УДК 81´322.4

Р.А. Гыйльмуллин,
А.Ф. Хөсәинов,
Ж.Ш. Сөләйманов

«TURKMT-7» – РУС ҺӘМ ТӨРКИ ТЕЛЛӘР АРАСЫНДА МАШИНА ТӘРҖЕМӘСЕ СИСТЕМАЛАРЫ КОМПЛЕКСЫН ТӨЗҮ ТУРЫНДА

В этой статье описываюся основные этапы работы над проектом TurkMT-7. Идея проекта «TurkMT-7» заключается в создании корпусов данных и нейросетевых машинных переводчиков для русско-тюркских языковых пар с ограниченными ресурсами. Достижение этой цели планируется за счет применения гибридного подхода к созданию многоязычного параллельного корпуса между русским и тюркским языками, использования различных подходов переноса знаний с одной языковой пары на другую, применения методов искусственного увеличения объема обучающих данных, а также разработка специализированных методов унификации параллельных данных на разных языках, основанная на агглютинативной природе выбранных тюркских языков (структурно-функциональная модель тюркской морфемы).

Ключевые слова: нейресетовой машинный перевод, многоязычные данные, тюркские языки.

The idea of the «TurkMT-7» project is to create datasets and neural machine translation systems for a set of Russian-Turkic low-resource language pairs. It's planned to achieve this goals through an hybrid approach to creating a multilingual parallel corpus between Russian and Turkic languages, studying the applicability and effectiveness of neural network learning methods (transfer learning, multi-task learning, back-translation, dual learning) in the context of the selected language pairs, as well as the development of specialized methods for unification of parallel data in different languages, based on the agglutinative nature of the selected Turkic languages (structural and functional model of the Turkic morpheme). In this article we describe the main stages of work on this project.

Keywords: neural machine translation, multilingual datasets, Turkic languages.

Автоматик рәвештә тәрҗемә системаларын төзү юнәлеше соңгы елларда бигрәк тә тиз үсеш алды. Аның төп сәбәпләреңең берсе – машиналы өйрәтүнең заманча ысулларын уңышлы куллану нәтиҗәсе. Әмма модельләрне өйрәтү өчен мәгълүмат җитмәгән очракта, нейрочелтәр технологияләрен дөньядагы иң танылган инглиз-алман, инглиз-кытай һ.б. парлар өчен куллангандагы кебек яхшы нәтиҗәләргә ирешү мөмкин түгел.

Аз ресурслы телләр өчен (татар теле дә бу исемлектә) аерым алынган кайбер мәсьәләләр өчен булган технологияләрне адаптацияләү һәм камилләштерү эшләре аеруча актуаль булып тора. Болар арасында «белем күчерү» (transfer learning, zero-shot learning) һәм өйрәтү

мәгълүматын ясалма рәвештә арттыру (мәсәлән, back-translation, dual learning) технологияләрен куллану аеруча уңышлы нәтиҗәләргә китерде. Әмма аз ресурслы төрки телләр төркеме булып саналган телләр өчен машина ярдәмендә күптелле (рус теленнән төрки телләргә һәм киресенчә) тәрҗемә системасын төзү турында тикшеренүләр моңарчы үткәрелмәде.

Әлеге проект җиде төрки телдән рус теленә һәм киресенчә рус теленнән җиде төрки телгә тәрҗемә итү системасын төзү өчен методик һәм программа чараларын эшләүгә юнәлтелә. Проект кысаларында куелган мәсьәләләр түбәндәгеләрдән гыйбарәт: өйрәтүче параллель мәгълүмат корпусларын туплау; төрки морфема функциональ моделе нигезендә тупланган параллель корпусларны унификацияләү методын эшләү; шулай ук белемнәрне күчерү һәм мәгълүмат күләмен ясалма рәвештә арттыру ысулларын кулланып, күптелле машина тәрҗемәчесен өйрәтү программа чараларын төзү. Болар ярдәмендә өйрәтүче мәгълүматларның аз булу проблемасын хәл итү планлаштырыла. Бу беренче тапкыр кырымтатар-рус тел парына тәрҗемә итү системасын төзергә ярдәм итәчәк. Моннан тыш, машина тәрҗемә системасы тагын алты тел пары өчен эшләячәк (татарча-русча, башкортча-русча, чувашча-русча, казакъча-русча, кыргызча-русча һәм үзбәкчә-русча). Әлеге тикшеренү күбесенчә Россия Федерациясе һәм БДБ илләре территориясендә яшәүче 57,93 млн кеше [Eberhard, David M., Gary F. Simons, and Charles D. Fennig, 2020] өчен файдалы булырга мөмкин.

