**А.С.Асанғали ¹\*, У.Б.Саурбек ¹, А.Қ.Калиева¹**

¹ Әл-Фараби атындағы Қазақ Ұлттық университеті

\*(e-mail: Aigerima\_brendly@mail.ru)

**Қазақ тілінің нейрондық машина аудармасы үшін белгісіз сөздердің аудармасын зерттеу.**

**Аннотация**

*Негізгі мәселе:* Қазақ тіліндегі сөздіктердің аздығы мен аудармалардың қазақ тілінде тура аудамауында. Яғни біз осы мақалада көрсетілген кемшіліктерді шешу мақсатында атқарылған жұмыстарды ашып көрсетеміз.

*Мақсаты:* нейрон тіліндегі машиналық аударма мәтіндерін танитын бағдарламалық қосымша құру. Нейрондық машина желісінде оқылымға жіберу мақсатында нәтиже алу.

*Әдістері:* берілген мәтіндерді тану үшін сипаттамалы әдіс, анализ, синтез жасаудың әдіс-тәсілдері қолданылады.

*Нәтижелері мен олардың маңыздылығы:* жасалған жүйе бойынша кез-келген графикалық мәтіндерді танып, жұмыс жасауды жеңілдету.Нейрон машина желісінде қазақ тіліндегі сөздік құрастыра отыра, синонимдес сөздер мен қазақ тілінің жұрнақ, жалғауларымен жұмыс жасаймыз. Бұл жоюамның маңыздылығы қазіргі қоғамға сай жаңа қазақ тіліндегі сөздік құра отыра, осы уақытқа дейінгі барлық технология қауымында еңбек етіп жүрген тұлғалардың зерттеуінде жүрген нейрон желісінде оқылым жүргізіп, оның алдағы уақытта үлкен нәжилелі болатынына үлесімді қоса аламын. Болашақта нейрондық машина желісінде қатесіз қазақ тіліндегі сөздікті ала алатындай кең ауқымды онлайн түрдегі аудармаларды қолдана алады.

**Кіріспе**

Neural Machine Translation (NMT) - бұл машиналық аудармаға жаңа тәсіл, ол дәстүрлі тәсілдермен салыстыруға болатын перспективалы нәтижелер көрсетті. Кәдімгі NMT жүйелеріндегі маңызды кемшілік олардың өте сирек кездесетін сөздерді дұрыс аудара алмауы болып табылады: NMT-дің сөздік қордан тыс (OOV) сөзді білдіретін бір реттік символы бар салыстырмалы түрде аз лексикасы бар. Бұл жұмыста біз осы мәселені шешудің тиімді әдісін ұсынамыз және енгіземіз. Біз NMT жүйесін сөздерді туралау алгоритмімен көбейтетін мәліметтерге үйретеміз, бұл NMT жүйесін шығаруға мүмкіндік береді, мақсатты сөйлемдегі әрбір OOV сөзі үшін, бастапқы сөйлемдегі сәйкес сөздің орнын. Бұл ақпарат кейінірек сөздіктің көмегімен әрбір OOV сөзін аударатын өңдеуден кейінгі кезеңде қолданылады. WMT’14-тен французшаға французшаға аудару бойынша біздің тәжірибелеріміз бұл әдісті қолданбайтын NMT эквивалентті жүйесінен 2,8 BLEU нүктесіне дейін айтарлықтай жақсартуды қамтамасыз ететіндігін көрсетті. 37.5 BLEU ұпайымен біздің NMT жүйесі WMT’14 жарыс тапсырмасында ең жақсы нәтижеден бірінші болып озды.

**Материалдар мен әдістері**

Біз кез-келген НМА жүйесіне оңай қолдануға болатын үш аннотация стратегиясын ұсынамыз (Калчбреннер және Блунсом, 2013; Суцкевер және басқалар, 2014; Чо және басқалар, 2014). Біз НМА жүйесін қара жәшік ретінде қарастырамыз және оны төмендегі модельдердің бірінде жазылған корпуста жаттықтырамыз. Біріншіден, түзетулер бақыланбайтын туралаумен жасалады. Әрі қарай, өңдеуден кейінгі сатыдағы сөздік аудармалар үшін қолданылатын сөздік сөздікті құру үшін туралау сілтемелерін қолданамыз. Егер сөз біздің сөздікте болмаса, жеке басын куәландыратын аударманы қолданамыз[1].

