УДК 81′33:811.512.145

М.М. Әюпов,
Р.Р. Сафина

РУС-ТАТАР МАШИНА ТӘРҖЕМӘЧЕСЕ ӨЧЕН ПАРАЛЛЕЛЬ ТЕКСТЛАРНЫҢ МӘГЪЛҮМАТ БАЗАСЫН ӘЗЕРЛӘГӘНДӘ ОЧРЫЙ ТОРГАН АВЫРЛЫКЛАР

В статье описываются проблемы, возникающие при подготовке данных для базы данных русско-татарских параллельных текстов и предлагаются пути их решения. База данных русско-татарских параллельных текстов в дальнейшем используется для обучения русско-татарского машинного переводчика.

Ключевые слова: русско-татарский машинный переводчик, база данных, татарский язык.

The article describes the problems that arise when preparing data for a database of Russian-Tatar parallel texts and suggests ways to solve them. The database of Russian-Tatar parallel texts is subsequently used to train a Russian-Tatar machine translator.

Keywords: Russian-Tatar machine translator, database, Tatar language.

«Татсофт» (https://translate.tatar/) рус-татар машина тәрҗемәчесе [Khusainov, Suleymanov, Gilmullin, 2020, p. 251–261] аерым сүзләр яки фразаларны түгел, ә бөтен җөмләне «аңлап» тәрҗемә итә торган иң беренче нейросеть машина тәрҗемәчеләре исәбенә керә. Хәзерге вакытта «Татсофт» үз аналоглары арасында (Яндекс һәм Google) тәрҗемә сыйфаты буенча иң яхшысы булып тора. Мондый системаны эшләү өчен зур күләмдә параллель җөмләләр (дистәләгән, йөзләгән миллион пар җөмләләр) таләп ителә. Дөньякүләм киң таралган инглиз, кытай телләре кебек зур телләрдән аермалы буларак, татар теле аз ресурслы телләргә керә һәм, шуңа бәйле рәвештә, рус-татар машина тәрҗемәчесен ясаган вакытта параллель текстларның мәгълүмат базасын тутыру иң катлаулы мәсьәлә булып тора. Әлеге мәсьәләне хәл итү өчен безнең тарафтан мәгълүмат җыю, эшкәртү һәм киңәйтү ысуллары һәм технологияләре эшләнде.

Машина тәрҗемәчесен өйрәтү өчен рус-татар параллель текстларының мәгълүмат базасын туплау түбәндәге этапларны үз эченә ала:

– рус һәм татар телләрендә текстлар җыю;

– текстларны җөмләләргә бүлү һәм ике телдәге параллель җөмләләрне үзара тигезләү;

– параллель җөмләләр белән тәрҗемә итүнең дөреслеген тикшерү;

– җөмләләрдә булган пунктуацион, грамматик, синтаксик һәм башка хаталарны төзәтү.

Рус һәм татар телләрендә текстлар җыю

Рус-татар параллель текстларының мәгълүмат базасын туплау эшенең беренче этабында бер телдәге электрон текстларны һәм ул текстларның икенче телгә тәрҗемәләрен җыярга кирәк. Китапларны, журналларны һәм башка басма продукцияне цифрлаштыру – текстлар җыюның бер ысулы. Безнең тикшерү күрсәткәнчә, татар телендәге күпчелек басма продукциясенең электрон нөсхәсе юк, булган очракта да – тулы түгел. Яки, әсәрнең электрон нөсхәсе бер телдә булса, икенче телгә тәрҗемәсе кәгазьдә генә теркәлгән. Шуңа күрә текстлар җыю өчен басма чыганаклардан файдалану зарурлыгы туа. Алга таба бу эшне башкару этапларын карап китик.

