

Г. Қалман<sup>1</sup>, М.А.Самбетбаева<sup>1</sup>, Е.С.Жұмабай<sup>2</sup>, У.Б. Кусайнова<sup>3</sup>, М.К.Айткенова<sup>3</sup>

<sup>1</sup>Л.Н.Гумилев атындағы Евразиялық ұлттық университеті, Астана, Қазақстан

<sup>2</sup>Астана халықаралық университеті, Астана, Қазақстан

<sup>3</sup>Абай Мырзахметов атындағы Көкшетау университеті, Көкшетау, Қазақстан;

E-mail: guljamal14@gmail.com

## К-EN NEAREST NEIGHBOR АЛГОРИТМ НЕГІЗІНДЕ ҚАЗАҚ ТІЛІНДЕГІ АНАФОРАНЫ ШЕШУ

**Андатпа.** Көбінесе есімдіктің шешімі ретінде пайда болатын анафораны шешу - бұл дискурстағы бұрынғы немесе кейінгі элементтерге сілтемелерді шешу мәселесі. Анафордың шешімі - бұл мәтінді іздеу, мәтінді жалшылау, диалогтарды түсіндіру, ақпарат алу және т.б. сияқты белсенді зерттеу саласы. Анафораның ағылшын және басқа да еуропалық тілдердегі шешу мәселесі жақсы зерттелген. Қазақ тілінде әлі күнге дейін зерттеу жұмыстары жүргізілмеген.

Бұл мақалада машиналық оқыту тәсілдерін қолдана отырып, анафорды шешудің қазақ тілі жүйесі келтірілген. Ережеге негізделген әдіс анафора мен ықтимал антецедент жұбын жинау үшін қолданылады. K-En Nearest Neighbor алгоритмі анафора сілтеме жасаған ең ықтимал кандидатты таңдау үшін қолданылады.

**Түйінді сөздер.** Анафора шешімі, машиналық оқыту, k-жақын көрші, морфологиялық белгілер, POS.

### Кіріспе.

Анафораның ажыратымдылығы – сілтеме өрнектер немесе дискурс жиынтығынан нысанға сілтемелерді шешу процесі. Анафора әдетте дискурстағы алдыңғы элементті көрсететін есімдік немесе сілтеме сөз болып табылады. Бұл табиғи тілдегі қарым-қатынаста кең таралған құбылыс. Анафораны ажырату - анафораның антецедентін анықтау және кейіннен анафораны оның антецедентімен ауыстыру процесі. Анафоралық қатынастың бірінші мүшесі антецедент, екінші мүшесі анафор немесе анафоралық деп атайды. Анафораны шешу жүйесін енгізу мәтіннің көп мағыналылығын ашуға, сөйлемдерді түсінуге және контекстің сәйкестігін тексеруге мүмкіндік береді. Осылайша, мұндай шешу жүйесі табиғи тілді өңдеудің (NLP) көптеген қосымшаларында қажет болды, мысалы:

- 1) Білім қорын құру үшін ақпаратты алу қолданбалар
- 2) Құжаттық зерттеулерге арналған қолданбалар.

Анафораны шешу үнді-еуропалық тілдерде және басқа көптеген тілдерде көп зерттеулер жүргізілді. Екінші жағынан, қазақ тіліне бағытталған зерттеулер өте аз. Анафораны ажырату жұмыстарының жоқтығы және қазақ тіліне арналған ресурстардың (аннотация құралдары ретінде, тегтелген корпус ретінде) тапшылығы бізді қазақ мәтіндерінде анафораны шешудің жаңа тәсілін ұсынуға түрткі болды. Қазақ тілі үшін анафоралық қатынастар жиі өзгеріске түседі, бұл құбылыс анафораны шешу тапсырмасын қиындатады. Тапсырманың күрделілігіне байланысты біз өзіміздің зерттеу саламызды есімдік анафорамен шектедік. Біздің жұмысымыздың мақсаты - әрбір анафора үшін анықталған үміткерлер тізімінен ең жақсы антецедентті табу. Осы мақсатқа жету үшін біз келесі қадамдарды ұсындық:

- 1) Есімдік анафораны анықтау.
- 2) Сілтемелік емес есімдіктерді жою.