Бұл тәсілде біз әртүрлі белгісіз сөздерді қайнар көзде және мақсатты тілде көрсету үшін бірнеше таңбалауыштарды енгіземіз, тек бір белгі таңбалауышын қолданудан гөрі. Біз белгісіз сөздерді бірдей таңбалауыштарды қайталай отырып, бастапқы сөйлемде лексикадан тыс сөздерін unk1, unk2, unk3 деп белгілейміз. Мақсатты тілде белгісіз сөздердің аннотациясы біршама нақтырақ: (а) белгісіз бастапқы сөзге тураланған әр белгісіз мақсатты сөзге сол белгісіз таңбалауыш тағайындалады:

en: The unk portico in unk ...

fr: Le p0 unkp−1 unkp1 de p∅ unkp−1 ... (1)

Позициялық барлық модель – PosAll(Positional All Model) моделінің мысалы. Әр сөзден кейін pd қатысты позициялық таңбалар немесе p∅. таңбалары бар. «Көшіру» моделі және (b) тегістелмеген немесе белгілі сөзбен тураланған белгісіз мақсатты сөз арнайы unk∅ қолданады. Мысал үшін (1)-ді қараңыз. Бұл аңғартпа бізге кез келген бос емес белгісін аударуға мүмкіндік береді[1].

Көшірілетін модель бастапқы сөйлемдегі белгілі сөздерге, мысалы, «portico» және «portique» сөздерінің жұптары сияқты белгілі мысалға келтірілген белгісіз мақсатты сөздерді аудару мүмкіндігімен шектеледі. Бұрынғы сөз бастауыш сөйлемде белгілі; ал соңғысы жоқ, сондықтан оны unk∅ деп белгілейді. Бұл көбінесе модельдердің бастапқы лексикасы мақсатты лексикадан әлдеқайда үлкен болғандықтан орын алады, өйткені үлкен бастапқы лексика арзан. Бұл шектеу бізді бастапқы уақыттағы және мақсатты сөйлемдер арасындағы толық түзетулерден тұратын аңдатпа үлгісін жасауға түрткі болды, өйткені жаттығулар уақытында толық түзетулер алуға болады.

Атап айтқанда, біз тек бір әмбебап марка таңбалауышын қолдануға ораламыз.

Алайда, мақсатты жағына, біз әрбір сөзден кейін pd позициялық белгіні қоямыз. Мұнда d салыстырмалы позицияны көрсетеді (d = −7,...,−1,0,1,... ,7) j нүктесінде мақсатты сөздің бастапқы мәнге i = j − d позициясындағы сөзге тураланғанын айқындау. Тым алыс орналасқан тураланған сөздер теңестірілмеген деп саналады және null token pn-мен суреттелген теңдестірілмеген сөздер [2].

PosAll моделінің басты кемшілігі - бұл мақсатты сөйлемнің ұзақтығын екі есеге көбейту. Бұл оқуды қиындатады және параметрді екі есе көбейту жылдамдығын баяулатады. Алайда, біздің өңдеуден кейінгі қадамымыз тек белгісіз сөздерді туралауға қатысты болатындығын ескерсек, тек белгісіз сөздерге түсініктеме беру орынды болады. Бұл белгісіз таңбалауыштарды (d үшін −7, ..., 7 немесе ∅ үшін) бір уақытта (а) сөздің белгісіз екендігін және (b) оның салыстырмалы позициясын d үшін пайдаланатын позициялық белгісіз модельді ынталандырады. PosAll моделі сияқты, біз белгілеуді белгілемейтін белгісіз мақсатты сөздер үшін қолданамыз. Біз әмбебап кодты барлық белгісіз белгілерге қолданамыз [3]. Аннотацияланған мысал

en: The unk portico in unk ...

fr: Le unkpos1 unkpos−1 de unkpos1 ...

Позициялық белгісіз модель - PosUnk моделінің мысалы: unkposd таңбалауыштарымен тек тураланған белгісіз сөздер жазылады. Оның баяу жылдамдығына қарамастан, PosAll моделі жақсы туралауды үйренуі мүмкін, өйткені ол көптеген сөздер мен олардың туралануы бойынша оқытылады.

