ОӘК 347.78.034

ТІЛДІК МОДЕЛЬ ЖӘНЕ МАШИНАЛЫҚ АУДАРМА

Тәжиева Дана Бекқызы
danchic99@mail.ru
Л.Н.Гумилев атындағы ЕҰУ Аударма теориясы мен практикасы кафедрасының магистранты,Нұр-Сұлтан, Қазақстан
Ғылыми жетекшісі – А.М. Кызырова

Жазбаша аударманы жасайтын қолжетімді интернет ресурстары күнделікті қолданысқа түскеннен бері машиналық аудармаға деген сұраныс пен оның зерттелу өзектілігі өсе бастады. Бұл мақалада заманауи компьютердің тілді талдап синтездеу қағидаттары сипатталып, практикалық көрініс үшін Google Translate автоматты аударма жүйесінің ағылшын тілінен қазақ тіліне аудармасы талданады.

Машиналық аударма немесе автоматты аударма толықтай компьютерлік жүйемен орындалатындықтан, бұл мәселені зерттейтін қолданбалы тіл білімінің тармағы компьютерлік лингвистика деп аталады. Компьютерлік лингвистика тіл қызметі мен тілдік ақпаратты модельдеуді және осыған қажетті бағдарламаларды зерттейді.

Компьютер дедуктивті логикалық жүйе негізінде қызмет атқаратындықтан, табиғи тілді, яғни адамның тілін және сәйкесінше индуктивті-ықтималдық ойлау үдерісін, компьютермен талдау әрі синтездеу алгоритм мен формула арқылы қайталау мүмкін емес. Тіл семиотикалық тұрғыдан кодтар жүйесі болғанымен, оның лингвистикалық көрінісі әлдеқайда күрделі. Ұлттық тіл нұсқасынан басқа адамның жеке тілі ажыратылады. Дара тұлғалардың сөйлесуі тілдің коммуникативтік құрал ретіндегі қызметін білдіреді. Мәтіндік ақпарат адамнан компьютерге компьютер мен адам арасындағы коммуникативтік жүйе арқылы ауысқанда, табиғи тіл өз көрінісі ретіндегі жазбамен алмасады. Ақпараттың компьютерден адам назарына кері ауысу үдерісінде жазба мәтіннің компьютердегі көрінісі қайтадан таңбалардың психикалық кескін үйлесімінің бұрынғы қалпына келтіріледі. Тілдік белгі мен оны білдіруші өзара психикалық пішін үйлесімінде ғана байланысты болмайды. Олар мазмұнды тұрғыдағы үнемі өзгерісте болатын экстралингвистикалық ұғымдармен ықтималдық байланыстарға ұшырайды. Тілдегі семантикалық құбылыстар негізінен табиғи тілдің мазмұндық болмысының толық формалдануына кедергі жасайды. [1, 21-25 бб.]

Лексемалардың ықтималдық байланысы компьютерлік бағдарламаның ішінде тілдік модельмен қамтамасыз етіледі. «Please turn your homework…» сөйлем басын оқығанда адамның ой өрісі келесі сөзді болжауға тырысады. Осы сөздер комбинациясының негізінде ойға «in»

немесе «over» сияқты сөйлемді толықтыратын және тілдік нормаға сәйкес варианттар келу мүмкін. Біз логикалық тұрғыда «refrigerator» не «the» сияқты лексемалардың мұнда ұйқаспайтынын білеміз. Енді компьютерді осындай мәнмәтіндік байланыстарды үйретуге қалай болатынын қарастырайық. [2, p.30]

Тілдердің түрленетін синтактикалық құрылымы тілдік модельдерді машиналық аударма үшін аса маңызды етеді. Мысал ретінде қытай-ағылшын, иероглиф және латиница таңбаларымен белгіленген тілдік жұптағы аударманы алайық.

