

«БД070300 – Ақпараттық жүйелер» мамандығы бойынша философия докторы (PhD) дәрежесіне іздену үшін ұсынылған Кәрібаева Айдана Сейілғазықызының «Нейронды машиналық аударма үшін қазақ тіліндегі мәтіндерді морфологиялық сегментациясының модельдері мен әдістерін құру және зерттеу» тақырыбындағы диссертациялық жұмысына ресми рецензенттің

СЫН-ШҚІРІ

р/н №	Критерийлер	Критерийлер сәйкестігі	Ресми рецензенттің ұстанымы
1.	Диссертация тақырыбының (бекіту күніне) ғылымның даму бағыттарына және/немесе мемлекеттік бағдарламаларға сәйкес болуы	<p>1.1 Ғылымның даму бағыттарына және/немесе мемлекеттік бағдарламаларға сәйкестігі:</p> <p>1) <u>Диссертация мемлекет бюджетінен қаржыландырылатын жобаның немесе нысаналы бағдарламаның аясында орындалған</u> (жобаның немесе бағдарламаның атауы мен нөмірі);</p> <p>2) Диссертация басқа мемлекеттік бағдарлама аясында орындалған (бағдарламаның атауы)</p> <p>3) Диссертация Қазақстан Республикасының Үкіметі жанындағы Жоғары ғылыми-техникалық комиссия бекіткен ғылым дамуының басым бағытына сәйкес (бағытын көрсету)</p>	<p>Диссертация тақырыбы «Ақпараттық, коммуникациялық және ғарыштық технологиялар» бағыты бойынша ғылым дамуының басым бағытына сәйкес келеді.</p> <p>1) Диссертациялық жұмыс PhD докторлық диссертация жоспарына сәйкес және Қазақстан Республикасы Білім және ғылым министрлігінің АР05131415 «Қазақ тілінің нейронды машиналық аудармасын құру және зерттеу» (2017-2020 жж., ғылыми жетекші: Тукеев У.А.) тақырыбы бойынша мемлекет бюджетінен қаржыландырылған жоба аясында жүргізілген.</p>
2.	Ғылымға маңыздылығы	Жұмыс ғылымға елеулі үлесін <u>қосады/қоспайды</u> , ал оның маңыздылығы <u>ашылған/ашылмаған</u> .	Жұмыс жасанды интеллекттің табиғи тілді өңдеу (Natural language processing) саласына, атап айтқанда қазақ тілінің машиналық аударма және қазақ тілін компьютерлік өңдеу зерттеу бағытына айтарлықтай үлес қосады. Зерттеу аясында алынған нәтижелер ғылыми тұрғыдан маңызды, өйткені қазақ тілін табиғи тілді өңдеуде күрделі морфологиясына байланысты қиын мәселелерін шешуге көмектеседі. Машиналық аудармада сегментация жасаудың маңыздылығы толығымен дәлелденіп ашылған.

Б.А.