**Нәтижелер**

Word2vec әдісі барысында белгісіз сөздерді шешу мақсатында алдымен корпус жинау жұмысын жасалынды. Яғни әртүрлі сала мамандық бойынша мәтіндерді жеке жеке сөйлем ретінде корпусқа жинақтадым. Қазіргі уақытта жиналынған корпустан әрине оң нәтижеге қол жеткізу өте күрделі сол себепті де диссертациялық жұмысымның соңына дейін BLUE метрикасыны бойынша жобамның шешімі табылғанша корпус жинау қажет болады. Sublimе text build приложениесін қолдана отырып, корпус жинактадым. Қазіргі уақытта 1 млн 15 мың сөйлем жинақтай алдым. Жиналынған корпустарды Word2vec gensim модулін пайдала отырып сөздікке қосып отырмын.

Gensim - бұл тақырыптық модельдеуге арналған табиғи тілді өңдеу кітапханасы. Оның көмегімен мәтіндерді өңдеуге, векторлық сөз модельдерімен жұмыс жасауға (мысалы, Word2Vec, FastText және т.б.) және тақырыптық мәтіндік модельдерді құруға болады.

Тақырыптық модельдеу дегеніміз - өңделген мәтін берілген негізгі тақырыптарды бөліп алу әдісі. Gensim пакеті LDA және LSI тақырыптарын модельдеудің негізгі алгоритмдерін жүзеге асырады

Gensim-тің тағы бір маңызды артықшылығы: ол бүкіл мәтінді жадқа жүктемей, үлкен мәтіндік файлдарды өңдеуге мүмкіндік береді.

Мәтіндік құжаттармен жұмыс істеу үшін Gensim сөздердінемесе сөйлемдердің бірегей идентификаторларға айналдыруды талап етеді. Ол үшін Gensim-де әр сөзді ерекше идентификатормен салыстыратын сөздік объектісін құру керек.

Ал сөздіктерді құрастыру үшін корпуста жазылынған текстерді corpora.Dictionary() обьектісі арқылы айналыдырып алдым. Сөздік объектісі әдетте сөздер пакеті деп аталатын нәрсені жасау үшін қолданылады. Gensim мамандандырылған осы лексика мен «сөздер пакеті» (Корпус) тақырыптық модельдеуге және басқа модельдерге кіріс ретінде пайдаланылады.

Жиналынған корпустарым келесідей түрде:

*Философиялық сутралар және оған түсініктемелер дәуірі.*

*Осы уақытқа дейінгі жиналған материалдарды қысқарту және жинақтау уақыты болды.*

*Сутралар қысқа әрі нұсқа текстер болғандықтан оларды қосымша түсіндіру қажет болды.*

*Үнді философиясының дәстүрінде бәрі мойындайтын және пайдаланатын іргелі ұғымдар бар.*

*Ең алдымен, бұл карма ұғымы.*

*Карма – жазмыш заңдылығы, тағдыр адамға істеген әрекетінің табиғатына қарай несібе сыйлайды.*

*Егер адам өзінің эгоистік пиғылдағы әрекетінен тазарып, тек жоғары құндылықтарға ұмтылып отырса, онда ол карманың кәріне ұшырамайды.*

*Карма ұғымы сансара туралы іліммен тікелей байланысты.*

*Сансара дегеніміз – жанның жай таппай жиһан кезуі, бір денеден екінші денеге ұдайы ауысып отыруы туралы ілім.*

*Қайтадан туу қатары үзіліп, сансарадан азат болуды мокша деп атайды.*

*Ведалар – бұл магиялық дұғалар мен салтанатты әуендердің, қасиетті әндер мен құрбан шалу формаларының жиынтығы.*

*................................................................................................................*

Осы жиналынған корпустарды Word2vec обученияға жіберу арқылы оны векторларға айналдырып отырмын.

*6af4 2490 97a6 3370 8537 a14b bb79 fa11*

*4c57 3dd6 efa8 0305 6a8e 8d02 c7c5 3605*

*ea94 b1b2 31ab 5fba 2784 b4df de72 60d8*

*8610 1b71 ec61 cdcd 79ec 432c 8646 50f6*

*cfb2 34fc 8dc1 0a54 8833 427c de46 89b9*

*c9f8 9400 ff8f a855 0179 5712 490b 06b5*

*eb66 5130 1cef dfa8 d9db 6142 d0b6 02c6*

*00e7 5489 193e 8d6e 4ffe 30e3 e4d0 fe48*

*fe3a 9877 8c15 ee0b 97e6 0f98 97e7 6c62*

*...................................................................*

Әрбір жиналған корпусқа жинаған 10000 мәтінімнен кейін обученияға жіберіп, есептеу үстіндеміз. Осы алынған векторлар арқылы нейронный машинаға жіберу арқылы BLUE метрикасы бойынша белгісіз сөздерді шешу жолы азайдыма пайызы соңы есептейтін боламын. Қазіргі уақытта корпус мөлшерінің аздық етуіне байланысты менің белгісіз сөздерімнің көрсеткіші жоғары пайызды құрамайды.

