«6D070300 – Ақпараттық жүйелер» мамандығының Phd докторанты Шормакова Асем Ноябревнаның «Ағылшын тілінен қазақ тіліне машиналық аударманың пост-редакциялау үлгілерін, әдістерін және программалық құралдарын зерттеу және өңдеу» тақырыбындағы диссертациялық жұмысына

АҢДАТПА

Зерттеу тақырыбының өзектілігі ақпараттық жүйелер саласындағы машиналық аударма мен пост-редакциялаудың заманауи дамуымен байланысты. Ақпараттық жүйелер қазіргі қоғам өмірінің барлық дерлік салаларында қолданылады. Сонымен қатар, ақпараттық технологиялар әр салада тиімділік пен өнімділікті арттырады және көптеген артықшылықтарға ие, сондықтан білім беру және басқа салаларда өсу үшін машиналық аударма өзекті болып табылады. Бүгінгі таңда пайдаланушылар үшін атап айтқанда, интерактивті ақпараттық жүйелер саласында машиналық аударма сапасы маңызды рөл атқарады.

Машиналық аударма – ақпараттық жүйелер саласындағы жасанды интеллекттің жетекші бағыттарының бірі. Машиналық аударма дүние жүзіндегі халықтар мен елдер арасындағы байланысты жақсартудың жаһандық мәселесін шешуде маңызды рөл атқарады. Машиналық аударманың сапасы жылдан-жылға артып келеді, бірақ кәсіби аударма сапасына әлі жеткен жоқ.

Машиналық аударманың сапасын жақсарту үшін ең маңызды және практикалық әдістердің бірі – пост-редакциялау үрдісі, яғни машиналық аударма сапасын жақсарту мақсатында машиналық аударманы түзету. Машиналық аударманы пост-редакциялау қолмен де, автоматтандырылған нұсқаларда да жүзеге асыруға болады. Машиналық аударманы қолмен пост-редакциялау – біршама еңбекті қажет ететін үрдіс. Автоматтандырылған пост-редакциялау машиналық аударма бағыты табиғи тілдердің машиналық аудармасының өзекті бағыттарының бірі болып табылады.

Соңғы жылдары машиналық аударма қолданушылар саны қарқынды өскен, әсіресе оқу орындары, жекеменшік кәсіпорындар, аударма орталықтары жиі қолданып жүр. Сонымен қоса шет елдегі компаниялардың басым көпшілігі машиналық аударманың көмегіне жүгініп жүр. Осыған қоса көптеген қолданушылар күнделікті тұрмыс жағдайында да машиналық аударманы қолданады.

Қазақ тілі машиналық аудармасы кәсіби аудармашылардың денгейіне жеткен жоқ әлі де, сол себептен қазіргі уақытта постредакциалау бағытын қолданып қазақ тілі машиналық аудармасының сапасын көтеру өте өзекті мәселе болып тұр.

Бұл жұмыстың ғылыми үлесі – ағылшын – қазақ сөйлемдерді туралау әдісімен талдау негізінде қате аударылған сөзді табу, табылған қате аударылған сөзге мағынасы жақын сөздердің тізімін (каталог) қалыптастыру және олардың ішінен лексикалық таңдау тапсырмасының технологиясын қолдана отырып, ең ықтимал дұрыс сөзді таңдауға негізделген қазақ тіліне арналған автоматты пост-редакциялау технологиясын өңдеу болып табылады.

Диссертациялық жұмыстың мақсаты. Бұл жұмыстың негізгі мақсаты – лексикалық таңдау негізінде машиналық аударманы автоматты түрде пост-редакциялау арқылы ағылшын тілінен қазақ тіліне машиналық аударманың сапасын арттыру.

Зерттелу есептері: Осы мақсатқа жету үшін 4 тапсырма қарастырылды:

1 – қазақ тіліндегі сөйлемнің қате аударылған сөздерін анықтау;

2 – қате аударылған сөздердің синонимдер каталогын автоматты түрде қалыптастыру;

3 – қате аударылған сөзді мағынасы жағынан жақын синониммен ауыстырып түзетілген сөйлемнің машиналық аудармасын шығару.