3.	Өзі жазу принципі	<p>Өзі жазу деңгейі:</p> <ol style="list-style-type: none"> 1) жоғары; 2) орташа; 3) төмен; 4) өзі жазбаған 	<p>Диссертациялық жұмысты орындаушының жұмысты жазу, сипаттауы, рәсімдеуі жоғары деңгейде. Кәрібаева А.С. зерттеу жасау барысында жоғары дербестік көрсеткен. Ізденуші қазақ тілінің нейронды машиналық аудармасында мәтінді морфологиялық сегменттеудің моделі мен әдісі әзірлеген, нейронды машиналық аударма жүйесінде оқыту мен тестілеуге арналған қазақ тіліндегі параллельді мәтіндер корпусын құрастырып, әзірленген модель мен әдістің тиімділігін анықтау мақсатында эксперименттер жүргізген.</p>
4.	Ішкі бірлік принципі	<p>4.1 Диссертация өзектілігінің негіздемесі:</p> <ol style="list-style-type: none"> 1) негізделген; 2) жартылай негізделген; 3) негізделмеген. <p>4.2 Диссертация мазмұны диссертация тақырыбын айқындайды</p> <ol style="list-style-type: none"> 1) айқындайды; 2) жартылай айқындайды; 3) айқындамайды <p>4.3. Мақсаты мен міндеттері диссертация тақырыбына сәйкес келеді:</p> <ol style="list-style-type: none"> 1) сәйкес келеді; 2) жартылай сәйкес келеді; 3) сәйкес келмейді <p>4.4. Диссертацияның барлық бөлімдері мен құрылысы логикалық байланысқан:</p>	<p>Зерттелген жұмыстың өзектілігі толығымен негізделген. Қазіргі таңда қазақ тілінің нейронды машиналық аудармасында сегментациялау жасау өзекті мәселе. Жұмыста ұсынылып отырған ғылыми жаңалығы, құрылған модельдер мен алгоритмдер толығымен жоғары деңгейде сипатталынып көрсетілген.</p> <p>Диссертация мазмұны зерттеу жұмысы кіріспеден, төрт тараудан, қорытындыдан, пайдаланылған әдебиеттер мен қосымшалардан тұрады. Жұмыстың мазмұны берілген зерттеу тақырыбына сәйкес келеді және зерттеу міндеттерін толық көлемде айқындайды.</p> <p>Диссертациялық жұмыстың мақсаты мен міндеттері тақырыбы мен мазмұнына сәйкес келеді.</p> <p>Жұмыста лингвистикалық ерекшеліктері негізінде қазақ тілінің нейронды машиналық аудармасының сапасын жақсарту үшін модельдер, алгоритмдер мен бағдарламалық қамтамасыздандыру құру мақсаты қойылып, жалғаулардың толық жиынтығына (CSE – Complete Set of Endings) негізделген қазақ тілінің морфологиясының тілдік моделін жетілдіру (қазақ тілінің мүмкін жалғаулар тізімін кеңейту), қазақ тілі морфологиясының CSE-моделі негізінде морфологиялық сегменттеу моделі мен алгоритмін құру, нейрондық машиналық аударма платформасында айқындалған тапсырмалар бойынша эксперименттер жүргізу және жасау міндеттері зерттеудің теориялық маңызы мен практикалық мәнін нақтылай түседі.</p> <p>Диссертация кіріспеден, 4 бөлімнен, қорытындыдан, 2 қосымшадан тұрады. Жұмыстың барлық бөлімдері мен мазмұны және қол</p>

Handwritten signature

		<p>1) <u>толық байланысқан;</u> 2) жартылай байланысқан; 3) байланыс жоқ</p>	жеткізген нәтижелерінің ішкі бірлігі логикалық тұрғыдан толық байланысқан, бір-бірімен сабақтастық сақтай отырып тұжырымдалған.
		<p>4.5 Автор ұсынған жаңа шешімдер (қағидаттар, әдістер) дәлелденіп, бұрыннан белгілі шешімдермен салыстырылып бағаланған: <u>1) сыни талдау бар;</u> 2) талдау жартылай жүргізілген; 3) талдау өз пікірін емес, басқа авторлардың сілтемелеріне негізделген</p>	Автор диссертациялық жұмыстың екінші бөлімінде машиналық аударма саласындағы сегментацияның әдістеріне талдау жасаған, практика жүзінде машиналық аударма облысында де-факто болып қабылданған ВРЕ сегментациялау әдісімен тәжірибелер жүргізген, алайда сапа жағынан аталған әдіс қазақ тілі сияқты агглютинативті тілдер үшін қолайлы емес екендігін анықтап, қазақ тілінің күрделі морфологиясын ескеретін және сегменттерге дұрыс бөлетін жаңа әдісті ұсынып, құрылған модель негізінде тәжірибелер жүргізген.
5.	Ғылыми жаңашылдық принципі	<p>5.1 Ғылыми нәтижелер мен қағидаттар жаңа болып табыла ма? <u>1) толығымен жаңа;</u> 2) жартылай жаңа (25-75% жаңа болып табылады); 3) жаңа емес (25% кем жаңа болып табылады)</p>	<p>Қазақ тілінің жалғаулардың толық жиынтығына (CSE – Complete Set of Endings) негізделген қазақ тілінің морфологиясының тілдік моделі жетілдіріліп, толықтырылған. Дегенмен қазақ тілінің CSE-моделін құру барысында сын есім, үстеу үшін жалғауларды тудыру бөлігінде жалғау ретінде қарастырылған аффикстер (Кесте 46) жұрнақ болғандықтан, оны жалғау ретінде анықталмағаны дұрыс. Алғаш рет қазақ тілі үшін CSE-моделіне негізделген морфологиялық сегменттеудің жаңа әдісі мен алгоритмі әзірленіп, белгілі нейрондық модельдер негізінде нейрондық машиналық аударма платформасында эксперименттер жүргізілген. Алынған ғылыми нәтижелер толығымен жаңа болып табылады.</p> <p>Ғылыми жұмыстың нәтижелері мен тұжырымдары 21 ғылыми еңбек негізінде жарық көрді. Оның ішінде 3 мақала Scopus базасында индекстелген жоғары көрсеткішті журналдарда, бақылау комитеті салалық білім және ғылым саласында ұсынылған журналдардағы 4 мақала, Web Science and Scopus мәліметтер базасына енгізілген 6 конференция мақалалары, және халықаралық конференциялардың материалдарындағы 8 басылымда сыннан өткен.</p>
		<p>5.2 Диссертацияның қорытындылары жаңа болып табыла ма? <u>1) толығымен жаңа;</u> 2) жартылай жаңа (25-75% жаңа болып табылады);</p>	<p>Жұмыстың қорытындылары толығымен жаңа және ол жарияланған мақалалар мен авторлық құқықпен қорғалатын объектіге рәсімделген куәлікпен негізделеді. Диссертациялық жұмысты орындау кезінде келесідей жаңа қорытындылар алынған: 1. Қазақ тілінің CSE-моделі құрылып, толықтырылды.</p>