Қазақ тіліндегі синонимдес жалғауларды жеке қарастырып, оларды gensim модулінде есептеп обученияға жіберу арқылы векторлық түрде сақтап алдым. Себебі жалғаулардың сөйлемдерде өте көп кездесуіне қарай алдағы уақытта жалғаулар кездесетін болса косинус векторларын сәйкестігіне қарай нейронды машинада автоматты түрде жалғауға сәйкес сөзді есептеу беру есебінде. Жалғауларды векторлық түрі:

*b77a 3c7b ad52 1761 146e c1e0 1d80 54aa*

*0eb7 2cde 80b2 de80 cb47 4059 8e80 cb8d*

*a02d 0da8 326b 6bea 6dce a943 dd99 386b*

*36eb 1d63 ac75 8eb1 d63a c758 eb1d 63ac*

*7506 4b3a c758 a240 0ea8 528c e726 732a*

*6fd3 c0c9 ad94 728e 51ca 6dca 3947 28cc*

*................................................................*

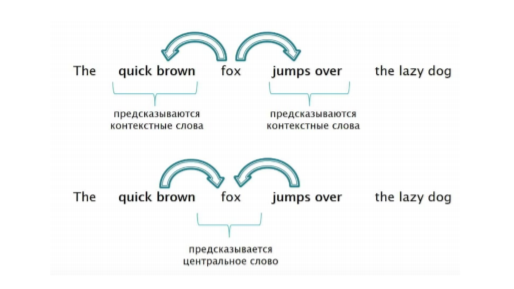
**Талқылау**

Word2vec әдісі - табиғи тілдегі сөздердің векторлық көріністерін алуға арналған жасанды нейрондық желі модельдерінің жиынтығының жалпы атауы. Табиғи тілдердің семантикасын дистрибутивтік семантика, машиналық оқыту және сөздерді векторлық бейнелеу негізінде талдау үшін қолданылады. «Word2vec» деп аталатын бағдарламалық жасақтама Google зерттеушілерінің тобымен 2013 жылы жасалған. Векторлық семантикалық модельдер жасауға арналған құралдар бұрын болған, бірақ word2vec алғашқы танымал іске асыруға айналды: ең алдымен, қолданудың қарапайымдылығы, бастапқы коды және жылдамдығы бойынша алдыңғы қатарлы тілдер санатында[4]. Қазіргі таңда Word2vec әдісімен өте танымал компаниялар зерттеу барысында сонымен қатар жаңаша қырларын есептеуде.

Бағдарлама келесідей жұмыс істейді: word2vec кіріс ретінде үлкен мәтіндік корпусты алады және әр сөзге векторды тағайындайды, шығудағы сөздердің координаталарын береді. Біріншіден, ол корпус сөздігін жасайды, содан кейін сөздердің векторлық көрінісін есептейді, кіріс мәтіндеріне «үйренеді». Векторлық бейнелеу мәтінмәндік жақындыққа негізделген: мәтінде бірдей сөздердің жанында пайда болатын сөздер (және, демек, мағынасы ұқсас) векторларға жақын болады (косинустық қашықтықта). Сөздердің алынған векторлық көріністері табиғи тілді өңдеу және машиналық оқыту үшін қолданыла алады. Мысалы ретінде жиі кездесетін сөз тіркестеріне қарай автоматты түрде машина векторларын есептеп, тиімді, үйлесімді сөзді таңдайды.

Word2vec - сөз ендірмелерін жасау үшін қолданылатын байланысты модельдер тобы. Бұл модельдер - сөздердің лингвистикалық контексттерін қалпына келтіруге үйретілген, екі қабатты таяз жүйке желілері. Word2vec өзінің мәтіні ретінде мәтіннің үлкен корпусын қабылдайды және векторлық кеңістікті шығарады, әдетте бірнеше жүз өлшемді, корпустағы әрбір ерекше сөзге кеңістіктегі сәйкес вектор тағайындалады. Сөз векторлары векторлық кеңістікте орналасқан, корпуста ортақ контекстке ие сөздер кеңістікте бір-біріне жақын орналасады.

