UDC 004.588
IRSTI 20.00.00

TRANSLATES INTO TURKISH LANGUAGES CREATION OF A MODEL OF NEURON MACHINE TRANSLATION

Translates into turkish languages creation of a model of neuron machine translation

Z. Omiotek[1], D.R. Rakhimova[2], A.Zh. Zhunussova[3]
[1]Lublin University of Technology, Lublin, Poland
[2,3]Al-Farabi Kazakh National University, Almaty, Kazakhstan

[1]ORCID ID: https://orcid.org/0000-0002-6614-7799, z.omiotek@pollub.pl,
[2]ORCID ID: https://orcid.org/0000-0003-1427-198X, di.diva@mail.ru
[3]ORCID ID: https://orcid.org/0000-0002-4350-6322, alia_94-22@mail.ru

Abstract. In our work we see that the quality of translation has improved due to the creation of models of translation from Turkish to English and from English to Turkic languages. Turkic-speaking languages are structurally similar. Therefore, studying one of the Turkic languages, you can assemble a corpus for other languages and apply it to the same model. This is done using the OpenNMT model (open neural machine translation). The article shows the morphological, lexical, semantic increase of BlEU (translation index) words and sentences of Turkic languages using OpenNMT. To increase the value of BLEU it is necessary to increase the base in the case. In addition, the work provides a detailed description of the construction of OpenNMT models. Experiments with the Kazakh language, one of the Turkic languages, were conducted and the results were obtained. Words in the Kazakh language taken from the news. The scientific work includes a review of the work of scientists who studied neural machine translation. It is shown that the results of this work outperform the work of other researchers. Having created the neural model of OpenNMT, you will see that the result of the translation is not the same as the online translation from Google, Yandex. OpenNMT also takes less time to read data and saves memory. The results of the experimental study show that the Kazakh-English and English-Kazakh language pairs gave good results in translation.

Keywords: OpenNMT, Turkic languages, neural machine translation.

Introduction

Every year we can see that the quality of machine translation in all languages is improving [1]. These languages include the group of Turkic-speaking languages. Languages that are part of the Turkic-speaking languages: Turkish, Turkmen, Tatar, Kazakh, Kyrgyz, Uzbek, Uyghur, Tuva. However, it is true that Google Translate, Yandex, Prompt machine translations still have many errors in the translation of some complex, negative sentences, idioms, scientific texts.

Assessment of the quality of translating simple sentences using a machine translation Google, Yandex Turkic languages are presented in table 1.

Table 1. Evaluation of the quality of translation of simple sentences by machine translation Google, Yandex for Turkic languages

Source text Turkic languages	Yandex MT	Google MT	Disadvantages
Дустым кичэ матур чалбар кигэн иде (Tatar language)	My friend kiche wore nice trousers	My friend was wearing nice pants last night	Google translator pays attention to punctuation. The Yandex translation of the word "kiche" is incorrect.
Сизни тәбрикләшкә иҗазәт бериң (Uyghur language)	No translation	Let me congratulate you	Uyghur translation is not included in Yandex.

			It is understood as the Tatar language.
Мені жерге қаратпа (Kazakh language)	don't put me on the ground	don't put me on the ground	Directly translated a regular expression
Бу тугрида гап хам булиши мумкин емас (Uzbek language)	It's all in tugrida and can not be found	This is out of the question	The Yandex translation could not translate the word "tugrida" and completely lost the meaning of the sentence
Birsey icmek istiyorum (Turkish language)	I want to drink birsey	I want to drink soething	The Yandex translation replaces the word "something" with the word "birsey"