Ж.С.

		3) жаңа емес (25% кем жаңа болып табылады)	2. Алғаш рет қазақ тілі үшін CSE-моделіне негізделген морфологиялық сегментацияның жаңа әдісі әзірленді және ұсынылды. 3. Қазақ тілінің нейронды машиналық аудармасында морфологиялық сегменттелген мәтінмен жүйені оқыту тәжірибелері жасалған, сонымен қатар құрылған морфологиялық сегментациялау әдісі қазақ тілінің нейронды машиналық аудармасында оқыту кезінде тым үлкен сөздіктен туындайтын жады мәселесін шешкен.
		5.3 Техникалық, технологиялық, экономикалық немесе басқару шешімдері жаңа және негізделген бе? <u>1) толығымен жаңа;</u> 2) жартылай жаңа (25-75% жаңа болып табылады); 3) жаңа емес (25% кем жаңа болып табылады)	Диссертациялық жұмыстың техникалық және технологиялық, шешімдері толығымен жаңа және негізделген. Интернеттің қолжетімділігі арқасында қолданушылардың машиналық аударманы қолдануға сұранысы арттыруда. Жұмыста алғаш рет қазақ тілінің CSE-моделіне негізделген морфологиялық сегментациялаудың жаңа әдісі және алгоритмінің программалық жүзеге асырылуы қазақ тіліндегі мәтіндер үшін нейрондық машиналық аударма нәтижесін жақсартатыны сөзсіз.
6.	Негізгі қорытындылардың негізділігі	Барлық қорытындылар ғылыми тұрғыдан қарағанда ауқымды дәлелдемелерде <u>негізделген/негізделмеген</u> (qualitative research және өнертану және гуманитарлық бағыттары бойынша)	Ғылыми тұрғыдан толық дәлелденген. Ұсынылған әдістің тиімділігі нейронды машиналық аудармада қолданылатын белгілі LSTM seq2seq нейрондық модельдердің негізінде экспериментальді түрде негізделген және тексерілген. Қазақ тілінің CSE-моделі құрылып, аталған модель негізінде стемминг, алғаш морфологиялық сегментацияның әдісі құрылды, құрылған әдістің тиімділігі ауқымды дәлелдемелермен негізделген. Нәтижесінде ұсынылған CSE негізіндегі сегменттеу байттық жұпты кодтауға (BPE) негізделген сегменттеумен салыстырғанда BLEU баллын орта есеппен қазақ-ағылшын және ағылшын-қазақ жұптары үшін 0,5 және 0,2 пунктке артқанын және нейронды қазақ тілінің машиналық аудармасын оқыту кезінде жады көлемін екі жарым есе кішірейгенін байқауға болады.
7.	Қорғауға шығарылған негізгі қағидаттар	Әр қағидат бойынша келесі сұрақтарға жауап беру қажет: 7.1 Қағидат дәлелденді ме? <u>1) дәлелденді;</u>	7.1 Диссертанттың жұмысы бойынша қорғауға шығарылатын негізгі келесі қағидаттары дәлелденді: қазақ тіліндегі сөздерді морфологиялық сегменттеудің жаңа моделі мен алгоритмі, қазақ тіліндегі сөздерді сегменттеудің ұсынылған моделі мен

Б.А.