他 向 记者 **介**绍了 **主要 内容** *He to reporters introduced main content Ол - тілшілер таныстыру негізгі мазмұн*

Тілдік модель аударма тілінің узусына сәйкес орын тәртібін және айқын тіркестер комбинациясына ізденіп, бастапқы варианттардың арасынан ең ұтымдысын ұсынады.

He introduced reporters to the main contents of the statement He briefed to reporters the main contents of the statement

He briefed reporters on the main contents of the statement

Сөйлемдер мен сөз тізбектерінің ықтималдылығын есептейтін тілдік модельдің ең қарапайымы әрі танымалы n-грамма деп аталады. Бұл – n-санды сөздердің тізбегі. 2-грамма немесе биграмма екі сөз тізбегін өңдейді: «please turn», «turn you're» және «your homework».

3-грамма немесе триграмма үш сөз тізбегін өңдейді: «please turn your» және «turn your homework». [2, p.30]

Биграмма деңгейінде ықтималдылық Марков қағидасы бойынша есептеледі, яғни келесі сөз бірінші сөзбен белгіленеді. Базадағы барлық лексемалар берілген сөзбен үйлесе бермейді. Сондықтан сөздің тізбекке үйлесуі 0 және 1 ықтималдылық арасында жатады. [2, p.32]

Алайда сөздер омонимиясы және полисемиясы басқа тілде оқырманға анық өнімді жасау үшін екі сөз тізбек жеткіліксіз болып табылады. Бағдарламаға берілетін коммуникациялық контекст n санымен белгіленеді. Шынайы аударма тілдік модель 3- грамма, 4-грамма, тіпті 5-грамма деңгейінде орындалады. Мұның арқасында компьютердің қателер саны азаяды. [2, p.34]

Құрылымы жақындау неміс және ағылшын тілдеріндегі төрт сөз тізбегін қарастырайық. Неміс тілінде кейбір сөздер жалқы есім болмағанымен міндетті түрде бас әріппен жазылу тиіс. Алайда табиғи тілді өңдеу сөздерді немесе токендерді ажыратқан кезде оларды кіші әріппен жазып жүйеге енгізеді. Сондықтан аударма коммуникация жағдаятыныңанықтығын қажет етеді.

das behaupten sie wenigstens that claim they at least

the she

you

Неміс тіліндегі «das» артикль де, есімдік те болу мүмкін, ал «sie» есімдігінің «ол» (әйел), «сіз» (бас әріппен Sie) және «олар» мағыналары ажыратылады. Бұл – полисемия, көпмағыналылық құбылысы, сол себепті бұл сөздердің мағынасын тізбектегі токендер арқылы айқындауға болады. Google Translate бұл сөздер комбинациясын «at least, that's what they say» деп аударады, яғни лексемалар орын тәртібін ағылшын тілінің узусына сәйкестіріп, үтір токенін қосы арқылы басқа тіл синтаксисін бейімдейді. [3, p.7]

Автоматты аударма жүйенің жұмыс істеу принципін біле отырып, енді машиналық аударманың сапасын бағалау үшін аударма теориясының эквиваленттілік мәселесін қарастырайық. Эквиваленттілік терминін енгізген ғалым В.Н. Комиссаров осы құбылыстың бес типін ажыратады. Эквиваленттілік деңгейлері мәтін мазмұнын бір тілден екінші тілге аудару үдерісіндегі сақталған элементтер бойынша жіктеледі. [4, 115 б.]

Бірінші деңгей: *«That's a pretty thing to say» - «Ұялсаңшы».*

Ағылшын және қазақ нұсқаларын салыстыра отырып, сөйлемнің коммуникативті мақсаты ғана сақталғаны анық. Түпнұсқаның өзі сөзбе-сөз аудармаға келмейді, себебі қазақ тіліндегі осы

лексика-семантикалық бірліктер тізбегі оқырманға ештеңе бере алмайды. Ал тиісті әсер тигізетін балама қазақ мәдениетінде анық ұғылады. [4, 115 б.]