Word2vec сөздердің үлестірілген бейнесін жасау үшін екі модельдік архитектураның бірін қолдана алады: үздіксіз сөз қаптары (CBOW) немесе үздіксіз Skip-gram (скип-грамм) (1-сурет). Сөздің үздіксіз архитектурасында модель ағымдағы сөзді қоршаған контекстті сөздер терезесінен болжайды. Контексттік сөздердің орналасу реті болжамға әсер етпейді (сөз орамы). Үздік-грамдық архитектурада модель ағымдағы сөзді қолдана отырып, контекстті сөздердің қоршаған терезесін болжайды[5]. Скип-грамматикалық архитектура жақын контекстті сөздерге қарағанда жақын контекстті сөздерді салмақтайды. Авторлардың ескертуіне сәйкес CBOW жылдамырақ, ал скип-грамм баяу, бірақ сирек кездесетін сөздер үшін жақсы жұмыс істейді.

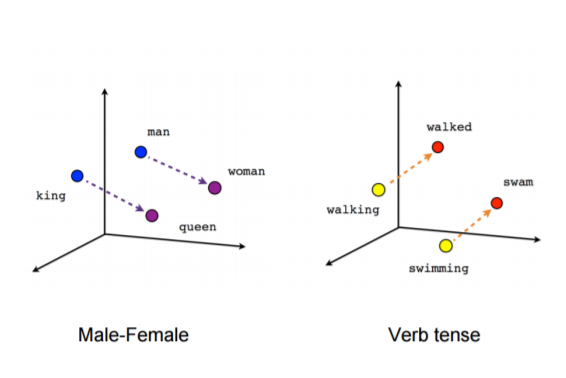


1-сурет.

Модель сөзі іргелес контекст берілген сөзді болжаса, грамматикалық скип моделі сөздің өзі берілген сөздің контекстін (немесе көршілерін) болжайды. Модель скип-граммен оқытылады, олар токендердің өтуіне мүмкіндік беретін n-грамм. Мәтіндік контекст грамматикалық скиптік жұптар жиынтығы арқылы ұсынылуы мүмкін (target\_word, context\_word), онда context\_word көршілес target\_word контекстінде пайда болады.

Word2vec шеңберінде табысты сөз енгізудің себептері нашар зерттелген. Голдберг пен Леви word2vec мақсаттық функциясы ұқсас контексте кездесетін сөздердің кірістірулерін (косинустың ұқсастығымен өлшенетін) ұқсас болуына әкеліп соқтырады және бұл Дж.Р.Ферттің үлестіру гипотезасына сәйкес келетіндігін ескертеді. Алайда, олар бұл түсіндірудің «өте қолмен» екенін атап өтіп, неғұрлым формальды түсініктеме берілетіндігін айтады.

Леви және басқалар. (2015) төменгі деңгейдегі тапсырмалардағы word2vec немесе соған ұқсас ендірулердің басым көпшілігінің өзі модельдердің нәтижесі емес, нақты гиперпараметрлерді таңдауы екенін көрсетеді. Осы гиперпараметрлерді анағұрлым «дәстүрлі» тәсілдерге ауыстыру ағынның төменгі жағындағы міндеттерде ұқсас нәтижелер береді. Жалпы word2vec әдісі жиі сөз тіркесі болып келетін сөздерді векторларын есептеу арқылы автоматты түрде машина өзі есептелініп таңдалынған сөзді шығарып береді (2-сурет).



2-сурет.

Бұл суреттен біз жиі тіркесте әйел –еркек , патша патшайым екендігін жақсы білеміз. Яғни патша қолданылған орында «р кезде патшайым қолданылады. Мысалы ағаш: алма ағашы, қайың ағашы қалыпты тіркестердің жиі қолданылуына қарай Word2vec әдісі де уйлесімділігіне қарай сөздерді орналастырып анықтайды.

Word2vec әдісі келесідей формула арқылы есептелінеді.