Иң элек (беренче адым) кирәкле әсәрнең сканлаштырылган нөсхәсен табарга яки, әгәр яхшы сыйфатлы нөсхә булмаса, әсәрне сканлаштырырга кирәк. Аннан соң (икенче адым) әсәр танып-белү махсус программалары ярдәмендә эшкәртелә. Өченче адымда текстан ярымавтомат рәвештә алга таба кирәк булмаган мәгълүмат (мәсәлән, рәсемнәр, бит номерлары) алына. Безнең тарафтан әлеге ысулны кулланып, 283 926 бит текстның электрон нөсхәсе эшләнде.

Электрон параллель текстларны тулыландыруның икенче ысулы – булган интернет ресурслардан файдалану. Аларга мөрәҗәгать иткәндә, анда тәкъдим ителгән материалның бертөрле түгел икәнен истә тотарга кирәк. Шулай ук бу чыганакларның ышанычлы булуына инану сорала. Бу ысулны куллану өчен татар һәм рус телләрендә бертөрле эчтәлектәге текстлар булган интернет ресурслардан автомат рәвештә мәгълүмат җыя торган программа эшләнде. Ул түбәндәге мөмкинлекләргә ия:

– татар һәм рус телләрендәге текстларны җыеп алу;

– җыелган текстлар арасында «татар телендә тәрҗемә» һәм «рус телендә тәрҗемә» бәйләнешләрен табу.

Әлеге программа ярдәмендә рус һәм татар телләрендә бәйләнешле 617 754 пар бит җыелды.

Электрон текстларны тулыландыруның өченче ысулы – татар җөмләләре җыелмасын ясалма арттыру. Моның өчен төрки морфотәрҗемәче, татар теленең барлык тел күренешләрен максималь исәпкә алып, башкорт, кырым татар телләре өчен бераз үзгәртеп яңартылды. Нәтиҗәдә, кагыйдәләргә нигезләнеп эшләүче башкорт-татар, татар-башкорт, татар-кырым татар һәм кырым татар-татар машина тәрҗемәчесе төрки телләрдә электрон корпусларны арттыру мөмкинлегенә ия булды. Ясалма алынган өстәмә электрон корпуслар машина тәрҗемәчесен өйрәтү өчен кулланыла, бу татар-рус һәм рус-татар машина тәрҗемәчесенең сыйфатын яхшыртырга ярдәм итә.

Текстларны җөмләләргә бүлү һәм параллель җөмләләрне тигезләү

Рус һәм татар телләрендәге параллель текстларның күбесеннән тәрҗемәсе туры килгән өлешләре турындагы мәгълүматны турыдан-

туры алып та булмый. Беренчедән, тәрҗемә ителгән һәм оригинал текстлардагы сүзләр арасында бер генә мәгънәле туры килү булмый, шулай ук грамматик төзелештә аерымлыклар, берничә мәгънәле сүзләр очрый. Икенчедән, бер җөмләнең тәрҗемәсе берничә җөмлә белән бирелә, һәм, киресенчә, берничә җөмлә бер җөмлә белән тәрҗемә ителә ала. Ниһаять, тәрҗемәдә төгәлсезлекләр булырга мөмкин, иң зур төгәлсезлек – төшеп калган сүзләр [Хакимов, Шаехов, 2021].

Шуңа күрә текстларны җөмләләргә бүлү – катлаулы мәсьәлә. Аны җөмләнең башын һәм ахырын күрсәткән синтаксик билгеләрне кулланып хәл итәргә мөмкин. Аерым тел лексикасы турындагы мәгълүматны – кыскартылмалар исемлеген һ.б. файдаланырга да була. Бусы – четереклерәк ысул.

Параллель җөмләләрне тигезләгәндә, җөмләләрдә сүзләрнең килү тәртибе лексик бәйләнешләрне табу өчен төп мәгълүмат чыганагы буларак кулланыла. Машина тәрҗемәчесе эшенең уңышы җөмләләрне тигезләү төгәллеге дәрәҗәсенә бәйле.