- 3) Әрбір анафораның ықтимал кандидаттық антецеденттерін анықтау.
- 4) Кандидаттар тізімін сүзгілеу.

Анафоралық қатынастың ең көп кездесетін түрі - есімдік анафорасы. Анафораның бұл түріне есімдіктің үшінші жақ түрі жатады, есімдіктердің ішінде ең көп анафоралық қызмет атқаратын жіктеу есімдігі мен сілтеу есімдігі болып табылады. Төмендегі мысалдардан жіктеу және сілтеу есімдігінің анафоралық қызметін көре аламыз.

Мысалы 1:

*Асқар* алматыдан келді, *ол* саған сайлықтар ала келіпті.

Бұл синтаксистік күрделі бірліктің бірінші сөйлемі мен екінші сөйлемін байланыстырып тұрған – бірінші сөйлемдегі *Асқар* сөзіне екінші сөйлемде *ол* жіктеу есімдігінің үшінші жақ түрі сілтеме жасау арқылы анафоралық қатынас болады. Сөйлемдегі *Асқар* антецедент, *ол* анафор болады.

Мысалы 2:

*Бәйтерек* – Астана қаласындағы сәулет құрылыс кешені, сәулет өнерінің бірегей туындысы. *Бұл*-Елорданың ең көрнекті ғимараттарының бірі.

Мысалдағы *бұл* сілтеу есімдігі бірінші сөйлемдегі “*Бәйтерек*” сөзіне сілтеме жасау арқылы анафоралық қатынас бола алады. Сөйлемдегі “*Бұл*” сөзі анафора ал “*Бәйтерек*” антецедент болады.

Анафоралық қатынасты шешу 1960 жылдардан бастап зерттеле бастады. Маңызды жұмыстарға Гоббс [1,2], Лаппин мен Лисс [3], Кеннеди мен Богураев [4], Митков [5,6,7] Тетрео[8] және Троуил [9] жұмыстарын атауға болады. Соңғы зерттеулерге шолу жасайтын болсақ.

Бұл жұмыста [10] биомедициналық мәтіндерде жиі кездесетін есімдік анафорасын UMLS онтологиясы мен SA/AO (субъект-әрекет/әрекет-нысан) модельдерін пайдалана отырып, анафора есімдігін шешу әдісі қолданылды, зерттеу нәтижесінде 92% алынды.

Келесі жұмыс орыс тіліндегі [11] анафораны шешу үшін аннотацияланған корпусты пайдаланды. Машиналық оқытуды пайдалана отырып корпустан “анафор антецедент” жұбын табу көзделді. Зерттеу нәтижесі “анафор антецедент” жұбын табу дәлдігі 75% құрады.

Қазақ тілі үшін анафораны шешу мәселесі енді ғана бастау алды, [12] мұнда қазақ тілінде референцияны шешуде алғаш рет жасалған зерттеу жұмысы, бұнда авторлар [15] ұсынылған тәсілге [13,14] біріктіру әрекеті жасалып көптілді жүйедегі сілтемелік қатынастарды шешу моделі ұсынылған.

Бұл мақалада қазақ тілі үшін анафораны шешу жүйесі қарастырылады. Бұл анафораны ажырату жүйесінде екі негізгі қадам бар. Бірінші қадам - *tengrinews* жаңалықтар топтамасы мен *Ғ.Мұстафинның* әңгімелерінің сілтеме өрнектері немесе дискурстары жиынтығынан есімдіктерді немесе зат есімді анықтау. Екіншіден, антецедент-анафора қатынасы анықталады, содан кейін есімдікке дұрыс және мүмкін антецедентті немесе үміткерді таңдайды.

Анафоралар мен сәйкес ықтимал антецеденттердің жинағы ереже негізінде және морфологиялық белгілермен анықталды. Ең ықтимал үміткер *k-Nearest Neighbor (k-NN)* әдісіне негізделген машиналық оқыту арқылы таңдалды.

Зерттеу нәтижелері және мәтіндегі есімдіктер бойынша сандық мәліметтер кесте (1,2 кесте) түрінде ұсынылды.