Table 2. Errors in machine translation of texts of different genres

Genre	The original text is in Russian / Kazakh	Google Translate	Yandex MT
Scientific	The study, published in the journal *Cell*, compared 'brain organoids' -- 3D tissues grown from stem cells which model early brain development -- that were grown from human, gorilla and chimpanzee stem cells. Similar to actual brains, the human brain organoids grew a lot larger than the organoids from other apes. Dr Madeline Lancaster, from the MRC Laboratory of Molecular Biology, who led the study, said: "This provides some of the first insight into what is different about the developing human brain that sets us apart from our closest living relatives, the other great apes. The most striking difference between us and other apes is just how incredibly big our brains are." During the early stages of brain development, neurons are made by stem cells called neural progenitors. These progenitor cells initially have a cylindrical shape	Cell журналында жарияланған зерттеуде мидың ерте дамуын модельдейтін бағаналы жасушалардан өсірілген «мидың органоидтары» - адам, горилла және шимпанзе жасушаларынан өсірілген 3D тіндер салыстырылды. Нақты миға ұқсас, адам миының органоидтары басқа маймылдардың органоидтарынан едәуір өскен. Зерттеуді жүргізген MRC молекулалық биология зертханасының қызметкері, доктор Маделин Ланкастер: «Бұл дамып келе жатқан адам миының бізді ең жақын туыстарымыздан, басқа да маймылдардан ерекшелендіретін ерекшеліктері туралы алғашқы түсінік береді. Біздің басқа маймылдардан айырмашылығы - бұл біздің миымыздың қаншалықты үлкен екендігі ». Мидың дамуының алғашқы	Cell журналында жарияланған зерттеуде "ми органоидтары" - мидың ерте дамуын модельдейтін бағаналы жасушалардан өсетін 3D тіндер-адам, горилла және шимпанзе бағаналы жасушаларынан өсірілген. Нағыз ми сияқты, адам миының органоидтары басқа маймылдардың органоидтеріне қарағанда әлдеқайда көп өсті. Зерттеуді басқарған MRC молекулалық биология зертханасының докторы Мадлин Ланкастер: "бұл бізді ең жақын тірі туыстарымыздан, басқа антропоидты маймылдардан ерекшелейтін дамып келе жатқан адам миынан не ерекшеленетіні туралы алғашқы түсінік береді. Біз бен басқа маймылдар арасындағы ең керемет айырмашылық-бұл біздің миымыз қаншалықты керемет." Мидың дамуының алғашқы кезеңдерінде нейрондарды нейрондық прекурсорлар деп аталатын бағаналы жасушалар жасайды. Бұл прекурсорлық жасушалар бастапқыда цилиндр тәрізді, бұл оларды бірдей

	that makes it easy for them to split into identical daughter cells with the same shape.	кезеңінде нейрондарды жүйке бастаушылары деп аталатын дің жасушалары жасайды. Бұл бастаушы жасушалар бастапқыда цилиндрлік пішінге ие, сондықтан олардың пішіні бірдей еншілес жасушаларға бөлінуін жеңілдетеді.	пішінді бірдей аналық жасушаларға оңай бөлуге мүмкіндік береді.
	They are served as features in our d-gram model replacing the traditional n-gram model.To obtain dependency parse trees, we first applied TreeTagger for tokenization and POS-tagging. Next, we fed the tagged output to the MaltParser that we had trained on the Russian National Corpora.	Олар дәстүрлі n-грамм моделін алмастыратын d-gram моделінің ерекшеліктері ретінде қызмет етеді. Тәуелділікті талдау үшін біз алдымен TreeTagger-ді токенизация және POS-тегтеу үшін қолдандық. Содан кейін біз Ресейдің ұлттық корпорациясында оқыған MaltParser-ге шығарылған өнімдерді жібердік.	Олар біздің d-грамм моделімізде дәстүрлі N-грамм моделін алмастыратын функция ретінде қызмет етеді. Тәуелділіктің синтаксистік талдау ағаштарын алу үшін алдымен treetagger-ті токенизация және POS тегтеу үшін қолдандық. Содан кейін біз Ресейдің ұлттық корпустарында оқыған MaltParser - дің белгіленген нәтижесін тамақтандырдық.
Fiction	Then a lifetime went by. I bought my own apartment in St. Petersburg and was walking around the parental abode collecting my clothes and books while begging my mother for various knickknacks and for old fabrics stored in even older suitcases.	Содан кейін бүкіл өмір өтті. Мен Санкт-Петербургтен өз пәтерімді сатып алдым және ата-анамның үйін аралап жүріп, киімімді және кітаптарымды жинап жүріп, анама түрлі соққылар мен одан да үлкен чемодандарда сақталған ескі маталарды сұрадым.	Содан кейін өмір өтті. Мен өзіме Санкт-Петербургтен пәтер сатып алдым және ата-анамның монастырында жүрдім, киім мен кітаптар жинадым, анамнан одан да ескі чемодандарда сақталған әртүрлі әшекейлер мен ескі маталарды сұрадым.
	What was it like? Maybe something like this: say you have babies, they grow up and turn forty, and you get used to this fact and live with it, and then you rummage in the closet and here he is, your first baby, the way he used to be—eighteen months old, not able to talk yet, sweet-smelling like oatmeal and applesauce, his face puffy from crying, lost but found, waiting for you all	Бұл қандай еді? Мүмкін, осындай нәрсе болуы мүмкін: сіздің балаларыңыз бар деп айтыңыз, олар өсіп, қырыққа келеді, ал сіз осыған үйреніп, онымен өмір сүресіз, сосын сіз шкафта сырласасыз, міне ол сіздің алғашқы балаңыз, ол бұрынғы әдісімен - он сегіз айлық, әлі сөйлей алмайтын, сұлы мен алма тәрізді хош иісті,	Бұл не болды? Мүмкін, сізде балалар бар делік, олар өсіп, қырық жасқа толады, сіз бұл фактке үйреніп, онымен бірге өмір сүресіз, содан кейін шкафқа кіріп кетесіз, міне, ол сіздің алғашқы балаңыз, ол бұрынғыдай-он сегіз айлық, әлі сөйлей алмайтын, сұлы майы мен алма сияқты тәтті иісі бар, көз жасы ісінген, жоғалған, бірақ табылған, сізді осы онжылдықтардың бәрін шкафта күтіп, Сізді шақыра

