ӘӨЖ 004.832.22

ПАРАЛЛЕЛЬ КОРПУСТЫ ҚҰРУДЫҢ ТИІМДІ ТЕХНОЛОГИЯЛАРЫ

Нұрбақыт Айдана
Dana.charlie98@gmail.com
Л.Н. Гумилев атындағы Еуразия ұлттық университеті Ақпараттық технологиялар факультеті «Информатика» мамандығының 2-курс магистранты, Нұр-Сұлтан, Қазақстан
Ғылыми жетекші - т.ғ.к., доцент Кудубаева Сауле Альжановна

Аңдатпа. Машиналық аударманы зерттеуде параллель корпустың маңыздылығы артып келеді. Екі тілді параллель корпусқа негізделген әртүрлі әдістер автоматты машиналық аударманың сапасын арттырып қана қоймай, сонымен қатар машиналық аудармада адам мен компьютердің өзара әрекеттесуін күшейте түсті. Интернетте көптеген тілдер бар, әр тіл өзіндік ерекшелікке ие, бұл кедергілер аударма қызметіне жоғары талаптар қояды. Бұл мақалада екі тілді параллель корпус құру әдісі қарастырылады, ол мыналарды қамтиды: 1) екі тілді параллель мәтіндер жинақтау; 2) тілдер арасында бір-бірімен сәйкестендіру; 3) мәтіндерді реттеу; 4) екі тілді параллель корпус құру.

Кілттік сөздер: Параллель корпус, туралау, жинақтау, машиналық аударма, Қазақ-Қытай параллель корпусы.

Корпус ғылыми іріктеуден өткен және өңделген ауқымды электронды мәтіндік кітапхананы білдіреді. Компьютерлік талдау құралдарының көмегімен ғылыми зерттеулер мен кәсіпорын бөлімшелері тіл бойынша тиісті теориялық және қолданбалы зерттеулер жүргізе алады. Екі тілді параллель корпус – бұл екі тілдегі дискурс, абзац және сөйлем деңгейіндегі теңестірілген мәтіндер.

Корпус табиғи тілді өңдеу технологиясының көптеген салаларының негізі болып табылады деп айтуға болады. Корпус тіліне сәйкес корпусты біртілді (Monolingual), қос тілді (Bilingual) және көптілді (Multilingual) корпус деп бөлуге болады. Корпустың жинақ бірлігі бойынша корпусты мәтінге, сөйлемге және сөз тіркесіне бөлуге болады. Екі тілді және көптілді корпусты корпустың ұйымдастырылуына қарай параллель (тураланған) корпус (Parallel Corpora) және салыстырмалы корпус (Comparable Corpora) деп бөлінеді. Параллель корпус аударма қатынасын құрайды, ол көбінесе машиналық аудармада, екі тілді сөздік құрастыруда және басқа қолданбалы салаларда қолданылады. Салыстырмалы корпус бір мазмұнды білдіретін әртүрлі тілдердегі мәтіндерді жинайды және көбінесе тілді салыстыру үшін қолданылады.

Ауқымды «корпустар» үлкен деректер дәуірінің «сүйіктісіне» айналды. Бүгінгі күні біз осы ауқымды мәтіндерден пайдалы ақпараттың барлық түрін ала аламыз. Жалпы айтқанда, корпус – бұл әртүрлі табиғи тілді өңдеудің негізгі жұмысы (мысалы, машиналық аударма, пиньинь және қытай таңбаларын түрлендіру, сөйлеуді тану, мәтінді жіктеу және кластерлеу, адам-машина сұрақтарына жауап беру жүйесі және т.б.). Онсыз қазіргі статистикалық әдістердің негізі жоқ. Әртүрлі тереңдікте өңделген нақты мәтіндер корпусы (жалпы мәтін, сегменттелген мәтін, таңбаланған мәтін, семантикалық мәтін, тарау бойынша тураланған мәтін, сөйлемді туралау мәтіні және т.б.) табиғи тілдің статистикалық қасиеттерін зерттеуге негіз болады. Корпустың маңыздылығын ескере отырып, «үлкен масштабты», «нақты» мәтіндік корпусты құру маңызды.

Құрылған корпусты мәтіннің туралау деңгейінен ажырату үшін оны сөз тіркесін туралау мәтіні (phrase alignment), сөйлемді туралау мәтіні (sentence alignment) және құжатты туралау мәтіні (document alignment) деп бөлуге болады. Олардың ішінде сөйлем деңгейінде теңестіру қазіргі табиғи тілді өңдеудің көптеген салаларында таптырмас рөл атқарады.

1) Екі тілді параллель корпустың базасын жинақтау: біріншіден теңестірілген параллель мәтіндерді қолдануға дайын барлық маңызды мәліметтерді табу; бірнеше тілде бірдей мәтіндері бар веб-сайттарды тексеріп шығу және оларды туралау үшін text aligner құралын пайдалану; бірнеше тілде бірдей мәтіндері бар веб-сайттардағы мәтіндерді тексеру үшін сценарийлерді пайдалану және оларды туралау үшін біріктірілген hunalign құралымен InterText құралын пайдалану[1].