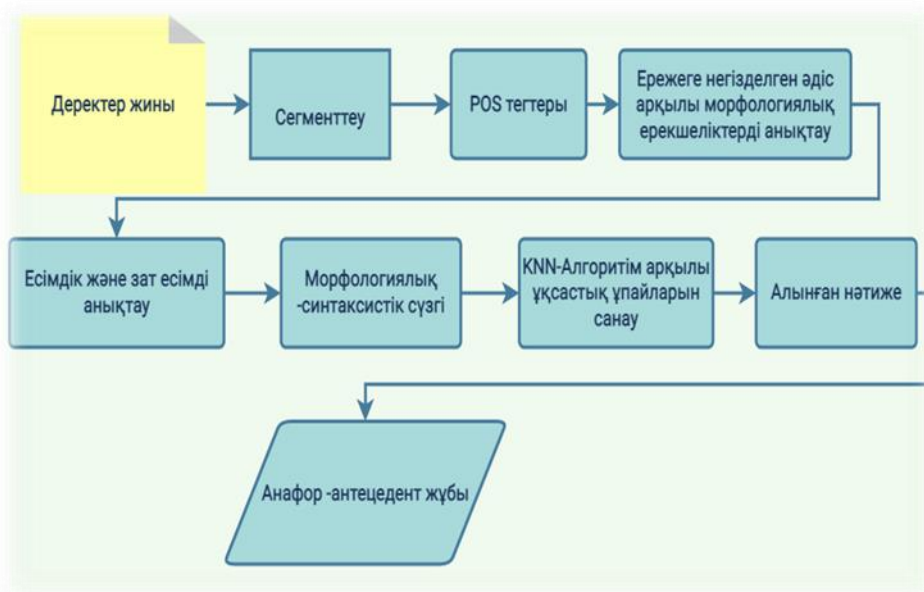
### **Материалдар мен тәсілдер.**

Бұл бөлімде қазақ тіліндегі анафораны ажырату жүйесінің архитектурасын ұсынады. Бұл жүйе ережеге негізделген сүзу модулін және машиналық оқыту модулін пайдаланады.

Қазақ тілінде сөйлемдегі сөздердің орын тәртібі, жалпы алғанда, тұрақты, орыс тілдегідей еркін емес, жалпы ереже бойынша баяндауыш сөйлемнің ең соңында, бастауыш одан бұрын, анықтауыш, пысықтауыш, толықтауыш өздері қатысты сөздерден бұрын тұрады.

Есімдіктер сөйлемде көбіне бастауыш, толықтауыш қызмет атқарады.

Бұл зерттеуде біз tengrinews жаңалықтар топтамасы мен F.Мұстафинның әңгімелерінен деректер жиынтығын аламыз. Бұл деректер жинағында жіктеу, сілтеу және өздік есімдіктерді шешіледі. Тиісті анафораны шешу жүйесі үшін субъект-объектіні сәйкестендіру қажет. Ұсынылған оқу деректер жинағы субъект, объект, сан, жанды немесе жансыз зат ретінде анықталады. POS тегтеріне арналған бұл жүйе қазақ тілі үшін Part-of-Speech Tagger үшін біз аннотацияланған корпусты пайдаланамыз, Есімдік пен зат есімді сөз тіркесі қазақ мәтініне арналған ережелерді қолдану арқылы таңдалады. Анықталған мүмкіндік мәні сөздеріндегі ең ықтимал үміткер k-Nearest Neighbors алгоритміне негізделген машиналық оқыту әдісін пайдалану арқылы таңдалады. Келесі ішкі бөлімдер әрбір құрамдас бөліктің негізгі қадамдарын, соның ішінде алдын ала өңдеу тапсырмасын егжей-тегжейлі талқылайды, морфологиялық сүзгілеу, ерекшелікті алу және жіктеу қадамы. 1-суретте көрсетілгендей, бұл жүйе екі негізгі бөліктен тұрады. Біріншіден, алдын ала өңдеу сатысында ережеге негізделген жүйені қолдана отырып, бастапқы деректер, морфологиялық белгілерді анықтау және есімдік пен зат есімді тіркестерді сәйкестендіру жиналады. Келесі қадамда есімдік пен зат есімнің әрбір жұбы k-Nearest Neighbor алгоритмі арқылы ұқсастық ұпайын алу үшін есептеледі. Содан кейін ол ең ұқсас болған анафораның алдыңғы жұбын шығарады.