Тикшеренү нәтиҗәләре машина тәрҗемәсенең сыйфатына күптөрле параметрларның (кардәш булган телләр буенча файдаланылган корпуслар күләме, ясалма рәвештә тупланган параллель мәгълүматларның кулланылышы, нейрочелтәр архитектурасын өйрәтү һәм сайлау ысулларын куллану) йогынты ясый торган дәрәҗәсе турында мәгълүмат бирәчәк.

Мәкаләнең 1 нче бүлегендә әлеге өлкәдәге тикшеренүләргә күзәтү тәкъдим ителде, 2 нче бүлек әлеге проект буенча эш планын үз эченә алды.

Соңгы елларда машина тәрҗемәсе системаларын төзүдә кулланыла торган ысуллар, технологияләр җитди үзгәрешләр кичерә. Дөньядагы иң зур кулланышлы телләр өчен дә, аз ресурслы телләр өчен дә машина тәрҗемәсе мәсьәләләрен чишү юнәлешендә тигез күләмдә шактый зур эшләр башкарыла. Аерым телләр өчен булган мәгълүмат күләме һәм сыйфатына карап, билгеле бер алгоритмнар җыелмасын һәм тәрҗемә ысулларын кулланырга мөмкин.

Машина тәрҗемәсенә караган мәсьәлә sequence-to-sequence *(бер элементлар эзлеклелеген икенче элементлар эзлеклелегенә күчерү моделе)* дип аталган [Sutskever, I, 2014] модель ярдәмендә чишелә, мәсәлән, код-лау һәм код-ачу архитектурасы (encoder/decodecoder архитектурасы) элементларын үз эченә алган рекуррент һәм төргәк нейрочелтәрләр. Иң яхшы сыйфатны күрсәткән модельләр шулай ук

«игътибар механизмын» (attention, self-attention – рекурент нейрочелтәрләрдә куллана торган ысул) үз эченә ала.

Өйрәтү процессын тизләтү һәм тәрҗемәче системаларның сыйфатын яхшырту максатыннан төзелгән нейрочелтәрләрнең төрле архитектуралары бар: рекуррент нейрочелтәрләр [Bahdanau, D., Cho, K., Bengio, Y., 2015], төргәк нейрочелтәрләр [Gehring, J., Auli, M., Grangier, D., Yarats, D., Dauphin, N.Y., 2017], Transformer һәм Evolved Transformer [Vaswani, A., 2017] (өйрәтү тизлеген арттыру өчен «игътибар механизмын» кулланучы модельләр). «Игътибар механизмы» да яңартылды: multi-hop attention, self-attention һәм multi-head attention (граф нейрочелтәрләрендә үзконтроль механизмнар) [Gehring, J., Auli, M., Grangier, D., Yarats, D., Yann N Dauphin, 2017; Paulus, R., Xiong, C., Socher, R., 2018; Johnson, M., 2017] вариантлары тәкъдим ителде.

Машина тәрҗемәсе системасын төзү технологиясен сайлау беренче чиратта өйрәтүче мәгълүматларның булуына һәм аларның күләменә бик нык бәйле. Чыганак һәм тәрҗемә телләре өчен зур күләмле бертелле корпусларның булуы машина тәрҗемәчеләрен төзү өчен unsupervised *(«укытучысыз өйрәнү»)* дип аталган алымны кулланырга мөмкинлек бирә. Әлеге алымның төп идеясе – ике тел өчен дә сүз/фразалардан торган бердәм векторлы киңлекне төзүдән гыйбарәт. Хәзерге вакытта әлеге алымның статистик [Artetxe, M., 2018], нейрочелтәр [Mikel Artetxe, Gorka Labaka, Eneko Agirre, Kyunghyun Cho, 2018] һәм гибридлы алымнар нигезендә гамәлгә ашыру вариантлары бар.