Параллель текстларны тигезләү өчен, беренче адымда ABBYY Aligner 2.0 программасы кулланылды. Ул төрле телләрдәге текстларда бер-берсенә туры килгән җөмләләрне таба, аларны үзара чагыштыра һәм автомат рәвештә тигезләнгән өлешләрне билгели. Аннан соң, икенче адымда, тигезләү вакытында киткән хаталарны төзәтү һәм нәтиҗәнең сыйфатын яхшырту өчен, тигезләнгән текстлар кулдан карап чыгылды. Әлеге эш кысаларында барлыгы 954 829 параллель җөмлә тигезләнде.

Параллель җөмләләрдә тәрҗемә ителүнең дөреслеген тикшерү

Сыйфатлы тәрҗемә оригинал текстның мәгънәсен төгәл тапшырырга тиеш. Яхшы сыйфатка ирешү өчен, гадәти тикшерүдән тыш, түбәндәгеләрне исәпкә алырга кирәк:

– терминнар барлык текстларда да бер үк төрле тәрҗемә ителергә тиеш;

– әгәр сүзнең берничә язылыш варианты булса, алар бердәм язылышка китерелергә тиеш;

– саннарның язылышы аерым тел кагыйдәләренә туры килергә тиеш.

Шулай ук җөмләләрдә хаталар юклыкка инану өчен, пунктуацион, грамматик һәм орфографик хаталарны төзәтеп чыгу сорала.

Әлеге эшләр кулдан башкарылды.

Мәгълүмат җыйганда очраган кыенлыклар

Рус-татар параллель текстларын туплаган вакытта бик күп авырлыклар белән очрашабыз. Аларның кайберләренә тукталып үтик.

Сканлаштырылган текстларны махсус танып-белү программалары ярдәмендә эшкәрткәндә, ышанычлы булмаган һәм танып белү программасы сүзлегендә булмаган сүзләр очраса, орфографик хаталар

барлыкка килә. Аерым килергә тиешле сүзләрнең кушылып язылуы – танылган текстларда шулай ук киң таралган очрак. Мондый хаталарны автомат рәвештә төзәтеп булмый, текстны кулдан карап чыккан вакытта гына аларны ачыкларга һәм төзәтергә мөмкин. Мондый төзәтүләрне ABBYY Aligner 2.0 программасы ярдәмендә параллель җөмләләрне тигезләгәнче башкарырга кирәк, чөнки әлеге хаталар тигезләү сыйфатына йогынты ясый.

Сканлаштырылган текстларны таныганда, рәсемнәр һәм рәсем асты язулары, таблицалар һәм таблица исемнәре, бит номерлары, төрле формулалар һ.б. алга таба кирәк булмаган мәгълүмат алып ташланырга тиеш. Моны гамәлгә ашыру өчен, документны текст форматында сакларга була, әмма бу очракта рәсем асты язуларын, бит номерларын алып ташлау катлауланачак, чөнки алар текстның бер өлешенә, ә таблицалар, үз функциональ мөмкинлекләрен югалтып, эзлекле рәвештә килүче абзацларга әйләнә. Шуңа күрә, алга таба кирәге булмаган мәгълүматны алып ташлау өчен, программа эшләргә карар кылынды, ул алынырга тиешле мәгълүматның түбәндәге үзенчәлекләренә таяна.

Әгәр дә бит номерлары өске колонтитулга кермәсә һәм бит астында килсә, сканлаштырылган текст танылганнан соң, алар аерым юлда килүче бер берәмлеккә үскән саннарны тәшкил итә һәм гади алгоритм ярдәмендә табыла. Әгәр дә бит номерлары өске колонтитулга керсә, алар колонтитуллар өчен булган алгоритмны кулланганнан соң юкка чыга.