1 сурет - Қазақ тіліндегі анафораны ажырату жүйесінің архитектурасы

Алдын ала өңдеу- бұл кезеңінде әртүрлі NLP әдістері қолданылады. Жүйедегі бұл модульдерге сөйлемдерді бөлу, POS-тегтеу, сөздерді сәйкестендіру, есімдік пен зат есімді анықтау кіреді. Біз мәтіннен POS тегтері күткен нәтижені жасау үшін Sentence Splitter қолданбасын қолданамыз. Sentence Splitter бізде берілген текстері токендерге бөлу операциясын орындайды және біз оны visual c++ ортасында жүзеге асырдық. Бұл есімдік шешу жүйесінде “||” белгісі сөйлемді бөлгіш ретінде қолданылады.

Біздің сөйлемді бөлгіштің нәтижесі - әрқайсысының бірегей сөйлем нөмірі бар сөйлемдер тізбегі. Сөйлем бөлігі (POS) – сөйлемдегі әрбір сөзге зат есім, етістік, сын есім

және т.б. сәйкес тілдік категорияларды тағайындау процесі. Сонымен қатар, әрбір сөзге берілетін морфологиялық белгілерді (сан, жол нөмірі, мағыналық түрі, есімдік түрі, грамматикалық рөлі) қамтамасыз етеді. Сан есімдер жеке (S), көпше (P) және ортақ (C) болып жіктеледі. Жол нөмірі анықталған іздеу шегіне қажет. Біз сондай-ақ негізінен семантикалық типті жанды немесе жансыз заттар ретінде анықтауымыз керек, өйткені біз негізінен анафора үміткерлері үшін жанды заттарды табамыз. Бұл зерттеу есімдіктердің жіктеу, сілтеу және өздік есімдік түрлерін шешеді. Зат есім де субъект, объект, жанды және жансыз сияқты грамматикалық рөлмен анықталады. Содан кейін біз есімдік анафора жұптары мен оның алдыңғы үміткерлеріне сәйкес келетін морфологиялық белгілері бар есімдіктер мен есімдіктерді қамтитын тізім жасаймыз.

Морфологиялық және синтаксистік сүзгі - Морфологиялық сүзгілеу қадамында біз «антецедент-анафор» жұптардың үлкен өлшемін (есімдік және антецедент) азайта аламыз. Жасалған тізімдегі әрбір анафора мен зат есім үшін жол нөмірі, семантикалық түрі, саны және жынысы сияқты келісім мүмкіндіктерін табамыз. Егер есімдік сөз тіркесі келісім мүмкіндіктеріне сәйкес келмесе, бұл сөз анафораның үміткерлер қатарынан алынады.

Анафора мен антецеденттің саны, септелуі, тегі және олардың арасындағы сөйлемдер саны, сонымен қатар анафора мен антецедент арасындағы зат есім сандары анықталады.

Синтаксистік сүзгі - бұл қадамында антецедент пен анафораның семантикалық рөлдері қарастырылады.

«антецедент-анафор» жұбын шығару - «антецедент-анафор» жұптарын шығарғанда ұпайлар бөлінеді және олардың ұпайлары бойынша сұрыптауға болады. Ең жоғары ұпайлары бар «антецедент-анафор» жұптар оқу деректер жинағына қосылады, ал қалғандарын елемеуге болады. «антецедент-анафор» жұптарын шығару табиғи тілді өңдеуде (NLP) өте маңызды тапсырма болып табылады. Машиналық оқыту (ML) әдістері негізінен деректердің анафор мен антецедент мән жұптарының мүмкіндік векторы ретінде ұсынылуын талап етті. Біз оларды ерекшеліктері бойынша үш санатқа топтастырдық: есімдік анафораның ерекшеліктері, алдыңғы үміткердің ерекшеліктері және екеуінің арасындағы байланыс ерекшеліктері. «антецедент-анафор» жұптары келесі 5 мүмкіндіктен тұрады. Олар:

F1– Candi-Line-No: Әрбір анафора үшін іздеу шекті жол нөмірлеріне зат есім-фразаның жол нөмірі қосылуы керек.