4 – жоғарыдағы үш тапсырманы біріктіріп пост-редакциялау технологиясын құрастыру.

Зерттеу әдістері: табиғи тілдерді өңдеу үлгілері мен әдістері.

Зерттеу нысаны: ағылшын тілінен қазақ тіліне машиналық аударманың мәтіндері.

Зерттеу пәні: ағылшын тілінен қазақ тіліне машиналық аударманы автоматты түрде пост-редакциялау.

Жұмыстың ғылыми жаңалығы.

1) Алғаш рет ағылшын – қазақ машиналық аударманың Post Edit – Lexical Choice (PE-LC) пост-редакциялау технологиясы әзірленді.

2) Ағылшын тілінен қазақ тіліне қате аударылған сөздерді табу әдісі кері аудармамен жетілдірілген.

3) Алғаш рет қате аударылған қазақ сөздердің синонимдер каталогын автоматты түрде қалыптастыру әдісі әзірленді.

4) Қате аударылған сөздің ықтималдығы жоғары синоним сөзді таңдау семантикалық текше әдісінің үлгісі мен алгоритмі бейімделді.

Зерттеудің теориялық құндылығы: Зерттеудің теориялық маңыздылығы ағылшын тілінен қазақ тіліне машиналық аударманы пост-редакциялау технологиясына мәтіндерді өңдеудің белгілі әдістерін әзірлеуде және біріктіруде болып табылады.

Зерттеудің практикалық құндылығы: Зерттеудің практикалық маңыздылығы ағылшыннан қазақшаға аударылған мәтінге пост-редакциялау технологиясын құру және бағдарламалық жабдықтау құралдарын өңдеп қолдану болып табылады.

Қорғауға шығарылатын негізгі жағдайлар:

1. Ағылшын–қазақ машиналық аударманың жаңа автоматты пост-редакциялау технологиясы.
2. Ағылшын тілінен қазақ тіліне қате аударылған сөздерді анықтаудың жетілдірілген әдісі.
3. Ағылшын тілінен қате аударылған қазақ сөздер синонимдерінің каталогын автоматты түрде қалыптастыру технологиясы.
4. Семантикалық текше негізінде ықтималдығы жоғары синонимді таңдау бейімделген әдісі.

Сенімділік дәрежесі мен апробациялау нәтижелері. Алынған нәтижелердің сенімділігін пост-редакциялау технологиясының эксперименттерінің нәтижелері, журналдардағы жарияланымдар мен халықаралық конференциялар материалдарындағы апробациялау нәтижелерінен көруге болады.

Жұмыстың ғылыми нәтижелері келесі халықаралық ғылыми конференциялар мен ғылыми семинарларда ұсынылып, талқыланды:

• ACIIDS 2017 интеллектуалды ақпарат және деректер базасы жүйелері бойынша 9-шы Азия конференциясы;

• Есептеу ұжымдық интеллект бойынша 11-ші халықаралық конференция ICCCI 2019;

• «Фараби әлемі» студенттер мен жас ғалымдардың халықаралық ғылыми конференциясы, Алматы, 2014, 2015, 2017, 2018 ж.

Сондай-ақ бұл тақырып әл-Фараби атындағы Қазақ ұлттық университетінің ақпараттық жүйелер кафедрасында және ақпараттық технологиялар факультетінің ғылыми семинарларында талқыланды.

Диссертациялық тақырыптың ғылыми бағдарламалармен байланысы. Диссертациялық жұмыс PhD докторлық диссертациясының жоспарына сәйкес және «Қазақ тіліндегі ақпараттық-аналитикалық деректерді іздеу жүйесін дамыту» гранттық қаржыландыру жобасының ғылыми-зерттеу жұмыстарының жоспарына сәйкес орындалды. (2018-2020 ж., мемлекеттік тіркеу нөмірі: No AP05132950). Диссертациялық жұмыс бойынша жүргізілген кейбір зерттеу нәтижелері осы жобаның 2018-2020 жылдарға арналған есептеріне енгізілген.