Өске колонтитул танылганнан соң аерым юлга урнаша һәм төп тексттан аерылып торган (гадәттә, кечкенәрәк зурлыктагы) шрифтлы була. Җәп санлы бит колонтитуллары барысы бертөрле булып, бу шулай ук так санлы бит колонтитулларына да кагыла. Колонтитуллар бер-берсеннән аларга бит номерлары кергән очракта гына аерылалар. Шул ук вакытта ул аерма колонтитулның башында яки ахырында, бит номеры өчен җавап биргән санда гына булачак.

Рәсемнәр, гадәттә, төп текст белән рәсем арасында килгән рәсем асты язуы белән бергә бара. Рәсем асты язуының тагын бер үзенчәлеге – аның зурлыгы төп текстныкыннан аерылып тора һәм күпчелек очракта башка шрифттан була.

Таблицалар рәсемнәр кебек эшкәртелә, аермасы – таблицага аңлатма язуының таблицадан соң түгел, ә аңа кадәр килүендә генә.

Шулай итеп, алга таба кирәк булмаган мәгълүмат махсус язылган программа ярдәмендә тексттан алып ташлана, бу очракта дөрес нәтиҗә 90% тәшкил итә.

Сүзне юлдан юлга күчерү текстның тышкы күренешен яхшырта. Шуңа күрә ул китап басу эшендә киң кулланыла. Сканлаштырылган текст танылганнан соң, сүз, юлдан юлга күчерелгән очракта, сүз башына, сызыкка, юл ахыры билгесенә һәм сүзнең күчерелгән өлешенә әверелә. Бу очракта бүленгән сүзнең кушма сүз яки сызыкча аша языла торган сүз булу ихтималын исәпкә алырга кирәк. Өстәвенә,

сызыкча аша язылган сүзләр чагыштырмача сирәк очрый, алар юлдан юлга күчкән сүзләрнең уртача 2% ын тәшкил итә. Шуңа күрә юлдан юлга күчкән сүзләрнең барысын да кушылып языла торган сүзләр дип карарга мөмкин булыр иде. Әмма рус-татар параллель текстларының мәгълүмат базасы әлеге текстлар хисабына тулылана барганга күрә, андагы хаталы сүзләр саны да артачак. Димәк, юлдан юлга күчкән сүзләрне эшкәртергә кирәк һәм моның өчен татар теленең морфологик анализаторы [Gatiatullin, Ayupov, 2015, p. 120–126] һәм түбәндәге алгоритм кулланыла.

Беренче адымда андый сүз кушылып языларга тиеш, дип фаразлана, шуңа күрә юлдан юлга күчкән сүздән сызык һәм юл ахыры билгесе алына; ул сүз морфологик анализаторга тапшырыла. Морфологик анализатор бу сүзне тамырга һәм морфемаларга таркатырга тырыша. Әгәр дә бу эш барып чыкса, сүз кушылып языларга тиеш санала. Сүзне тамыр һәм аффикслар тезмәсе рәвешендә күрсәтеп булмаган очракта, морфологик анализаторга сүз сызыкча аша язылып бирелә. Бу юлы сүз дөрес дип табылса, әлеге сүз сызыкча аша языла дип исәпләнә.

Мисал өчен 1 нче рәсемдәге беренче юлны алыйк. Башта морфологик анализаторга кушылып язылган *хатынкыз* сүзен бирәбез, әлеге сүз морфологик анализатор базасында табылмады дигән җавап алачакбыз, димәк, сүз хаталы язылган. Икенче адымда сызыкча аша язылган *хатын-кыз* сүзен бирәбез. Морфологик анализатор бу сүзне тамыр сүз дип җавап кайтарачак, димәк әлеге сүз дөрес язылган. Шуңа күрә ул сызыкча аша языларга тиеш, дигән фикергә киләбез.