F2– Ana-Ante-Number: Оның мүмкін мәндері 0, 1, 2. Егер анафора мен алдыңғы сан сәйкес келсе 2 мәнді береді, әйтпесе 0 мәнін. Егер зат есім белгісіз болса, 1.

F3– Same-Sent: анафора мен алдыңғы үміткер бір сөйлемде болса, 0 мәнін қайтарады, кері жағдайда 1 болады.

F4 – Dist: анафора мен алдыңғы үміткер арасындағы сөйлем қашықтығы.

F5 – Freq: егер үміткер іздеу шегінде біреуден көп қайталанса, 0 мәнін қайтарады, кері жағдайда 1 болады.

K-Nearest Neighbor алгоритмы - K - Nearest Neighbor алгоритмі ( k-NN) – объектіні автоматты түрде жіктеуге немесе регрессияға арналған метрикалық алгоритм.

Ол деректерді іздеу, статистикалық үлгіні тану, суретті өңдеу сияқты көптеген қолданбалы салаларда қолданылады. Бұл объектіге жақын орналасқан оқу деректері негізінде объектілерді жіктеуді орындау әдісі.

Жіктеу әдісін пайдаланған жағдайда объект осы элементтің **k** көршілері арасында ең көп таралған классқа тағайындалады, оның класстары бұрыннан белгілі.

Дұрыс антецедент алу үшін **k** үшін оң бүтін санды көрсету керек [6]. Жақын көршінің **k** классификаторы, әдетте, тест үлгісі мен берілген оқу үлгілері арасындағы Евклид қашықтығына негізделген. Евклид нүктелері арасындағы қашықтық ( $a_1, a_2, \dots, a_p$ ) және ( $b_1, b_2, \dots, b_p$ ) ретінде анықталады.

$$d(a, b) = \sqrt{\sum_{p=1}^n (a_p - b_p)^2}, \quad (1)$$

мұндағы  $d(a, b)$  - тізімдегі есімдіктер мен зат есімдердің жұптары арасындағы қашықтық;

$n$  - алынған белгілердің саны. KNN алгоритмінің  $k$  мәні-таңдалған сөзге жақын коллекциядағы сөздердің қажетті санын көрсететін коэффициент.

Анафораны шешу алгоритмі - бұл алгоритм келесі ретпен орындалады:

1) “Антецедент анафора” жұбын табуда бірінші қадам антецедентке сәйкес анафораны табу, анафор табылмаған жағдайда нәтиже 0 қайтарады.

2) Анафора мен антецедент арасындағы анафор болып табылатын барлық зат есімдерді немесе есімдіктерді іздеу. Олардың саны мен септелуіне анафора сәйкес келуі керек. Іздеу аймағы алдын ала анықталған сөздер санымен шектеледі. Үміткер антецедентпен анафора арасындағы зат есімдер мен есімдіктерді табу, морфологиялық сүзгі жұмысында көрсетілгендей барлық келісім мүмкіндіктері бір біріне сәйкес келуі керек. Болмаған жағдайда нәтиже 0 қайтарады.

3) Үміткер антецедент пен анафораның семантикалық рөлдері бір біріне сәйкес келуі керек. Болмаған жағдайда нәтиже 0 қайтарады.

4) 2 және 3 қадамдар оң нәтиже берген жағдайда, дұрыс антецедент болу ықтималдығын есептеледі.

Төменде “Антецедент анафора” жұбын табу алгоритмі көрсетілген.

```
Input: P training examples( $a_i, x_j$ )  
number of iterations T  
init:  $\vec{q} \leftarrow \vec{0}$  ;  
fort  $\leftarrow 1$  to T ,  $i \leftarrow 1$  to P do  
 $\hat{x}_i \leftarrow \operatorname{argmax}_{c \in \operatorname{ant}(a_i)} F(c, a_i) \cdot \vec{q}$   
If  $\hat{x}_i \neq x_i$  then  
 $\vec{q} = \vec{q} + f(a_i, x_j)$   
End  
End  
Output: pair "antecedent-anaphora" ->  $\vec{q}$ 
```

2 сурет-Алгоритм: “Антецедент анафора” табу алгоритмі

Мұндағы  $F$  – мүмкіндікті шығару функциясы;

$a_i$  – анафоралық өрнек;

$x_j$  – шынайы антецедент;

$q$ - “Антецедент анафора” жұбы.