Тәрҗемә сыйфатын яхшырту өчен өйрәтүне өлешчә укытучы катнашында башкарганда бертелле корпуслардан файдалануның төрле вариантлары тәкъдим ителде: semi-supervised («өлешчә өйрәнү») [Munteanu, D.S., 2004].

Бертелле мәгълүматларны куллануның тагын бер ысулы – ул код-ачу (decoder) [Caglar, G., 2015] өчен җавап бирә торган системаның бер өлешен тел моделе белән тулыландыру. Әлеге алым IBM-ның элекке хезмәтләрендә [Brown, P.F., 1990] күрсәтелгән. Соңрак бастырылган хезмәтләрдә тәрҗемә теле өчен кулланылган өстәмә тел моделе статистика нигезендә төзелгән системалар өчен тәрҗемәнең табигыйлыгын һәм төгәллеген яхшыртырга мөмкинлек бирә, дип күрсәтелгән иде. Алга таба әлеге стратегия нейрочелтәрләр нигезендә [He, W., 2016] төзелгән тәрҗемә системалары өчен дә кулланылды. Кулланудан тыш кодны ачу вакытында, нейрочелтәрле телләр һәм тәрҗемә модельләре [Caglar, G., 2017] яшерен халәтен берләштерү исәбенә эчке дәрәҗәдә уңышлы интеграцияләнергә мөмкин. Моннан тыш, нейрочелтәр архитектурасы күпмәсьәләле өйрәнү (multi-task learning) һәм параметрларны уртак өйрәнү (parameter sharing) [Domhan, T., 2017] ысулларын кулланырга мөмкинлек бирә.

Әлеге хезмәтләрдә [Cheng, Y., 2016] авторлар бертелле мәгълүматларны автомат рәвештә кодлаштыру өчен ярдәмче мәсьәләне

өстәргә тәкъдим итәләр. Бу ысул чыганак җөмләләр эзлеклелеге нигезендә ике якка да тәрҗемә итүне тәэмин итә.

Аз ресурслы телләр өчен тәрҗемә сыйфаты синтетик мәгълүматлар хисабына яхшырырга мөмкин [Currey, A., 2017]. Анда чыганак теле җөмләләре тәрҗемә теле җөмләләренең гади күчермәсе буларак төзелә.

Өйрәтү мәгълүматларын автомат рәвештә эффектив арттыру ысулы (data-augmentation) да мәгълүм [Sennrich, R., 2015]. Ул back-translation *(кирегә тәрҗемә)* (текст буенча алга таба КТ) дип атала: чыгарылыш теленнән чыганак теленә тәрҗемә итеп, яңа тәрҗемәләрне чыганак тел җөмләләре белән катнаштырып, өйрәтүче мәгълүматлар күләмен арттыру өчен кулланыла торган ысул. Шул исәптән параллель корпус күләме дә арттырыла. Соңыннан шушы корпус машина тәрҗемәсен өйрәтү өчен файдаланыла.

