ТЕХНИЧЕСКИЕ НАУКИ
И ТЕХНОЛОГИИ

АҚПАРАТТЫҚ ЖҮЙЕЛЕР
ИНФОРМАЦИОННЫЕ СИСТЕМЫ
INFORMATION SYSTEMS

DOI 10.51885/1561-4212_2024_1_117
МҒТАА 20.19.27

**Е.С. Жұмабай[1], Г. Қалман[2], Е.А. Сидорова[3], А.А. Дауренова[4]**
[1]Жұмабек Тәшенов атындағы жалпы орта мектебі, Аршалы, Қазақстан
*E-mail: erzhan_93kz@list.ru*
[2]Л.Н. Гумилев атындағы Еуразиялық ұлттық университеті, Астана, Қазақстан
*E-mail: guljamal14@gmail.com**
[3]А.П. Ершов атындағы информатика жүйелер институты, Новосибирск, Ресей.
*E-mail: e.sidorova3@g.nsu.ru*
[4]Абай Мырзахметов атындағы Көкшетау университеті, Көкшетау, Қазақстан.
*E-mail: seralieva_a_a@mail.ru*

# РЕФЕРЕНЦИЯЛЫҚ ҚАТЫНАСТЫ ШЕШУ МОДЕЛІН ӘЗІРЛЕУ

# РАЗРАБОТКА МОДЕЛИ РЕШЕНИЯ РЕФЕРЕНЦИАЛЬНЫХ ОТНОШЕНИЙ

# DEVELOPMENT OF A REFERENCE RELATIONSHIP SOLUTION MODEL

***Аңдатпа.*** *Дәстүрлі оқытуға негізделген кореференцияны шешу моделдері екі ескертпенің референтті немесе референтті емес екендігін анықтау үшін жұп үлгісін (анафор-антецедент жұбын) оқыту арқылы жұмыс істейді. Тұжырымдамалық тұрғыдан қарапайым және түсінуге оңай болғанымен, аталған жұптық модель лингвистикалық тұрғыдан өте күрделі. Осы күрделі процесті жақсарту мақсатында сілтеме жұбының моделін жақсартуға тырыстық, бірінші модель берілген кореференция үшін алдыңғы ескертулерді бағалау үшін рейтинг моделі (mention-ranking model), екінші модель: алдыңғы кластерлеудің берілген референтке сәйкес келетіндігін анықтау үшін объектіні еске түсіру моделі. Біз атап өтілген рейтинг моделі және объектіні еске түсіру моделінің жақсы жақтарын біріктіретін және осы екі модельге қарағанда теориялық жағынан тиімдірек болатын негізгі кореференциялық шешімге кластерлік тәсілді қолданамыз. Сонымен қатар, кластерлік тәсіл арқылы кореференция мен анафораны бірге шешу тәсілін де ұсынамыз. Моделді сынақтан өткізу үшін қазақ тіліндегі корпусты пайдаландық (https://qazcorpus.kz ), эксперименттік нәтижелерін өзара салыстырғанда кластерлік рейтинг моделі жақсы нәтиже көрсетті.*

***Түйін сөздер****: кореференция, кластерлеу, референция, анафора, бағалау моделі, жұптық модель.*

***Аннотация.*** *Традиционные модели решения кореференции, основанные на обучении, работают путем обучения модели пары (пары анафор-антецедент), чтобы определить, являются ли две заметки референтными или нет. Хотя это концептуально просто и легко понять, упомянутая парная модель очень сложна с лингвистической точки зрения. Чтобы улучшить этот сложный процесс, мы попытались улучшить модель эталонной пары, первая модель: используя модель оценки (mention-ranking model) для оценки предыдущих замечаний для данной анафоры, вторая модель: путем обучения модели отзыва объекта, чтобы определить, соответствует ли предыдущий кластер данной ссылке. Мы предлагаем кластерный подход к решению основной кореференции, который сочетает в себе лучшие аспекты отмеченной модели оценки и модели объектного воспоминания и теоретически более эффективен, чем эти две модели. Кроме того, мы также предлагаем способ совместного решения кореференции и*

*анафоры с помощью кластерного подхода. Для тестирования модели nur.kz мы использовали серию новостей, экспериментальные результаты показали высокую производительность инструментов кластерного рейтинга по сравнению с конкурирующими подходами.*