### Нәтижелер.

Бұл жүйеге арналған деректер tengrnews жаңалықтар топтамасы мен Ғ.Мұстафинның әңгімелерінен алынды. Қазақ тіліндегі есімдіктердің барлығы анафорлық қызмет атқармайды, есімдіктердің ішіндегі ең көп қолданылатын және анафорлық қызметке жиі түсетін жіктеу есімдігін және сілтеу, өздік есімдіктерін қарастырамыз.

Tengrnews жаңалықтар топтамасынан жалпы 65-ге жуық деректерді, Ғ.Мұстафинның әңгімелерінен 50 деректерді қарастырдық, онда жалпы анафораны шешуге арналған 376 есімдік бар. Осылайша біз 4000-ға жуық есімдіктермен (2 кестеде көрсетілген) анафоралық қатынас шешілді.

Анафора шешімі жіктеу мәселесі ретінде қарастырылды. Өнімділікті өлшеу үшін дәлдік (Precision), қайта шақыру (Recall) және F-өлшемі (F-measure) қолданылды. Өнімділікті өлшеу формуласы 2,3,4 формулаларда көрсетілді.

Зерттеу нәтижелері 1-кестеде көрсетілген.

$$Precision = \frac{TP}{(TP+FP)} \quad (2)$$

$$Recall = \frac{TP}{(TP+FN)} \quad (3)$$

$$F_1 = \frac{2*Recall*Precision}{(Recall+Precision)} \quad (4)$$

1 кесте - Нәтижелерді салыстыру

Деректер жины	<i>Precision</i>	<i>Recall</i>	$F_1$
tengrnews	0,84	0,56	0,67
Ғ.Мұстафинның әңгімелері	0,68	0,79	0,75

Келесі кестеде оқу жаттығу кезіндегі жалпы есімдіктердің сандық көрсеткіштері көрсетілген.

2 кесте – Есімдіктердің сандық көрсеткіштері

Есімдіктер	Саны	%
1 жақ	4560	17.85%
2 жақ	5645	24.72%
3 жақ	12766	60.38%
Жекеше және көпше түрлері		
Жекеше	7045	22.48%
Көпше	458	3.51%
Жалпы саны	13424	63.01%

Жұмыстың мақсаты мәтіндердегі “Антецедент анафора” жұбын шығару болған, зерттеу барысында біз қолданған оқу деректерінде жалпы 79 дәл нақты жұбын таба алдық.

### **Талқылау.**

Бұл мақалада біз анафоралық қатынасты шешуге жалпы теориялық зерттеулер жүргіздік, [1-8] жасалған зерттеулерден біз ағылшын және орыс тілдеріндегі анафораны шешуде ең көп қолданылған әдістерін және қазіргі кездегі қолданылатын әдістеріне өзара салыстыра отырып зерттеу жүргіздік.

Кезекті зерттеуімізде  $K$  - Nearest Neighbor алгоритмі ( $k$ -NN) қолдану және аннотацияланған тексттерді қолдану әдістеріне біріктіру әрекеті жасалды.

Ұсынылған алгоритм қазақ тіліндегі есімдік анафорасын шешу ережелерін кеңейту болып табылады. Алгоритм ерекшелігі морфологиялық және синтаксистік өңдеу сатысында мәтінді талдаудың есептеу және мұнда  $K$  - Nearest Neighbor алгоритмін ( $k$ -NN) сәтті қолданылуында. Анафоралық қатынастарды шешудің эксперименталды зерттеуі жүргізілді. Эксперименталды зерттеу *tengrinews* жаңалық топтамаларымен Ғ.Мұстафин әңгімелеріне жүргізілді, әрбір деректер жинағы үшін есімдіктердің үш түріне нақты сандық зерттеу жүргізіліп олардың сөйлемдегі анафоралық қызыметі жаңалықтар топтамасы үшін 67%, Ғ.Мұстафин әңгімелері үшін 75% құрады. Бұндай нәтиже дамушы тіл қазақ тілі үшін үлкен нәтиже болмақ.