		<p>2) шамамен дәлелденді; 3) шамамен дәлелденбеді; 4) дәлелденбеді 7.2 Тривиалды ма? 1) ия; <u>2) жоқ</u> 7.3 Жаңа ма? 1) ия; 2) жоқ 7.4 Қолдану деңгейі: 1) тар; 2) орташа; <u>3) кең</u> 7.5 Мақалада дәлелденген бе? 1) ия; 2) жоқ</p>	<p>алгоритмінің тиімділігін растайтын қазақ тілінің нейронды машиналық аудармасы бойынша тәжірибелердің нәтижелері. Жұмыста шығарылған қағидаттар толығымен дәлелденген. Қорғауға шығарылған негізгі тұжырымдар тривиальді емес (7.2), жаңа (7.3) және қолданудың кең (7.4) деңгейіне ие, мақалаларда дәлелденген (7.5). 7.4 Диссертациялық жұмыста ұсынылған CSE-моделінің қолдану деңгейі кең. Құрылған модельді негізге ала отырып, басқа түркі тілдер тобындағы тілдерге қолдануға болады. Аталған модель негізінде қазақ тілі үшін стемминг, морфологиялық сегментация есептері шешілуде. 7.5 Зерттеу жұмыстарының нәтижелері 21 мақала түріндегі жарияланымдармен негізделген. Оның 9-ы Web of Science және Scopus базаларында индекстелген.</p>
8.	Дәйектілік принципі Дереккөздер мен ұсынылған ақпараттың дәйектілігі	<p>8.1 Әдістеменің таңдауы - негізделген немесе әдіснама нақты жазылған 1) ия; 2) жоқ</p> <p>8.2 Диссертация жұмысының нәтижелері компьютерлік технологияларды қолдану арқылы ғылыми зерттеулердің қазіргі заманғы әдістері мен деректерді өңдеу және интерпретациялау әдістемелерін пайдалана отырып алынған: 1) ия; 2) жоқ</p> <p>8.3 Теориялық қорытындылар, модельдер, анықталған өзара байланыстар және заңдылықтар эксперименттік зерттеулермен дәлелденген және расталған (педагогикалық ғылымдар</p>	<p>Диссертациялық жұмыста қолданылған әдіснаманың таңдауы негізделген және әдіснама нақты жазылған. Зерттеу жұмысында комбинаторикалық тәсілдің көмегімен қазақ тілінің жалғауларының есептеуіш моделі әзірленіп, белгілі нейрондық модельдер көмегімен эксперимент жүргізілген.</p> <p>Зерттеу жұмысын орындау кезінде Python бағдарламалау тілі, Tensorflow кітапханасының нейронды машиналық аударманы оқытудың NMT жүйесі, машиналық аударманың сапасын бағалауда қолданылатын BLEU көрсеткіші қолданылып, маңыздылығы дәлелденді. Аударма жүйесін оқыту үшін параллельді корпус Қазақстан Республикасының ресми сайттарынан деректер жиналып, өңделген. Нейронды машиналық аударманы оқыту кезінде 109 772 параллельді қазақ-ағылшын және ағылшын-қазақ сөйлемі өңделген.</p> <p>Теориялық тұжырымдар мен қорытындылар, құрылған модельдер мен олардың өзара байланыстары эксперименттік зерттеулермен дәлелденген және расталған. Диссертациялық жұмысты орындау кезінде қазақ тілінің нейронды машиналық аудармасында сегменттеудің танымал әдісі byte-pair encoding (BPE)-пен</p>

Ж.Б.