	these decades behind the storage chest in the closet, unable to call for you—and now you are finally reunited.	жылағаннан беті ісінген, жоғалған, бірақ табылған, сізді осы оншақты жыл бойы шкафтағы сақтау қоймасының артында күтіп тұрған, сізді шақыра алмай - міне, енді сіздер қайта қауыштыңыздар	алмадыңыз—міне, сіз қайта қосылдыңыз.

In the table above, we can see some mistakes in the texts of scientific and literary genres in some complex sentences, in repeated sentences spoken by one author.

Let's take a look at the mistakes made by machine translators using one of the Turkic-speaking languages, a corps of 20,000 lines consisting of Kazakh and English:

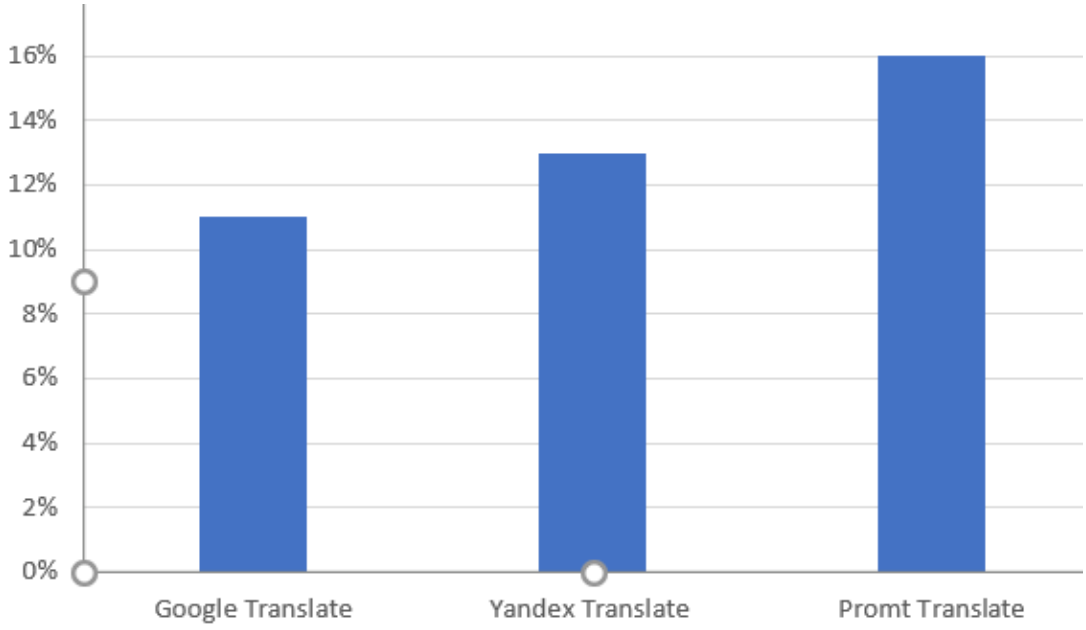

Figure 1- Errors in machine translations

Main part

The article uses the OpenNMT (Open Neural Machine Translation) model of neural machine translation to improve the quality of machine translation.