### **Қорытынды.**

Бұл мақалада ережеге негізделген және  $K$ -ең жақын көршілес алгоритмді пайдалана отырып, қазақ анафорасының ажыратымдылық жүйесін ұсынды. Бұл тәсіл ережеге негізделген және машиналық тәсілдің күшті жақтарын пайдаланады. Бұл зерттеудің нәтижесі қазақ тілі үшін анафораны жақсырақ шешу құралын жасауға пайдалы болады.

Бұл жүйе есімдіктің үш түрін шеше алады: жіктеу есімдік, сілтеу есімдік, өздік есімдік.

Ережеге негізделген және машиналық оқыту тәсілінің үйлесіміне негізделген ұсынылған анафораны ажырату жүйесі қазақ тілі үшін толығымен тиімді және жақсы нәтижесін берді. Болашақтағы жұмыс анафораның басқа түрлеріне және сөйлемдегі катафоралық және корреференциялық сияқты сілтемелік қатынастарды шешуге және нейрондық әдіс немесе тереңдетіп оқыту әдістерін қолдана отырып шешуді арқылы гибридті тәсіл жасалады деп күтілуде.

## **ӘДЕБИЕТТЕР**

- [1] Hobbs J.R. Resolving pronoun references, 1978, *Lingua* 44, pp. 339-352.
- [2] Hobbs J.R. Pronoun resolution, Technical Report, Department of Computer Science, City College, City University of New York 76-1. 1976.
- [3] Lappin S, Leass H.J. An Algorithm for Pronominal Anaphoric Resolution, 1994, *Computational linguistics* 20, pp. 535-561
- [4] Kennedy C, Boguarev B. Anaphora for everyone: pronominal anaphora resolution without a parser. 16th International Conference on Computational Linguistics (COLING'96) Denmark, 1996, pp.113-118.
- [5] Mitkov R. Robust Pronoun Resolution with Limited Knowledge. 18th International Conference on Computational Linguistics, Canada, 1998, Volume 2, pp. 869–875
- [6] Mitkov R. Outstanding Issues in Anaphora Resolution. Lecture Notes in Computer Science 2004, pp.110-125.

[7] Mitkov R. Anaphora resolution: to what extent does it help NLP applications? *Anaphora: Analysis, Algorithms and Applications*. Springer Verlag, Berlin Heidelberg, 2007, pp.179-190.

[8] Tetreault J, Allen J. An empirical evaluation of pronoun resolution and clausal structure. 2003 International Symposium on Reference Resolution. Venice, Italy, 2003, pp. 1–8.

[9] Tetreault J, Allen J (2004) Dialogue Structure and Pronoun Resolution. *Lecture Notes in Computer Science* 1793, 2004, pp.515-525.

[10] Trouilleux F. Insertions et interprétations des expressions pronominales. *Actes de l'Atelier Chaînes de référence et résolveurs d'anaphores. TALN 2002 Nancy*. 2002, pp.1-11.

[11] Yu-Hsiang Lin and Tyne Liang. 2004. Pronominal and Sortal Anaphora Resolution for Biomedical Literature. In *Proceedings of the 16th Conference on Computational Linguistics and Speech Processing, The Association for Computational Linguistics and Chinese Language Processing (ACLCLP)*, Taipei, Taiwan. 2004, pp. 101–109

[11] Zhumabay, Y., Kalman, G., Sambetbayeva, M., Yerimbetova, A., Ayapbergenova, A. Bizhanova, A. Building a model for resolving referential relations in a multilingual system. *Eastern-European Journal of Enterprise Technologies*, 2022. V 2(2 (116), pp. 27–35.

[12] Garanina, N. O., Sidorova, E. A., Seryi, A. S. Multiagent Approach to Coreference Resolution Based on the Multifactor Similarity in Ontology Population. *Programming and Computer Software*, 2018, 44 (1), pp.23–34.