Зәмзәмия гаепне заманага да тага. Мөселман хатын-кызларына куелган таләпләр кысасында яшәсә, ул шундый көнгә калыр идемени? Гүзәл зат, бөек ана, ир-атлар белән хатын-кызларның хокуклары тигез, дип, күккә чөеп, хатын-кызны бозып бетерделәр. Бу мактауны күтәрә алмыйча, күпләр тайгак юлдан китеп барды. Зәмзәмия үзен әнә шундыйларның берсе итеп саный. Аның фикеренчә, аракы эчкән, тәмәке тарткан, ят жирләрен күрсәтергә атлыгып торган, ирләр муенына асылынырга хирыс хатын-кызлар иң модалы затка әйләнделәр. Тыйнаклык, гүзәллек, сөйкемлелек, хатын-кыз горурлыгы турында сүз катарга урын калмады…

1 нче рәсем. Сканлаштырылган текст үрнәге

1 нче рәсемдәге соңгы юлны карап үтик. Бу очракта беренче адымда морфологик анализаторга *калмады* сүзе бирелә. Җавап итеп, сүз *кал-* тамыры һәм *-ма* һәм *-ды* морфемаларына таркалып кайта, ягъни андый сүз бар һәм ул дөрес язылган. Димәк, әлеге сүз кушылып языларга тиеш.

Юлдан юлга күчкән сүзнең башында яки ахырында сызыкча килгән очраклар да була. Аларда юлдан юлга күчерү өчен хезмәт иткән сызыкча һәм юл ахыры билгесе автомат рәвештә алып ташлана.

Бу алгоритмны куллану, дөрес танылмаган сүзләрдән кала, сүзне юлдан юлга күчерүнең барлык очракларын дөрес эшкәртергә мөмкинлек бирә.

Машина тәрҗемәчесен өйрәтү өчен сыйфатлы параллель текстлар кирәк, әмма әлеге таләп һәрвакытта да канәгатьләндерелми. Бигрәк тә бу Интернеттан алынган мәгълүматларга кагыла. Кайвакыт уртаклыклары аз яки бөтенләй булмаган җөмлә парлары очрый. Шуңа күрә, сыйфатлы машина тәрҗемәчесе булдыру өчен, рус-татар параллель текстларының мәгълүмат базасын мондый сыйфатсыз парлардан чистартырга туры килә.

Йомгаклау. Безнең тарафтан эшләнгән мәгълүмат җыю, эшкәртү һәм киңәйтү ысуллары рус-татар параллель текстларының зур күләмле һәм сыйфатлы базасын төзергә мөмкинлек бирде. Бу, үз чиратында, «Татсофт»ның бүгенге көндә, үз аналоглары (Google, Яндекс) белән чагыштырганда, рус һәм татар телләре арасында тәрҗемә итү сыйфаты буенча иң яхшы тәрҗемәче булуына сәбәп булды. Алга таба да рус-татар параллель текстларының мәгълүмат базасы күләмен арттыру буенча эш дәвам иттереләчәк. Димәк, русчадан татарчага татар машина тәрҗемәсе тагын да камилләшәчәк.

Әдәбият

Khusainov A., Suleymanov D., Gilmullin R. (2020) 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. P. 251–261. URL: https://doi.org/10.1007/978-3-030-59535-7_18.

Хакимов Б.Э., Шаехов М.Р. Проблема эквивалентности параллельных предложений в тестовом корпусе для русско-татарского машинного переводчика // Proceedings of the 9th International Conference on Turkic Languages Processing (TURKLANG-2021). (Tyva, September 21–23, 2021). Tyva, 2021.

Gatiatullin A., Ayupov M. Modifications of morphological analysis programs for the problems of multilingual search // Proceedings of the International Conference «Turkic Languages Processing» TurkLang-2015. Kazan, 2015. P. 120–126.

Әюпов Мәдехүр Мәсхүт улы,
ТР ФА Гамәли семиотика институты фәнни хезмәткәре,
Казан федераль университеты өлкән укытучысы

Сафина Рәзинә Рәфкать кызы,
ТР ФА Гамәли семиотика институты кече фәнни хезмәткәре