2) Автоматты туралау: Параллель мәтін – бұл мәтін және оның аудармасы (бұл жағдайда ол биттік мәтін деп аталады) немесе аудармалар арқылы құрылған жиынтық. Параллель мәтінді теңестіру – бұл биттік мәтіннің әрбір жартысының блоктары немесе таңбалауыштары арасындағы сәйкестіктерді анықтау міндеті. Туралanған параллель мәтіндер қазіргі уақытта кең ауқымда қолданылады: білім саласы, машиналық оқыту, табиғи тілді өңдеу және т.б. Битмәтіндер әдетте екі мәтін арасында туралауды орындау арқылы алынады. Бұл теңестіруді әдетте абзац, сөйлем немесе тіпті сөз деңгейінде жасауға болады [2].

3) мәтіндерді реттеу келесі қадамдардан турады:

– интернетте параллель мәтіндерді қарап шығу;
– жиналған мәтіндерді тазалау және пішімдеу;
– сөйлемдерді бөлу;
– сөйлемдерді туралау;
– Адамның қатысуы арқылы қолмен тексеру.

Тазалау, пішімдеу және сөйлемдерді бөлуден кейін бізде екі тілдегі сөйлемдердің екі тізімі пайда болады, олар бір-бірінің аудармасы бола алады, бірақ әртүрлі аударма себептеріне байланысты сөйлемдердің сәйкес келмеуі мүмкін.

Деректерді алдын ала өңдеу параллель корпусты дайындаудағы маңызды және негізгі қадам болып табылады [3].

Жиналған параллель деректердің пішімдері, тыныс белгілері және басқа да маңызды емес мазмұны құрамы болғандықтан, одан қолдануға болатын корпусты дайындау өте қиын және көп уақытты қажет етеді. Алдын ала өңдеу барсында әр тілдегі қажет емес сілтемелер, сандар, белгілер және шетелдік мәтіндер жойылады. Сонымен қатар, таңбаларды қалыпқа келтіру, сөйлемді таңбалау, сөйлемді теңестіру және тазалау орындалады.

– Таңбаларды қалыпқа келтіру. Мағынасы бірдей сөздердің әртүрлі сөздер болып қабылданбауы үшін біз осы ұқсас қызметі бар таңбалар жиынын бір жиі қолданылатын таңбаға ауыстырамыз.

– Токенизация немесе сегменттеу – тыныс белгілерін сөздерден бөлу сияқты қарапайым процестерді немесе морфологиялық өңдеулерді қолдану сияқты күрделірек процестерді қамтитын кең ұғым. Тыныс белгілерін ажырату және лексиканы сөздерге немесе қосалқы сөздерге бөлу сөздік қорды азайтуға және әрбір сөзге мысалдар санын көбейтуге, белгілі бір тілдер үшін аударма сапасын жақсартуға көмектесетіні дәлелденді. Токенизация сөздер арасында бөлгішсіз тілдермен жұмыс істегенде қиынырақ. Екі тіл де сөз деңгейіндегі токенизацияны пайдаланады. Бұл кезеңде орындалған негізгі тапсырма сөздерді тыныс белгілерінен ажырату болды.

– True-Casing біз бұл тапсырманы корпустағы әрбір сөйлемнің дұрыс бас әріппен жазылуын қамтамасыз ету үшін орындаймыз. Бұған жету үшін біз Moses кірістірілген truecaser сценарийін қолдандық [4].

– Тазалау қадамы бос жолдарды жою үшін орындалады; таңбалар мен сөздер арасында артық бос орындарды болдырмау; және параллель корпустағы өте ұзақ сөйлемдерді бір уақытта кесіп тастау және алып тастау [5].

4) Екі тілді параллель корпус құру.

Екі тілді параллель корпус аударманы оқытудың таптырмайтын және маңызды анықтамалық материалы және жұмыс платформасы болып табылады. Ол мұғалімдер мен студенттерге түсіндіру және еліктеу үшін сөздер, сөз тіркестері және сөйлемдер деңгейінде бай екі тілді аударма мысалдарымен қамтамасыз ете алады. Бұл екі тілді оқыту және меңгеру үшін маңызды ресурс. Корпустың параллель құрылысы – бөлшектерге назар аударуды қажет ететін жұмыс. Қолданыстағы құрылыс процесінің негізінде сәйкес тілдердің сипаттамаларын түсіну және корпусты алдын-ала өңдеуде жақсы жұмыс жасау ұсынылады. Сынақ аударма процесі аудармашының сапасын және корпус жағдайын толық көрсете алады, бұл түпнұсқа корпусты уақтылы түзетуге және жобаның орындалу барысын бақылауға маңызды әсер етеді.

Қолданылған әдебиеттер тізімі

1. New Kazakh parallel text corpora with on-line access Zhandos Zhumanov1 , Aigerim Madiyeva2 and Diana Rakhimova3 al-Farabi Kazakh National University, Laboratory of Intelligent Information Systems, Almaty, Kazakhstan.

2. A survey on parallel corpora alignment Conference Paper · January 2011.

3. Rauf, S., Holger, S.: Parallel sentence generation from comparable corpora for improved SMT. Machine translation 25(4), 341–375 (2011)

4. Lita, L. V, Ittycheriah, A., Roukos, S., Kambhatla, N.: tRuEcasIng. In: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, pp. 152—159. Sapporo, Japan (2003).

5. Parallel Corpora Preparation for English-Amharic Machine Translation Conference Paper June 2021.