Екінші деңгей: *«He answered the phone» - «Он поднял трубку» - «Ол қоңырауға жауап берді».*

Берілген аудармаларда коммуникативті мақсат пен жағдаят сипаттамасы жеткізілгенімен, үш тілде бір әрекетті сипаттау үшін үш түрлі тіркес қолданылады. [4, 115 б.]

Үшінші деңгей: *«Scrubbing makes me bad tempered» – «Еденді жуу саладарынан менің көңіл-күйім бұзылады».*

Осы типте коммуникативті мақсат пен жағдаят сипаттамасымен қатар оларды жеткізуде қолданылған түпнұсқа құралдары да сақталады. [4, 116 б.]

Төртінші деңгей: *«I told him what I thought of her» - «Я сказал ему свое мнение о ней».*

Мұнда аталған элементтерге тағы түпнұсқаның синтактикалық құрылымы қосылады. Ағылшын және орыс тілдерінің сөйлем мүшелерінің орын тәртібі көбінесе сәйкес келетіндіктен, аудармашы осы типке жататын мәтінді құрастыра алады. Ал қазақ тілінің баяндауышы міндетті түрде сөйлемнің соңында тұру керектігінен, сөйлемнің баламасы бір эквиваленттілік деңгей төмен (Мен оған сол әйел туралы пікірімді айттым). [4, 116 б.]

Бесінші деңгей: *«The house was sold for 10 thousand dollars» – «Дом был продан за 10 тысяч долларов» (Үй 10 мың долларға сатылды).*

Бесінші типте түпнұсқаның тілдік бірліктері аудармада сақталып берілген. Бұл ең толық аударма болып есептеледі. [4, 116 б.]

Эквиваленттілік деңгейі белгілі бір тілдік жұптың лингвистикалық ерекшеліктерімен ғана емес, мәтін стилі, мазмұны және реципиентімен белгіленеді. С. Пинкердің The Language Instinct кітабында осы тіл мен ойлау үдерісінің байланысын көрсету үшін түрлі жанр-стильді мәтін мысалдары келтіріледі. Машиналық аударманың қабілетін бағалау мақсатында публицистикалық мәтін үлгісін талдап көрейік.

«Cherries jubilee on a white suit? Wine on an altar cloth? Apply club soda immediately. It works beautifully to remove the stains from fabrics.» [5, p.15]

«Cherries jubilee» атауы американдық ағылшын тілінде «отқа түскен қара түсті шие мен бренди немесе кирш қосылған ваниль балмұздақты» білдіреді. [6] «Altar cloth» тіркесіндегі

«алтарь» сөзі бірнеше мағынағы ие: тура және ауыспалы мағынада қолданылатын құрбандық орны; христиан шіркеуіндегі Евхаристияны тойлаудағы үстелге ұқсас конструкция; табынушылық немесе салт жоралар үшін қолданылатын үстел немесе орын. [7] Тіркестердің барлығы екі зат есімнен тұрғандықтан, автоматты аударма биграмма деңгейінде оларды жеке атау ретінде тану тиіс. Алайда мәтіннің коммуникативті мақсаты американдық христиандарына тұрмыстағы кеңес беру болса және парақтағы белгілер шектелумен келсе, реалияны сипаттап түсіндірунің қажеті әрине жоқ. Осы лексико-семантикалық бірліктер нақты АҚШ мәдениетіне қатысты болып, қазақ оқырманымен иллокутивті және перлокутивті деңгейлерінде түсінілмеуі мүмкін. Сонымен қатар ағылшын және қазақ тілдерінің сөйлемді құрастырудағы сөздер орын тәртібі айрықша болып келеді. Сондықтан қазақ тіліндегі аударма эквиваленттіліктің екінші деңгейінде орындалу тиіс.