Куллану ягыннан КТ гади, чөнки ул машина тәрҗемәсен өйрәтү алгоритмнарына үзгәреш кертүне таләп итми. Аз ресурслы телләр өчен өйрәтүче мәгълүматлар күләмен арттыру мәсьәләсеннән башка, машина тәрҗемәсен аерым предмет өлкәсенә яраклаштыру өчен шулай ук бертелле корпуслар да эшкә җигелә [Bertoldi, N., Federico, M., 2009]. Соңгы елларда дөнья күргән хезмәтләрдә [Sennrich, R., 2016; Lample, G., 2018] КТ ысулын яхшырту өчен синтетик парларны «нур буенча эзләү» (beam search) яки «комсыз эзләү» (greedy search) ысуллары ярдәмендә яңа җөмләләр генерацияләүдән баш тартырга тәкъдим ителде. Әйтеп кителгән алгоритмнар апостериор максимумны *(maximum a posteriori)* (текст буенча алга таба АМ) эзләргә, ягъни җөмләне модель нигезендә иң зур ихтималлык белән табарга мөмкинлек бирә. Әмма АМ ны куллану тәрҗемәләр аскорпусы төрлелегенең кимрәк булуына китерергә мөмкин, чөнки күпмәсьәләлек очракларында алгоритм һәрвакыт иң ихтимал вариантны сайлаячак [Ott, M., 2018]. Альтернатива буларак, аерым авторлар «очраклы сайлау» (random sampling) ысулын кулланырга тәкъдим итәләр [Imamura, K., 2018]. Әлеге ысул генерацияләнгән җөмләләр парының лексик төрлелеген сакларга мөмкинлек бирә. Шул ук вакытта сирәк булган тәрҗемә вариантларын төшереп калдыру өчен, өстәмә кагыйдәләр кулланырга мөмкин [Graves, A., 2013]. Моннан тыш, һәр җөмлә өчен чыганак җөмләләренең берничә вариантын да генерацияләп була. Мөһим үзгәреш [Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, Trevor Cohn, 2018] хезмәтендә бар: авторлар, финал системаларының сыйфатын яхшырту өчен, өйрәтүче корпусның синтетик өлешен итератив рәвештә өйрәтергә/өстәргә тәкъдим итте.

Аерым фәнни хезмәтләрдә ике тел өчен бертелле корпуслар булган очракта, тәрҗемәләрнең сыйфатын ничек яхшыртып булуы күрсәтелгән [Cheng, Y., 2016]. Ике корпусны берьюлы куллану ВТ алымыннан «икеләтә өйрәтү» (dual learning) ысулына күчәргә мөмкинлек бирә. Бу очракта өйрәтү бер үк вакытта тәрҗемәнең ике юнәлешендә

дә бара, ике юнәлештә дә өйрәтүче корпус күләмен һәм синтетик җөмләләр өлешен акрынлап арттыру өчен, ВТ алымы кулланыла.

Byte-pair encoding *(байтлар парын кодлаштыру)* (текст буенча алга таба БПК) [Gade, F., 1994] алымы аеруча игътибарга лаек. Ул төп элемент буларак бер сүздән кимрәк булган элементларга нигезләнгән машина тәрҗемәсе мәсьәләсе өчен кулланыла. БПК [Sennrich, R., 2016] нигезендә сегментацияләүне куллану ачык сүзлек ярдәмендә тәрҗемә итү мәсьәләсен чишәргә мөмкинлек бирә (система теләсә нинди сүзләрне тәрҗемә итә ала, шул исәптән, өйрәтүче корпусларда булмаганнарын да). БПК алымы баштан ук мәгълүматны кысу алгоритмы буларак төзелгән, ләкин сүзләрне сегментка бүлү өчен түбәндәгечә җайлаштырылган: өйрәтүче сүзлекнең һәр сүзе сүз азагының махсус символы белән тәмамлана торган символлар тезмәсе булып тора; барлык символлар элементлар сүзлегенә өстәлә; символларның иң еш очраган парлары билгеләнә, табылган тезмәләр элементлар сүзлегенә өстәлә һәм корпуска берләшә. Процедура бирелгән берләшү операцияләр саны үтәлгәнчә кабатлана.