[13] Сидорова Е. А., Гаранина Н. О., Кононенко И. С. Многместные онтологические отношения в задаче разрешения. Шестнадцатая национальная конференция по искусственному интеллекту с международным участием КИИ-2018. 279-287С

[14] Kibrik, A. A. Reference and Working Memory. *Current Issues in Linguistic Theory*, 1999. pp.29.

## REFERENCES\*

[13] Sidorova E. A., Garanina N. O., Kononenko I. S. Mnogomestnye ontologicheskie otnosheniya v zadache razresheniya. Shestnadcataya nacional'naja konferenciya po iskusstvennomu intellektu s mezhdunarodnym uchastiem KII-2018. 279-287S

**Gulzhamal Kalman**, doctoral student, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan, guljamal14@gmail.com

**Madina Sambetbayeva**, PhD, docent, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan, madina\_jgtu@mail.ru

**Yerzhan Zhumabay**, doctoral student, Astana International University, Astana, Kazakhstan, erzhan\_93kz@list.ru

**Ulzhan Kussainova**, master, senior lecturer, Abay Myrzakhmetov Kokshetau University, Kokshetau, Kazakhstan, ulzhan-92-92@mail.ru

**Makhabat Aitkenova**, master, senior lecturer, Abay Myrzakhmetov Kokshetau University, Kokshetau, Kazakhstan, Mahabat\_89\_is@mail.ru

## SOLUTION OF ANAPHORA IN THE KAZAKH LANGUAGE BASED ON THE ALGORITHM K K-EN NEAREST NEIGHBOR

**Abstract.** Resolving anaphora, which often occurs as a resolution of a pronoun, is a matter of resolving references to earlier or later elements in the discourse. The solution of anaphora is searching the text, summarizing the text, explaining dialogues, getting information, etc. such an active field of research.

The resolution of anaphora in English and other European languages is well studied. Research works have not been carried out in the Kazakh language.

This article presents the Kazakh language system for solving anaphor using machine learning methods. A rule-based method is used to collect anaphora and possible antecedent pairs. The K-En Nearest Neighbor algorithm is used to select the most likely candidate referred by the anaphora

**Keywords.** Anaphora decision, machine learning, k-nearest neighbor, morphological features.

**Гулжамал Калман**, докторант, Евразийский национальный университет им. Л.Н. Гумилева, Астана, Казахстан, guljamal14@gmail.com

**Мадина Самбетбаева**, PhD, доцент, Евразийский национальный университет им. Л.Н. Гумилева, Астана, Казахстан, madina\_jgtu@mail.ru

**Ержан Жұмабай**, докторант, Астанинский международный университет, Астана, Казахстан, erzhan\_93kz@list.ru

**Улжан Кусаинова**, магистр, старший преподаватель, Кокшетауский университет имени Абая Мырзахметова, Кокшетау, Казахстан, ulzhan-92-92@mail.ru

**Махабат Айткенова**, магистр, старший преподаватель, Кокшетауский университет имени Абая Мырзахметова, Кокшетау, Казахстан, Mahabat\_89\_is@mail.ru

## РЕШЕНИЕ АНАФОРЫ В КАЗАХСКОМ ЯЗЫКЕ НА ОСНОВЕ АЛГОРИТМА К K-EN NEAREST NEIGHBOR

**Аннотация.** Разрешение анафоры, которая часто встречается как разрешение местоимения, заключается в разрешении ссылок на более ранние или более поздние элементы дискурса. Решением анафоры является поиск текста, резюмирование текста, объяснение диалогов, получение информации и т. д. такая активная область исследований.

Разрешение анафоры в английском и других европейских языках хорошо изучено. Исследовательские работы на казахском языке не проводились.

В данной статье представлена система казахского языка для решения анафоры с использованием методов машинного обучения. Метод, основанный на правилах, используется для сбора анафоры и возможных антецедентных пар. Алгоритм K-En Nearest Neighbor используется для выбора наиболее вероятного кандидата, на который ссылается анафора.

**Ключевые слова.** Решение анафоры, машинное обучение, k-ближайший сосед, морфологические признаки.

\*\*\*\*\*