Әрбір басылымды дайындаудағы докторанттың үлесі. Жарияланған мақалалар мен ғылыми еңбектер диссертация тақырыбы бойынша зерттеу нәтижелерін сипаттайды. Ғылыми жұмыс барысында 14 ғылыми жұмыс жазылды, оның ішінде: Scopus индексі бар журналда 1 ғылыми мақала жарыққа шықты:

1. Shormakova A., Zhumanov Z.H., Rakhimova D. "Post-editing of words in Kazakh sentences for information retrieval". *Journal of Theoretical and Applied Information Technology,* 2019, 97(6), p. 1896–1908. (Scopus 2021: Q4, CiteScore-1.3; Percentile- 30%)

Қазақстан Республикасы Білім және ғылым министрлігі Білім және ғылым саласындағы бақылау комитеті ұсынған журналдарда 4 мақала шықты:

1. Абеустанова (Шормакова) А.Н. "Машиналық аударманың нарықтағы және Қазақстандағы күйі". *ҚазҰТУ хабаршысы* № 6(106), 2014. –150-152 б.
2. Абеустанова (Шормакова) А.Н. "Қазақ тіліндегі көпмағыналы сөздердің бірін анықтаудың бір болжамы". *ҚазҰТУ хабаршысы* №4(110) 2015. –625-628 б.
3. Абеустанова (Шормакова) А.Н. "Ағылшын тілінен қазақ тіліне аударылған қазақша қате сөздерді анықтау және баламалар каталогын құру". *ҚазҰТУ хабаршысы* №6 2017. –313-317 б.
4. Шормакова А.Н. "Екі табиғи тілдегі аударылған мәтінді туралау". *ҚазҰТУ хабаршысы*, №4(128), 2018. –344-349 б.

Scopus негізінде индекстелген халықаралық ғылыми-тәжірибелік конференциялар жинақтарында 2 ғылыми мақала жарияланған:

1. Abeustanova (Shormakova) A., Tukeyev U. "Automatic Post-editing of Kazakh Sentences Machine Translated from English". *Studies in Computational Intelligence/Advanced Topics in Intelligent Information and Database Systems*, vol. 710 – Springer International Publishing, 2017. – p. 283-295. (Scopus 2021: Q4, CiteScore-1.8; Percentile- 27%).
2. Rakhimova D., Assem S. "Problems of Semantics of Words of the Kazakh Language in the Information Retrieval". *Lecture Notes in Computer Science* (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2019, 11684 LNAI, p. 70–81. (Scopus 2021: Q2, SJR=0.25, CS=2.1, Percentile-50%)

Халықаралық ғылыми конференциялар жинақтарында 6 және ғылыми-техникалық журналда 1 мақала жарияланған:

1. Shormakova A. "Machine translation and post-editing". *Материалы международной конференции студентов и молодых ученых «Мир науки»*, 17-19 апреля 2013г. – Алматы: Қазақ университеті, 2013. – с. 222

2. Шормакова А.Н. "Информатика терминдерінің мемлекеттік тілге аудару ерекшеліктері". *Материалы III международного конгресса студентов и молодых ученых «Мир науки»*, 23-28 апреля, 2009г.-Алматы: Қазақ университеті,- с. 249.

3. Шормакова А.Н., Тукеев У.А. "Технология машинного перевода с обучением английского языка на казахский язык". *Материалы международной конференции студентов и молодых ученых «Мир науки»*, 23-26 апреля 2012г. – Алматы: Қазақ университеті, – с. 154.

4. Sundetova A., Forcada M.L., Shormakova A., Aitkulova A. "Structural transfer rules for English-Kazakh machine translation in the free/open-source platform Apertium", in *Proceedings of the I International Conference on Computer Processing of Turkic Languages (TurkLang-2013)* (Astana, 3-4 oct. 2013) , p. 322-331.

5. Шормакова А.Н., Айткулова А. "Добавление новой англо-казахской языковой пары в платформу машинного перевода Апертиум". *51-я Международная научная студенческая конференция «Студент и научно-технический прогресс»*, Новосибирск, 12-18 апреля 2013, Секция "Информационные технологии".- с. 241.