Проектны гамәлгә ашыру планы һәм әлеге этап нәтиҗәләре

Рус теле һәм төрки телләр төркеме өчен өйрәтүче параллель мәгълүматларны туплау мәсьәләсен үз эченә алган төрле ысуллар җыелмасы ярдәмендә чишәргә тәкъдим ителә. Шул исәптән: йомгаклау корпусын ике телле Интернет-чыганаклардан тулыландыру (яңалыклар порталы, ирекле лицензияле электрон китапханәләр һәм башкалар), проектта сайланган телләр өчен тәрҗемә ителгән китап басмаларын цифрлаштыру, параллель мәгълүматларның инде төзелгән чыганакларын берләштерү. Әлеге мәсьәләне хәл итү өчен, төрле эшләр башкару сорала: мәгълүматлар чыганакларын билгеләү буенча эксперт эше; анализ өчен кирәкле булган ысулларны гамәлгә ашыру, шул исәптән документлар буенча тигезләү алгоритмнарын төзү, җөмләләр буенча тигезләү, эвристик кагыйдәләр җыелмасы нигезендә фильтрлау, шулай ук булган яхшы сыйфатлы аскорпуслар нигезендә параллель пар җөмләләр фильтрациясенең интеллектуаль модельләрен куллану.

Машина тәрҗемәсе моделен төзүнең төп мәсьәләсе нейрочелтәрләр алымы нигезендә башкарылачак. Башлангыч этапта Transformer нейрочелтәре архитектурасын куллану планлаштырыла. Аның төп үзенчәлеге – «игътибар механизмын» кулләүнә (multi-head attention) һәм төргәкле һәм рекуррент катламнарның булмавында.

Проектта игълан ителгән тел парларының күбесе өчен (казакъ-рус, татар-рус парларыннан тыш) күрсәтелгән алымны модельләр өйрәтү өчен уңышлы куллану мөмкин түгел иде. Күптелле тәрҗемәчене төзүгә юнәлтелгән комплексны алым бу проблеманы түбәндәгеләр ярдәмендә хәл итәчәк:

• белемнәрне бер тел парыннан икенчесенә күчерүнең төрле алымнарын куллану (мәсәлән, ресурслар белән тулырак тәэмин

ителгән тел пар/парларына өйрәтелгән fine-tuning нейрочелтәрләре; чыганак тел өчен нейрочелтәр архитектурасын «сүзләр катламы» (word embedding) ысулы нигезендә төзү һәм аны бердәм күптелле нейрочелтәрне өйрәтү өчен куллану);

• проектта игълан ителгән барлык телләр өчен сүз өлешләрен гомуми формага китерү буенча эш башкару (мәсәлән, барлык телләр өчен гомуми byte-pair encoding-элементлар);

• өйрәтү мәгълүматын ясалма рәвештә арттыру ысулларын куллану (back-translation) (бертелле корпусларны тәрҗемә итү нигезендә параллель мәгълүматлар күләмен арттыру өчен, тәрҗемәченең арадаш версияләрен куллану) һәм ысулларны, лексик яктан баерак тәрҗемә булсын өчен, «нур буенча эзләү» (beam search) ысулы урынына «очраклы сайлау» (random sampling) ысулын кулланып, модификация эшләрен башкару;

• төрле тел парлары өчен җыелган параллель корпусларны унификацияләү ысулларын эшләү, бу бер тел пары өчен төзелгән зуррак күләмдәге корпусларны башка парлар өчен тәрҗемә моделен өйрәткәндә тулырак файдаланырга мөмкинлек бирәчәк. Төрки морфеманың структур-функциональ моделен [32] төрле төрки телләрдә аффикслар тарафыннан башкарыла торган грамматик рольләрнең үзара бәйләнеше турындагы мәгълүматны өйрәтүче мәгълүматларда куллану максатыннан тәкъдим ителә. Әлеге мәгълүмат корпус әзерлегенең башлангыч этабында төрле телдәге морфемалар өчен бердәм элементлар формалаштырырга мөмкинлек бирәчәк.

Шулай итеп, тәрҗемә моделен өйрәтүгә һәм өйрәтүче мәгълүматларны әзерләүгә кагылышлы эшләрнең төп этаплары бер үк вакытта машиналы өйрәтү һәм кагыйдәләргә нигезләнгән (төрки морфеманың структур-функциональ моделендәге телләр өчен мәгълүматлар базасы) алымнарга таяна.

