corpora.

Acknowledgements

This research performed and financed by the grant Project IRN AP05132950 "Development of an information-analytical search system of data in the Kazakh language", awarded to The Republican State Enterprise (RGP) on the right of economic management (PVC) «Institute of Information and Computational Technologies».

REFERENCES:

1. Tukeyev, U., Automaton models of the morphology analysis and the completeness of the endings of the kazakh language. Proceedings of the international conference «Turkic languages processing» TURKLANG-2015 September 17–19, Kazan, Tatarstan, Russia, 2015. 91-100 pp.

2. Anna Currey, Antonio Valerio Miceli Barone, and Kenneth Heafield. Copied monolingual data improves low-resource neural machine translation. In Proceedings of the Second Conference on Machine Translation, pp. 148–156, Copenhagen, Denmark, September 2017. Association for Computational Linguistics..

3. Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, Trevor Cohn.Iterative Back-Translation for Neural Machine Translation. Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, pages 18–24. Melbourne, Australia, July 20, 2018. Association for Computational Linguistics.

4. Di He, Yingce Xia, Tao Qin, LiweiWang, Nenghai Yu, Tieyan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In D. D. Lee, M.

Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, Advances in Neural Information Processing Systems 29, pages 820–828..

КЛАССИФИКАЦИЯ СЕМАНТИЧЕСКИХ РОЛЕЙ В СЕМАНТИЧЕСКОЙ РАЗМЕТКЕ ЭЛЕКТРОННОГО КОРПУСА ТЕКСТОВ ТУВИНСКОГО ЯЗЫКА[4]

А.Б. Хертек, В.С. Ондар, Тувинский государственный университет,
г. Кызыл, Россия, khertek.ab@yandex.ru

В статья содержит описание классификации семантических ролей для семантической разметки Электронного корпуса текстов тувинского языка. Для составления инвентаря семантических ролей были использованы понятия классов семантических функций: адъекты, актанты и сирконстанты. Классы актантов и сирконстантов представлены подклассами, каждый из которых включает по несколько конкретных семантических ролей.

Ключевые слова: *семантические роли, актанты, сирконстанты, адъекты, семантическая разметка.*

THE CLASSIFICATION OF SEMANTIC ROLES IN THE SEMANTIC MARKUP OF THE ELECTRONIC TEXT CORPUS OF TUVAN LANGUAGE

A.B. Khertek, V.S. Ondar, Tuvan State University,
Kyzyl, Russia, khertek.ab@yandex.ru

In this article provides a description of the classification of semantic roles to semantic markup of the Electronic text corpus of Tuvan language. To compile the inventory of semantic roles has been used the concept of classes of semantic features: adjecti, actants and circostanti. Classes of actants and circonstances represented by subclasses, each of which includes several specific semantic roles.

Key words: *semantic role, actant, circonstance, objects, semantic markup.*

[4] Работа выполнена в рамках Госзадания Министерства науки и высшего образования РФ № 34.3876.2017/4.6

Ilmiy-ommabop nashr

Muharrir: Ulug‘ BEK
Texnik muharrir: Bobur Hamroyev
Sahihalovchi dizayner: Ulug‘bek Urunov
Musahhih: Sevinch Ahmedova

Litsenziya raqami: AI 310. 2017-yil 24-noyabr sanasida berilgan.
Bosishga 2018-yil 30-dekabr sanasida ruxsat etildi.
Bichimi: 60 x 84 1/8; Shartli bosma tabog‘i: 38,00.
Nashriyot hisob tabog‘i: 38,37. 01-sonli buyurtma.

ISBN 978-9943-5635-1-3

Original maket «NAVOIY UNIVERSITETI» nashriyot-matbaa
uyida tayyorlandi va www.Turklang.uz saytiga pdf shaklida joylashtirildi.

Nashriyot manzili: Toshkent shahri, Yusuf Xos Hojib ko‘chasi, 103-uy.
Tel.: +998 (94) 639-0344; (97) 344-0241; (90) 909-5401.
Web: navoiy-uni.uz E-mail: navoiyuniversiteti@mail.ru

₪ 388 ₪

С 23
УДК 811.512.1 (063)
ББК 81.2 ТЮРК (я43)

Шестая Международная конференция по компьютерной обработке тюркских языков «TurkLang 2018». (Труды конференции) –Ташкент: Издательско-полиграфический дом «NAVOIY UNIVERSITETI», 2018. – 390 с.

Научно-популярное издание

Редактор: Улуг БЕК
Технический редактор: Бобур Хамраев
Дизайнер-верстальщик: Улугбек Урунов
Корректор: Севинч Ахмедова

Регистрация лицензии: AI 310. Выдана 24 ноября 2017 года.

Разрешено в печать 30.12.2018.
Формат: 60 x 84 1/8; Усл. печ. лист: 38,00.
Издат. лист: 38,37. Заказ №01.

ISBN 978-9943-5635-1-3

Оригинал макет изготовлен в издательско-полиграфическом доме «NAVOIY UNIVERSITETI» и размещен на сайте www.turklang.uz

Адрес издательства: г. Ташкент, ул. Юсуф Хос Хожиб, дом 103.
Tel.: +998 (94) 639-0344; (97) 344-0241; (90) 909-5401.
Web: navoiy-uni.uz E-mail: navoiyuniversiteti@mail.ru

Отдел маркетинга ИПД ГУП «NAVOIY UNIVERSITETI»:
+998 (97) 701–5401.

₪ 389 ₪