		<p>бойынша даярлау бағыттары үшін нәтижелер педагогикалық эксперимент негізінде дәлелденеді):</p> <p>1) ия; 2) жоқ</p>	<p>салыстыра отырып тәжірибелер жасалынған. Алынған эксперименттік тәжірибелер нәтижесінде ізденуші жұмысында құрған CSE-негізіндегі морфологиялық сегментация машиналық аударма жүйесінде қолданылатын BLEU көрсеткішінде жақсартулар алған. Атап айтқанда, CSE-негізіндегі морфологиялық сегментация ВРЕ-ға қарағанда қазақ-ағылшын және ағылшын-қазақ нейронды машиналық адармасында +0,5 және +0,2 көрсеткішіне жақсарту көрсеткен, сонымен қатар құрылған морфологиялық сегментациялау моделі негізінде сөздіктің көлемін 2,5 есе қысқартқан, бұл өз кезегінде есептеу экспериментінің уақыты қысқартқанын және техникалық ресурс қажеттілігін азайтқанын көрсетті.</p>
		<p>8.4 Маңызды мәлімдемелер нақты және сенімді ғылыми әдебиеттерге сілтемелермен <u>расталған</u> / ішінара расталған / расталмаған</p>	<p>Маңызды мәлімдемелер ғылыми әдебиеттерге сілтемелермен расталған. Пайдаланған әдебиеттер тізімі зерттеу саласына сәйкес. Дегенмен табиғи тілді өңдеумен айналысатын отандық ғалымдардың еңбектері толық зерттелмеген. Мысалы, Шәріпбай А.Ә., Бекманова Г. Т., Есенбаев Ж. А. және басқа да ғалымдардың еңбектері зерттеліп, шолу жасалмаған.</p>
		<p>8.5 Пайдаланылған әдебиеттер тізімі әдеби шолуға жеткілікті/жеткіліксіз</p>	<p>Пайдаланылған әдебиеттер тізімі орындалған диссертациялық жұмыстың зерттеу саласын толық қамтиды.</p>
9	Практикалық құндылық принципі	<p>9.1 Диссертацияның теориялық маңызы бар:</p> <p>1) ия; 2) жоқ</p>	<p>Диссертациялық жұмыстың теориялық маңызы бар. Диссертациялық жұмыс табиғи тілдерді өңдеуде маңызды және өзекті есептердің бірі болып саналады. Зерттеу жұмысында алғаш рет машиналық аударма үшін күрделі морфологиясы бар қазақ тілі үшін морфологиялық сегментациялаудың жаңа модельдері ұсынылған.</p>
		<p>9.2 Диссертацияның практикалық маңызы бар және алынған нәтижелерді практикада қолдану мүмкіндігі жоғары:</p> <p>1) ия; 2) жоқ</p>	<p>Диссертациялық жұмыстың практикалық маңызы бар. Жұмыста орындау барысында құрылған CSE-моделі негізіндегі морфологиялық сегментацияның, стемминг модельдері мен алгоритмін табиғи тілдерді өңдеудің есептерінде практикалық қолдану өте маңызды болып саналады. Құрылған модельдерді ақпаратты іздеу және алу, морфологиялық талдау, машиналық аударманың әртүрлі есептерін шешуде қолдануға болады.</p>
		<p>9.3 Практикалық ұсыныстар жаңа болып табылады?</p>	<p>Қазақ тілінің лингвистикалық ерекшеліктерін ескере отырып, CSE-моделі құрылып, толықтырылған. Ұсынылған морфологиялық</p>

		<p>1) толығымен жаңа; 2) жартылай жаңа (25-75% жаңа болып табылады); 3) жаңа емес (25% кем жаңа болып табылады)</p>	<p>сегментациялаудың әдісін қазақ тілінің нейронды машиналық адармасында тәжірибе жүргізу арқылы практикалық қолданысы көрсетілген. Құрылған CSE-моделін басқа да түркі тілдес тілдер тобындағы басқа тілдерге қолданыла алуы құрылған модельдер мен алгоритмінің күрделі морфологиясы бар, аглютинативті тілдер үшін практикалық маңызы бар екендігін көрсетті. CSE-моделі негізіндегі морфологиялық сегментациялаудың қазақ тілінің нейронды машиналық аударма жүйелерінде қолданылуының практикалық жағынан мүмкіндігі жоғары.</p>
10.	Жазу және ресімдеу сапасы	<p>Академиялық жазу сапасы: 1) жоғары; 2) орташа; 3) орташадан төмен; 4) төмен.</p>	<p>«Нейронды машиналық аударма үшін қазақ тіліндегі мәтіндерді морфологиялық сегментациясының модельдері мен әдістерін құру және зерттеу» тақырыбындағы диссертациялық жұмысы диссертациялық жұмысқа қойылатын талаптарға сәйкес дайындалған. Диссертациялық жұмысты жазу және ресімдеу сапасы жоғары, ресімдеу құрылымы мен ережелері сақталған. Диссертациялық жұмыс мәтінінде орфографиялық қателер мен стилистикалық қателер кездеседі. Аталған ескертулер жұмыстың құндылығын төмендетпейді.</p>

Кәрібаева Айдана Сейілғазықызының «Нейронды машиналық аударма үшін қазақ тіліндегі мәтіндерді морфологиялық сегментациясының модельдері мен әдістерін құру және зерттеу» тақырыбындағы диссертациялық жұмысы «Ғылыми дәрежелерді беру ережесінің» талаптарына сәйкес келеді және оның авторы **Кәрібаева Айдана Сейілғазықызын «6D070300 – Ақпараттық жүйелер» мамандығы бойынша философия докторы (PhD) дәрежесіне лайық деп санаймын.**

Ресми рецензент:

**Л.Н. Гумилев атындағы Еуразия Ұлттық университеті,
электрондық оқытуды және құжат айналымын сүйемелдеу секторының менеджері,
философия докторы (PhD)**

Ергеш Б.Ж.