Йомгаклау этабында кулланылган алымнарны, җайлауларны, нейрочелтәр гиперпараметрларының кыйммәтен һәм кулланылган өйрәтүче мәгълүматларның күләмен ачыклау максатыннан машина тәрҗемәсе системаларының төрле вариантлары өчен тест үткәреләчәк.

Бу мәкаләдә без әле башланып киткән җиде пар тел өчен машина тәрҗемә системасы комплексына караган «TurkMT-7» исемле проектны тәкъдим иттек. Аны гамәлгә ашыруның төп үзенчәлеге – машиналы өйрәтү буенча заманча технологияләр һәм төрки телләр үзенчәлекләренә нигезләнгән алымнар ярдәмендә башкаруда.

Әдәбият

Eberhard, David M., Gary F. Simons, and Charles D. Fennig (eds.). 2020. Ethnologue: Languages of the World. Twenty-third edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.

Sutskever, I. Sequence to sequence learning with neural networks [Text] / Ilya Sutskever, Oriol Vinyals, and Quoc V.Le. // Advances in Neural Information Processing Systems. 2014. P. 3104–3112.

Bahdanau, D., Cho, K., Bengio, Y. Neural machine translation by jointly learning to align and translate [Text] / Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio // International Conference on Learning Representations (ICLR). 2015.

Gehring, J., Auli, M., Grangier, D., Yarats, D., Dauphin, N.Y. Convolutional sequence to sequence learning [Text] / Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin // International Conference of Machine Learning (ICML). 2017.

Vaswani, A. Attention is all you need [Text] / Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin // Conference on Advances in Neural Information Processing Systems (NIPS). 2017.

Gehring, J., Auli, M., Grangier, D., Yarats, D., Yann N Dauphin. Convolutional sequence to sequence learning [Text] / Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin // International Conference of Machine Learning (ICML). 2017.

Paulus, R., Xiong, C., Socher, R. A deep reinforced model for abstractive summarization [Text] / Romain Paulus, Caiming Xiong, and Richard Socher // International Conference on Learning Representations (ICLR). 2018.

Johnson, M. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation [Text] / Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viegas, Martin Wattenberg, Gregory S. Corrado, Macduff Hughes, Jeffrey Dean // Transactions of the Association for Computational Linguistics, Vol. 5. 2017. P. 339–351.

Artetxe, M. Unsupervised Statistical Machine Translation [Text] / Mikel Artetxe, Gorka Labaka, Eneko Agirre // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018. P. 3632–3642.

Artetxe, M. Unsupervised Neural Machine Translation [Text] / Mikel Artetxe, Gorka Labaka, Eneko Agirre, Kyunghyun Cho // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018. P. 3632–3640.

Lample, G. Phrase-Based & Neural Unsupervised Machine Translation [Text] / Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, Marc'Aurelio Ranzato // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018.

Munteanu, D.S. Improved machine translation performance via parallel sentence extraction from comparable corpora [Text] / D.S. Munteanu, A. Fraser, and D. Marcu // ACL. 2004.

Caglar, G. On using monolingual corpora in neural machine translation [Text] / Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio // arXiv preprint arXiv:1503.03535. 2015.

Brown, P.F. A statistical approach to machine translation [Text] / Peter F. Brown, John Cocke, Stephen Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin // Computational Linguistics, 16. 1990. P. 79–85.

Koehn, Ph. Statistical phrase-based translation [Text] / Philipp Koehn, Franz Josef Och, and Daniel Marcu // Conference of the North American Chapter of the Association for Computational Linguistics (NAACL). 2003.

He, W. Improved neural machine translation with smt features [Text] / Wei He, Zhongjun He, Hua Wu, and Haifeng Wang // Conference of the Association for the Advancement of Artificial Intelligence (AAAI). 2016. P. 151–157.

Caglar, G. On integrating a language model into neural machine translation [Text] / Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, and Yoshua Bengio // Computer Speech & Language, 45. 2017. P. 137–148.