***Ключевые слова**: кореференция, кластеризация, референция, анафора, модель оценки, парная модель.*

***Abstract.** Traditional learning-based coreference resolution models work by learning a pair pattern (anaphore-antecedent pair) to determine whether the two notes are referent or not. Although conceptually simple and easy to understand, the named pair model is very complex linguistically. In order to improve this complex process, we tried to improve the link pair model, the first model: by using the mention-ranking model to evaluate previous mentions for a given anaphora, the second model: by learning the object recall model to determine whether the previous cluster corresponds to a given reference. We propose a cluster approach to the basic coreference solution, which combines the best aspects of the noted evaluation model and the object recall model and is theoretically more effective than these two models. In addition, we also offer a way to solve coreference and anaphora together through a cluster approach. To test the model nur.kz we used a news package, and the experimental results showed a higher performance of cluster rating tools compared to competing approaches.*

***Keywords**: coreference, clustering, reference, anaphora, mention-ranking model, entity-mention model.*

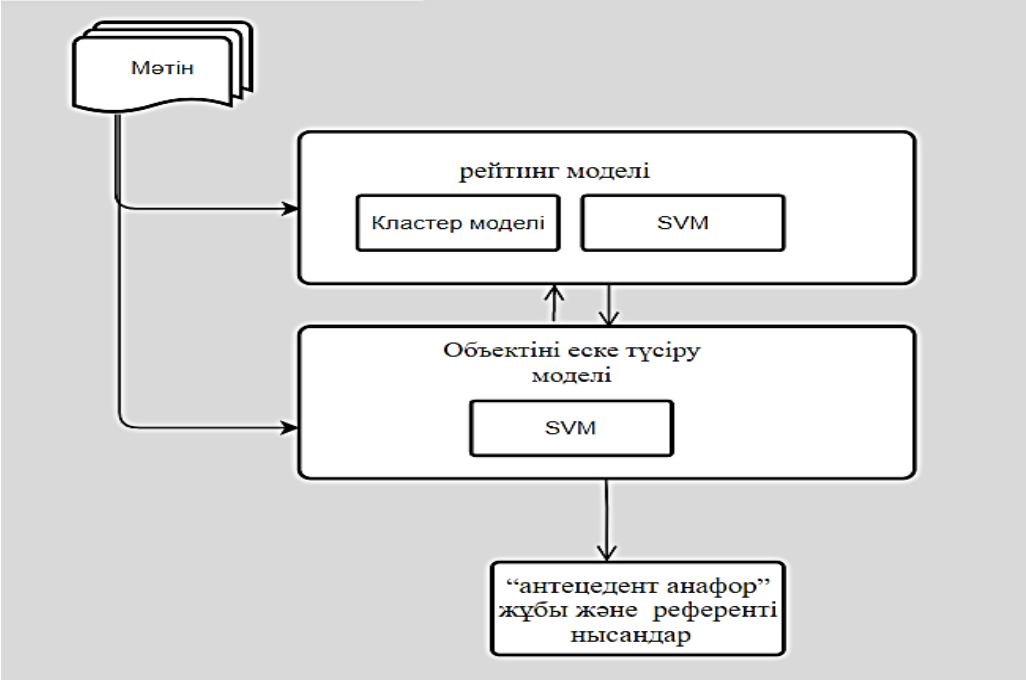
*Кіріспе*. Зат есім сөз тіркестеріндегі кореференцияны шешу – мәтіндегі немесе диалогтағы зат есім тіркесінің бір нақты объектіге немесе тұжырымдамаға қатысты екенін анықтау міндеті. Есептеу тұрғысынан алғанда, кореференцияны кластерлеу мәселесі болып табылады, оның мақсаты ескертулер жиынын негізгі сілтеме кластерлеріне бөлу болып табылады, мұнда әрбір кластер бір-бірімен байланысқан референттерді қамтиды. Біз бұл жұмысқа анафора терминін қосамыз. Анафора – бұл дискурста белгілі бір объектіге (немесе нысандарға) сілтеме жасау құралы болып табылады, сілтеме анафор деп, ал сілтеме жасайтын объект (немесе нысан) оның референті немесе антецеденті болып табылады.