Section 1. Since the 2000s, Kazakh scholars have been studying the models of machine translations in the translation from Kazakh into other languages. The first research in Kazakhstan was conducted by Tukeyev UA can be seen in the works of [2]. In subsequent works, in 2014-2018, Tukeyev UA, who created a model for statistical machines. students Rakhimova DR [3] and Kartbayev A.Zh. [4] can be noted. Tukeev UA In his works he studied the morphological structure of words in the Kazakh language and created a model of distinguishing suffixes and suffixes [5]. In recent works, Kazakh scientists have used neural machine translation to distinguish morphology. We can see that the translation rate has increased in this work. According to this work, the Kazakh-English parallel corpus was used.

In the present study, data were used, as in [6]. They are produced by the OpenNMT method in a neural machine.

Overview of the structure and model of OpenNMT neural machine translation

OpenNMT is open code created by neurons. Thanks to OpenNMT, data does not take up much

memory, machine translation training takes less time. Implementation of training consists of three stages [7]:

1) Data preparation. It consists of words, phrases and sentences compiled from Kazakh and English, respectively, created in parallel in two files.

Data preparation is performed using the code onmt_build_vocab -config toy_en_de.yaml -n_sample 10000. The location of the file must be specified. Write in the YAML configuration file:*data:*

corpus_1:
path_src: `toy-ende/src-train.txt`
path_tgt: `toy-ende/tgt-train.txt`
valid:
path_src: `toy-ende/src-val.txt`
path_tgt: `toy-ende/tgt-val.txt`

Add the following information to the YAML configuration file:

`Model training:`
src_vocab: `toy-ende/run/example.vocab.src`
tgt_vocab: `toy-ende/run/example.vocab.tgt`
 # Train on a single GPU
world_size: 1
gpu_ranks: `[0]`
 # Where to save the checkpoints
save_model: `toy-ende/run/model`
save_checkpoint_steps: 500
train_steps: 1000
valid_steps: 500

`In this model we prepare a learning step and a neural model. Then we run this model on code onmt_train -config toy_en_de.yaml.`

In translation, it translates 1000 words or sentences from a pre-rendered file. To translate it, write the code onmt_translate -model toy-ende / run / model_step_1000.pt -src toy-ende / src-test.txt -output toy-ende / pred_1000.txt -gpu 0 –verbose.

Section 2. Carry out experimental research on the translation of one of the Turkic languages into Kazakh and show the results

One of the Turkic-speaking languages, Kazakh, was used in the experiment. Kazakh-English and English-Kazakh pairs of neural machine translations were used. A total of 109970 phrases and sentences were compiled and created. The calculation process lasted for 3 days.

Table 3. The result obtained in OpenNMT

Language pair	Speed (tok /s)	BLEU
Kazakh-English	4185	20.56
English-Kazakh	4185	20.05

As a result, it can be seen that the BLEU value is high for Turkic-speaking languages. To give a better result, you need to increase the data in the case.

Conclusion

Summarize the results obtained and note the work that is still under consideration.

By creating an OpenNMT neural model, you can see that getting a translation is no less than online translation from Google, Yandex. However, it can be seen that the translation result can be improved only by increasing the data in the cases. OpenNMT also takes less time to read data and saves memory. As a result of experimental research, it can be seen that the Kazakh-English and English-Kazakh language pairs gave good results in translation. The next study is to obtain translations of other Turkic languages according to this model. For this purpose, a corpus will be

assembled for other Turkic-speaking languages.

References

[1] Zhumanov Zh.M., Tukeyev U.A**.** Development of machine translation software logical model (translation from Kazakh into English language). Reports of the Third Congress of the World Mathematical Society of Turkic Countries, (2009). 1. 356-363.

[2] U. Tukeyev , A. Karibayeva & Z h. Zhumanov. Morphological segmentation method for Turkic language neural machine translation. 2020.

[3] Controllable Invariance through Adversarial Feature Learning. Xie, Qizhe, Dai, Zihang, Du, Yulun, Hovy, Eduard, Neubig, Graham. 2017.

[4] Mikel L. Forcado and Ramon P. Neco Recursive Hetero-Associative Memories for Translation. International Work-Conference on Artificial and Natural Neural Networks, IWANN'97 Lanzarote, Canary Islands, Spain, (1997). 453-462.

[5] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic Regularities in Continuous Space Word Representations. The Association for Computational Linguistics. In HLTNAACL. 2013. 746–751.

[6] Nal Kalchbrenner, Phil Blunsom. Recurrent Continuous Translation Models. Proceedings of the Conference on Empirical Methods in Natural Language Processing, Seattle, Washington, USA, (2013). 1700–1709.