6. Тукеев У.А., Абеустанова (Шормакова) А.Н., Сундетова А. "Ағылшын-қазақ тілдік жұбы үшін Apertium платформасындағы сөйлемді синтаксистік құрылымдық түрлендіру ережелері және мәселелері" . *IV международная научно-практическая конференция:* (секция «Искусственный интеллект»). Қоғамды ақпараттандыру IV Халықаралық ғылыми-практикалық конференция еңбектері, Астана 2014 , 127-129 б.

7. Шормакова А.Н. Қазақ тіліндегі автоматтандырылған синонимдер тізімін құру [Мәтін] / А.Н. Шормакова, У.А. Тукеев // Механика және технологиялар / Ғылыми журнал. – 2022. – №3(77). – Б.44-49. https://doi.org/ 10.55956/AEQO3045

Диссертация құрылымы мен көлемі. Диссертациялық жұмыс кіріспеден, алты бөлімнен, қорытындыдан, пайдаланған әдебиеттер тізімінен және екі қосымшадан тұрады. 77 беттік машинамен жазылған мәтінді құрайды, оның ішіне 12 кесте, 7 сурет кіреді.

Кіріспеде диссертациялық жұмыстың өзектілігін негіздейді. Жұмыстың мақсаты, зерттеу жұмысының объектісі мен пәні тұжырымдалды. Ғылыми жаңалығы мен практикалық маңыздылығы анықталды. Жүргізілген зерттеу нәтижелері сипатталған. Зерттеу және жариялау нәтижелерін апробациялау туралы ақпарат берілген.

Бірінші бөлімде машиналық аудармаға және автоматты пост-редакциялауға шолу берілген. Диссертациялық жұмысқа қатысты қолданылған терминдер мен ұғымдар келтірілген. Пост-редакциялау бойынша бар, жаңа ғылыми жұмыстар сипатталған. Осы тақырып бойынша ғылыми еңбектерге сілтеме жасалып, шолу жасалды.

Екінші бөлімде PE-LC пост-редакциялау технологиясының құрылымы мен алгоритмі сипатталған. Диссертацияда қойылған төрт тапсырма туралы қысқаша ақпарат берілген. Зерттеу жұмысында ұсынылған PE-LC технологиясының жалпы алгоритмі жазылған.

Үшінші бөлімде Бірінші тапсырманың шешімі жан-жақты қарастырылды: аударылған сөйлемдегі қате аударылған сөзді анықтау. Ағылшын тілінен қазақ тіліне қате аударылған сөздерді анықтаудың жетілдірілген әдісі сипатталған.

Төртінші бөлімде қате аударылған сөздерден жасалған синонимдердің автоматты каталогын (тізімін) жасаудан тұратын екінші тапсырма сипатталды. Каталогты автоматты түрде қалыптастыруға арналған құралдар мен сілтемелер таныстырылды. Каталогты автоматты түрде қалыптастыру тапсырмасының қате аударылған сөздердің синонимдерінің мысалдары келтірілген.

Бесінші бөлімде үшінші тапсырма, яғни қате аударылған сөздің лексикалық таңдау мәселесі сипатталған. Семантикалық текше әдісінің жетілдірілген үлгісі мен алгоритмі негізінде берілген қате аударылған сөзге ең қолайлы синонимді таңдау ұсынылған. Табылған қате аударылған сөздерге семантикалық текшені құру кезінде кестелер мен мысалдар есептеулері келтірілген.

Алтыншы бөлімде ұсынылған PE-LC технологиясы мен оның Google Translate-пен салыстыру эксперименттерінен кейін алынған нәтижелер көрсетілген. Ұсынылған жұмыстағы жақсартуларды анықтау үшін эксперименттік мәліметтердің статистикалық мәнділігі есептеліп көрсетілді. Зерттеу жұмысының нәтижелерін салыстыру үшін бірнеше құралдар мен көрсеткіштер пайдаланылды.

Қорытындыда диссертацияда алынған негізгі нәтижелер тұжырымдалды.