Соңғы онжылдықта кореференцияны шешудегі зерттеулер эвристикаға негізделген тәсілдерден машиналық оқыту әдістеріне біртіндеп ауысты. Бұл ауысым табиғи тілді өңдеу дәуірінің (NLP) басталуымен және аннотацияланған корпустардың жалпыға қол жетімділегімен түсіндіруге болады. Кореференцияны шешуге арналған машиналық оқытудың ең ықпалды тәсілдерінің бірі – жіктеуге алгоритмдерін және лингвистикалық білімді пайдаланып шешу. Мұнда зат есім тіркесті кореференцияны шешуді жіктеу алгоритмі, лингвистикалық білімді және кластерлеу алгоритмі арасындағы өзара әрекеттесуді тиімді пайдалану арқылы кореференцияны шешу жүйелерін жақсартуға болатынын көрсетті [1]. Келесі жұмыста ең ықтимал антецедентті табу үшін кластерлеу алгоритмі және референт пен антецедент жағдайын сипаттайтын лингвистикалық білім базасы керектігі атап өтілді [2]. Атап айтқанда, кейбір модельдер мәтінде референт жұбының бірлескен референт екенін немесе жоқтығын анықтау үшін қолданылады. Алайда, осы моделдер жасаған жұптық классификациялар (қазіргі уақытта жұпты еске түсіретін модель ретінде белгілі) кореференция қатынасына тән транзитивтілік қасиетін қанағаттандырмауы мүмкін, өйткені модель (A, B) coreferent, (B, C) coreferent және (A,C) coreferent емес деп жіктей алады. Нәтижесінде жұптық жіктеу және осы сілтемелердің бөлінуін құру бойынша кластерлеу механизмі қажет болады. Кореференцияны шешуде қолданылған жұптық модельде айтарлықтай жақсы нәтижелер алдынды [3, 4], бірақ кейбір кемшіліктері болды, атап айтсақ, шешілетін референтке (бұдан әрі белсенді ескертпе) әрбір үміткер басқалардан тәуелсіз қарастырылатындықтан, бұл модель тек қана кандидат антецедентінің белсенді атауға қатысты қаншалықты жақсы екенін анықтайды, бірақ кандидаттың антецеденті басқа кандидаттарға қатысты қаншалықты жақсы екенін анықтай алмайды [5, 6]. Басқаша айтқанда, ол қай кандидаттың антецеденті

ең ықтимал деген сыни сұраққа жауап бере алмайды. Осындай кейбір моделдердің кемшіліктерін болдырмау мақсатында, осы жұмыста біз рейтинг моделін (mention-ranking model) және объектіні еске түсіру моделін (entity-mention model) қолданамыз.

Біз осы бағытта бұған дейін бірнеше зерттеулер жасаған едік, келесі жұмыста [7] машиналық оқыту негізінде қазақ тіліндегі есімдік анафорасын шешу алгоритмін ұсынып жақсы нәтиже алынған еді, одан кейінгі зерттеулер кореференцияны кластерлеу әдісімен[8] және k-nearest neighbor алгоритмын [9] пайдаланып кореференцияны шешу тәсілдері ұсынылды. Рефренцияны шешудегі қазіргі кезде көп қолданылатын әдістердің бірі корпустық жиынтықты пайдаланып көп тілді жүйеде референциялық қатынасты шешу моделі [10] жасалды. Келесі жұмыста [11] қазақ тіліндегі есімдік анафорасын шешудің SVM тірек векторлық машиналық оқыту алгоритмін құрдық. Жоғарыдағы аталаған жұмыстарда анафорамен кореференцияны шешу жұмыстары әр түрлі алгоритмдер мен моделдер көмегімен жеке-жеке зерттелген, ал біз осы жұмыс аясында анафора мен кореференцияны шешу мәселесін моделдердің артықшылықтарын ескере отырып біріктіру процесін жүзеге асырамыз. Осы мақсатқа жету үшін анафора кандидаттардың антецедентке жұбы болу ықтималдығын анықтайтын рейтинг моделін, кореференцияны шешуде референті болуын анықтайтын объектіні еске түсіру моделін (entity-mention model) қолданамыз.

Анафораны шешуде «антецедент анафор» жұбын табу болып табылады. Осы мақсатқа жету үшін біз рейтинг моделіне (mention-ranking model) кластерлік рейтинг моделін және SVM пакетінен SVMlight алгоритмін оқыту [12], ал кореференцияны шешуде объектіні еске түсіру моделін (entity-mention model) және SVM қолданамыз. Зерттеу жұмысында қолданылған моделдің жалпы жұмыс істеу принципі 1-суретте көрсетілген.




**1-сурет**. Референцяны шешу моделі

*Зерттеу әдістері*. Жұмыс барысында қойылған мақсаттарға байланысты векторлық машина алгоритмі, мәтінді сегментке бөлу, кандидат пен антецедентті лексикалық

граматикалық талдау және кластерлеу әдістері қолданылды.