(2)

**Қорытынды**

Ортаға дәл келтіруге негізделген әдістеме қазіргі NMT жүйелерінің негізгі кемшіліктерінің бірін азайтуға, тіпті жеңуге болатындығын, яғни олардың сөздік құрамына кірмейтін сөздерді аудара алмайтындығын көрсетті. Біздің техниканың басты артықшылығы - бұл кез-келген NMT жүйесіне ғана емес, сонымен қатар терең LSTM моделіне ғана емес, сонымен қатар Суцкевер және т.б. (2014). Егер NMT жүйесі машиналық аударманың қазіргі заманғы көрсеткіштеріне қол жеткізсе, біздікі сияқты әдіс қажет.

Біз WMT’14 аударма тапсырмасында біздің техникамыз әртүрлі NMT жүйелерінде 2,8 BLEU нүктелеріне дейін дәйекті және түбегейлі жетілдірілетінін эмпирикалық түрде көрсеттік. Ең бастысы, 37,5 BLEU ұпайымен біз ең жақсы MT жүйесін WMT`14 конкурстық деректер жиынтығынан озған алғашқы NMT жүйесін құрамыз.

**Пайдаланылған әдебиеттер тізімі**

[Turian соавт., 2010] Джозеф Туриан, Лев Ратинов және Йошуа Бенгио. Сөздік бейнелер: жартылай бақыланатын оқытудың қарапайым және жалпы әдісі. Есептеу лингвистикасы қауымдастығының 48-ші жыл сайынғы жиналысының материалдарында, 384–394 беттер, 2010 ж.

[Калчбреннер және Блунсом, 2013] Н.Калчбреннер және П. Блунсом. 2013. Қайталанатын үздіксіз аударма модельдері. EMNLP-де.

[Суцкевер және басқалар, 2014] Суцкевер И., О. Виняльс және Q. В. Ле, 2014: Нейрондық желілермен жүйелі оқуға дайындық. Ақпаратты өңдеудің нейрондық жүйелеріндегі жетістіктер (NIPS).

[Чо және басқалар, 2014] Статистикалық машинаны аудару үшін rnn  кодтаушы-декодерін қолдана отырып, фразалық сөйлемдерді үйрену. EMNLP-де.

Интернет желісінен <https://vvs-info.ru/helpful_information/poleznaya-informatsiya/potentsialnye-potrebiteli/>

**А.С. Асангали ¹ \*, У.Б. Саурбек ¹, А.К. Калиева ¹**

¹ Казахский национальный университет имени аль-Фараби

**Исследование перевода неизвестных слов для нейронного машинного перевода с казахского языка.**

Основная проблема - отсутствие словарей на казахском языке и отсутствие прямого перевода на казахский язык. То есть опишем проделанную работу по устранению выявленных в этой статье недостатков.

Цель: создать программное приложение, распознающее тексты машинного перевода на нейронный язык. Получение результата для чтения в сети нейронных машин.

Методы: для распознавания заданных текстов используются описательные методы, методы анализа, синтеза.

Результаты и их важность: распознать и облегчить работу любого графического текста в разработанной системе.В нейронной машинной сети мы работаем с синонимичными словами и суффиксами казахского языка, составляя словарь на казахском языке. Важность этого решения в том, что я могу создать новый словарь казахского языка в соответствии с требованиями современного общества, прочитать нейронную сеть, которую до сих пор изучают люди, работающие во всех технологических сообществах, и внести свой вклад в ее дальнейшее развитие. В будущем нейронная машина сможет использовать широкий спектр онлайн-переводов, чтобы вы могли без ошибок получить словарь на казахском языке.

**A.S. Asangali ¹ \*, U.B. Saurbek ¹, A.K. Kalieva ¹**

¹ Al-Farabi Kazakh National University

**Study of the translation of unknown words for the neural machine translation of the Kazakh language.**

The main problem is the lack of dictionaries in the Kazakh language and the lack of direct translation into the Kazakh language. That is, we will describe the work done to eliminate the shortcomings identified in this article.

Purpose: to create a software application that recognizes machine translation texts into a neural language. Obtaining a result for reading in a network of neural machines.

Methods: descriptive methods, methods of analysis, synthesis are used to recognize given texts.

Results and their importance: to recognize and facilitate the work of any graphic text in the developed system. In the neural machine network, we work with synonymous words and suffixes of the Kazakh language, compiling a dictionary in the Kazakh language. The importance of this decision is that I can create a new dictionary of the Kazakh language in accordance with the requirements of modern society, read the neural network that people working in all technological communities are still studying, and contribute to its further development. In the future, the neural machine will be able to use a wide range of online translations so that you can get a dictionary in Kazakh without errors.