Domhan, T. Using targetside monolingual data for neural machine translation through multi-task learning [Text] / Tobias Domhan, Felix Hieber // Conference on Empirical Methods in Natural Language Processing (EMNLP). 2017.

Cheng, Y. Semi-supervised learning for neural machine translation [Text] / Y. Cheng, W. Xu, Z. He, W. He, H. Wu, M. Sun, and Y. Liu // arXiv:1606.04596. 2016.

Currey, A. Copied Monolingual Data Improves Low-Resource Neural Machine Translation [Text] / Anna Currey, Antonio Valerio Miceli Barone, and Kenneth Heafield // Proc. of WMT. 2017.

Sennrich, R. Improving neural machine translation models with monolingual data [Text] / Rico Sennrich, Barry Haddow, and Alexandra Birch // arXiv preprint arXiv:1511.06709. 2015.

Bertoldi, N., Federico, M. Domain adaptation for statistical machine translation with monolingual resources [Text] / Nicola Bertoldi and Marcello Federico // Workshop on Statistical Machine Translation (WMT). 2009.

Sennrich, R. Improving neural machine translation models with monolingual data [Text] / Rico Sennrich, Barry Haddow, and Alexandra Birch // Conference of the Association for Computational Linguistics (ACL). 2016.

Lample, G. Unsupervised machine translation using monolingual corpora only [Text] /Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato // International Conference on Learning Representations (ICLR). 2018.

Ott, M. Analyzing uncertainty in neural machine translation [Text] / Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato // Proceedings of the 35th International Conference on Machine Learning. Vol. 80. 2018. P. 3956–3965.

Imamura, K. Enhancement of encoder and attention using target monolingual corpora in neural machine translation [Text] / Kenji Imamura, Atsushi Fujita, and Eiichiro Sumita // Proceedings of the 2nd Workshop on Neural Machine Translation and Generation. 2018. P. 55–63.

Graves, A. Generating sequences with recurrent neural networks [Text] / Alex Graves // arXiv, 1308.0850. 2013.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, Trevor Cohn. Iterative backtranslation for neural machine translation [Text] / Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn // Proceedings of the 2nd Workshop on Neural Machine Translation and Generation. 2018. P. 18–24.

Cheng, Y. Semisupervised learning for neural machine translation [Text] / Yong Cheng, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu // Conference of the Association for Computational Linguistics (ACL). 2016.

Gade, F. A New Algorithm for Data Compression [Text] / Philip Gage // C Users J., 12(2):23–38, February.1994.

Sennrich, R. Neural Machine Translation of Rare Words with Subword Units [Text] / Rico Sennrich, Barry Haddow, and Alexandra Birch // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016). Berlin, Germany. 2016.]

Сулейманов Д.Ш., Гатиатуллин А.Р., Альменова А.Б., Баширов А.М. Многофункциональная модель тюркской морфемы как база данных для лингвопроцессоров // Филология и культура. 2016. № 2 (44). С. 143–151.

Сөләйманов Ж.Ш., А.Р. Гатиатуллин, Гыйльмуллин Р.А., Хөсәенов А.Ф. Татар теле һәм яңа инфокоммуникацион технологияләр // Фәнни Татарстан. 2020. № 2. Б. 92б.

Khusainov A., Suleymanov D., Gilmullin R. The Influence of Different Methods on the Quality of the Russian-Tatar Neural Machine Translation. In: Kuznetsov S.O., Panov A.I., Yakovlev K.S. (eds) Artificial Intelligence. RCAI 2020. Lecture Notes in Computer Science, vol 12412. Springer, Cham. 2020. P. 251–261.

Гыйльмуллин Ренат Абрек улы,
физика-математика фәннәре кандидаты,
ТР ФА Гамәли семиотика институты директоры

Хөсәенов Айдар Фаил улы,
техник фәннәр кандидаты, ТР ФА Гамәли семиотика институты директор урынбасары

Сөләйманов Җәүдәт Шәүкәт улы,
техник фәннәр докторы, ТР ФА Гамәли семиотика институты фәнни җитәкчесе – баш фәнни хезмәткәре