*Референцияны шешу моделдері.* Бұл кезеңде біз кореференция мен анафораны шешудегі моделдерді қалай оқыту керектігін айтамыз. Моделді оқыту барысында біз төмендегі мәтін сегментін қолданамыз. Мәтінді сегменттеу сөйлемдегі референциялық қатынастарды тез табу үшін қолданылады.

|СОФІЯ КОВАЛЕВСКАЯ (1850 - 1891) |$^{1}_{1}$- Ресейдегі тұңғыш |әйел|$^{1}_{2}$ - |профессор|$^{1}_{3}$ және әлемдегі |тұңғыш әйел математик|$^{1}_{4}$. |Ол|$^{1}_{5}$ қатты дененің қозғалмайтын нүкте айналасында айналу есебінің шешілу мүмкіндігінің үшінші классикалық жағдайын ашты.

**2-сурет**. Мәтін сегменті

Сегменттегі әрбір m ескертпесі $[m]^{cid}_{mid}$ деп белгіленеді, мұндағы mid – сілтеме идентификаторы, ал cid-m кластер идентификаторы. Көріп отырғанымыздай, ескертулер бес жинаққа бөлінген, «София Ковалевская»**,** «әйел», «профессор», «тұңғыш әйел математик», «ол», бір кластерде, ал қалған ескертулердің әрқайсысы өз кластерінде.

Жоғарыда айтып өткендей, объектіні еске түсіру моделі белсенді ескертпе mk mj кандидатының алдыңғы қатарлы мәніне референті немесе референті емес екендігін шешетін жіктеуші болып табылады. Әрбір і данасы (mj, mk) mj және mk білдіреді. Біздің іске асыруымызда і данасы үшін 15 функциядан тұрды (1-кестеде көрсетілген).

**1-кесте**. Корференцияны шешуге арналған функциялар жиынтығы

| № | mj сипаттайтын мүмкіндіктер | «а» кандидаттың антецеденті |
|---|---|---|
| 1 | Pronoun_1 | егер mj есімдік болса Y; есімдік болмаса N |
| 2 | subject_1 | егер mj зат есім болса Y; зат есім болмаса N |
| 3 | Nested_1 | егер mj кірістірілген NP болса Y; кері жағдайда N |
| | Mk сипаттайтын мүмкіндіктер | |
| 4 | number 2 | жекеше немесе көпше түр |
| 5 | Pronoun_2 | егер mk есімдік болса Y; кері жағдайда N |
| 6 | Nested_2 | егер mk кірістірілген NP болса Y; кері жағдайда N |
| 7 | Pro typee_2 | номинативті жағдай mk, егер бұл есімдік болса; кері жағдайда NA |
| | mj, кандидаттың антецеденті және mk арасындағы байланысты сипаттайтын мүмкіндіктер | |
| 8 | head match | егер референттер бірдей есімдік болса C; басқа жағдайда I |
| 9 | str match | егер референттер бір жолда болса C; басқа жағдайда I |
| 10 | substr match | егер бір референт екіншісінің ішкі жолы болса C; басқа жағдайда I |
| 11 | pro str match | егер екі референт де есімдік болса және бір жолда болса C; басқа жағдайда I |
| 13 | pn str match | егер екі референт де жалқы есім болса және бір жолда болса C; басқа жағдайда I |
| 14 | number | егер референт саны сәйкес келсе C; басқа жағдайда I; |

| | | |
|---|---|---|
| | | егер бір немесе екі референтке арналған санды анықтау мүмкін болмаса NA |
| 15 | span | егер референттің ешқайсысы екіншісін қамтымаса C;<br>басқа жағдайда I |

Жіктеу моделін пайдаланған кезде, SVM оң нүктелерді теріс нүктелерден бөлетін гипер жазықтықты (яғни, сызықтық классификатор) оқытуды мақсат етеді. Максималды маржа гипер жазықтық w · x − b = 0 арқылы анықталады, мұндағы x – деректер нүктесін көрсететін мүмкіндік векторы, ал w (салмақ векторы) және b (скаляр), осы параметрлерді шешу үшін келесі оңтайландыру мәселелері қарастырылады. Шекараларды анықтау төмендегі формуламен анықталады.