Аударма теориясынан анық болғандай, толық сөйлем бір сөзбен берілу мүмкін жағдайларда тілдік модельдің жұмысы қиындатылады. Тілдік модель жұмысының шынайы өнімін талдап, шыққан қателерінің себептерін анықтау үшін біз осы үзіндіні Google Translate сервисіне жүктейміз. Бұл сервистің интернет нұсқасы (GNMT) нейрондық машиналық аударма болып табылады. Бұл автоматты аударманың түрі жасанды сана негізінде қызмет етіп, уақыт өте келе интернетке жүктелген деректер бойынша үйренуге қабілетті. [8, с.90]

«Ақ костюмдегі шие мерейтойы? Құрбандық үстеліндегі шарап? Клуб содасын дереу жағыңыз. Бұл матадан дақтарды кетіру үшін әдемі жұмыс істейді». [9]

«Cherries jubilee» бір тіркес болып аударылудың орнына жеке лексема ретінде берілгендіктен, Google Translate автоматты аудармасы толық сөйлем ретінде ұйқаспай, коммуникация жағдаятын ғана емес, коммуникативті мақсатты да сауатты жеткізе алмай тұр. Құрбандық үстелі altar лексеманың сөздіктегі алғашқы анықтамасы арқылы аударылғанымен, қазақ тілінің нормалық ережелерін сақтайды. Алайда шараптың дақ ретіндегі коннотациясы жоғалып кетіп, машиналық аударма қысқа мәтіннің мәніне әле де жетпей тұр. Келесі сөйлем бұйрық райда тұрып, сөзбе-сөз аудармада орындалғанда барабар нұсқа берер еді, бірақ тілдік модель американдық ағылшын тілінде «сода қосылған судың» атауы болып табылатын «club soda» [6] тіркесін тағы да бөлек лексема болып аударады. Аударманың соңғы сөйлемді эквиваленттіліктің үшінші деңгейіне жатқызуға болар еді, бірақ «әдемі жұмыс істеу» сөздер тізбегі қазақ тілінде тіркес ретінде қолданылмайды.

Тілдік модельдің Марков қағидаты қазақ тілінен ағылшын тіліне автоматты аударма жасауда өзінің толық әлеуетін аша алмады. Бұл машиналық аудармадағы осы тілдердегі базаның толықсыздығын, жасанды сананың мәтінді аударма қажет ететін аудармашылық эквиваленттіліктің типін таңдай алмауын және бір мәдениеттің баламасыз лексикасын танымайтынын көрсетеді. Біз машиналық аударма сапасын көтеру үшін қазақ тіліндегі мәтін, тіркес пен сөздер базасын толықтырып, Google Translate жасанды санасына кеңейтілген ақпарат көлеміне бейімделуге уақыт беруді жөн көреміз.

Қолданылған әдебиеттер тізімі

1. Жұбанов А.Қ. Компьютерлік лингвистикаға кіріспе : оқу құралы / А.Қ. Жұбанов. - 2-бас. - Алматы : КИЕ, 2013. – 204 б.
2. Jurafsky D., Martin J.H. Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Draft of December 30, 2020.
3. Koehn P. Neural Machine Translation. Cambridge: Cambridge University Press, 2020. — 408 p.
4. Аударма теориясы мен практикасының негіздері [Text] : оқулық / Әбішева, Клара Мухамедиярқызы, Рахымжанов, Қанат Хисматұлы; ҚР Білім және ғылым министрлігі ; "Тұран-Астана" университеті ; Қазіргі лингвистикалық білім - Астана : [б. ж.], 2011 . - 242 б.
5. Pinker, Steven, 1954-. The language instinct / [Steven Pinker], p. cm. Originally published: New York : W. Morrow and Co., cl994
6. https://www.collinsdictionary.com 26.03.2022
7. https://www.merriam-webster.com 26.03.2022
8. В. А. Нуриев, Архитектура системы нейронного машинного перевода, Информ. и её примен., 2019, том 13, выпуск 3, 90–96. https://doi.org/10.14357/19922264190313
9. https://translate.google.com 01.06.2021