$$arg\ min\frac{1}{2}||w||^2 \quad (1)$$

$$y_i(w, x_i - b) \geq 1, \quad 1 \leq i \leq n$$

мұндағы $y_i \in${+1, -1}-$x_i$-ші жаттығу нүктесінің $i$ класы. Осы оңтайландыру тапсырмасындағы әрбір $x_i$- деректер нүктесі үшін $x_i$ – дұрыс жіктелуіне кепілдік беретін дәл бір сызықтық шектеу бар екенін ескеру. Атап айтқанда, әрбір теңсіздіктің оң жағындағы 1 мәнін қолдану, әрбір $x_i$ мен гипержазықтық арасында белгілі бір қашықтықты (яғни, маржа) қамтамасыз етеді. Маржа салмақ векторының ұзындығына кері пропорционал екенін көрсетуге болады. Демек, салмақ векторының ұзындығын азайту маржаны ұлғайтумен тең. Алынған SVM классификаторы қатты маржа SVM ретінде белгілі: маржа «қатаң анықтау» керек, өйткені әрбір деректер нүктесі гипержазықтықтың дұрыс жағында орналасқан болуы керек.

$$arg\ min\frac{1}{2}||w||^2 + c\sum_i \zeta_i \quad (2)$$

$$y_i(w, x_i - b) \geq 1 - \zeta_i, \quad 1 \leq i \leq n$$

$y_i \in${+1, -1}-$x_i$-ші жаттығу нүктесінің $i$ класы, $c$ – жаттығу қатесі мен маржа өлшемін теңестіретін реттеу параметрі, $\zeta_i$ - $x_i$-дің қате жіктелу дәрежесін білдіретін теріс емес бос айнымалы; атап айтқанда, егер $\zeta_i > 1$ болса, онда і деректер нүктесі гипержазықтықтың дұрыс емес жағында болады. Бұл SVM деректер нүктелерінің гипер жазықтықтың дұрыс емес жағында пайда болуына мүмкіндік беретіндіктен, ол жұмсақ маржа SVM ретінде де белгілі.

Осы оңтайландыру мәселесін ескере отырып, біз оңтайлы гипер жазықтықты табу үшін SVMlight алгоритмін пайдаланамыз. Яғни, алгоритм сынақ барысында белсенді референтті оның алдыңғы үміткер антецедентпен салыстырылады, сынақ нәтижесінен алынған жұп SVM классификаторына ұсынылады, ол екі ескертудің кореферентті болу ықтималдығын көрсететін мәнді қайтарады. 0-ден жоғары мәндері бар сілтеме жұптары кореферентті болып саналады, кері жағдайда, жұп кореферентті емес болып саналады.

Рейтинг моделін оқыту – рейтинг моделі mk белсенді ескертуінің барлық үміткерлеріне рейтинг қояды. Рейтинг моделін оқыту үшін біз Joachims SVMsvmlight Ranker-learning [13] алгоритмін қолданамыз. Бұл моделдің алдыңғы моделден айырмашылығы, үміткерлер арасындағы референттік дәрежені есептейді, яғни объектіні еске түсіру моделі шығарып берген референті үміткердің ең нақты антецедент дәрежесін ала аламыз [14].

Жоғарыда келтірілген мысал бойынша, $i$ [София Ковалевская**,** әйел], $i$ [София Ковалевская профессор], $i$ [София Ковалевская, тұңғыш әйел математик] осы үміткерлерде рейтинг 2-ге тең болады, ал қалған жағдайларда 1-ге тең. Моделді оқыту барысы төмендегідей. T жиынын белгілеп аламыз және $T \in (x_{jk}, y_{jk})$ болады, мұндағы, $x_{jk}$ – анафориялық ескертпе mk және кандидат антецедент mj арқылы жасалған мүмкіндік

векторы, $y_{jk}$ – оның дәрежелік мәні. Моделді оқытпас бұрын SVM ranker-learning алгоритімі $T$ жиынынан $T'$ жиынын келесідей шығарып аламыз, мұндағы мақсат Т'-де қате жіктеу санын азайтатын гипер жазықтықты табу керек. $(x_{ik}, y_{ik})$ және $(x_{jk}, y_{jk})$, мұндағы $y_{ik} \neq y_{jk}$. $T'$ үшін жаңа жаттығу жиынын жасап аламыз, $T \in (x_{ijk}, y_{ijk})$, мұндағы $x_{ijk} = x_{ik} - x_{jk}$, $y_{ijk} \in \{+1, -1\}$.

SVM ranker-learning алгоритмі шешуге тырысатын шектеулі оңтайландыру мәселесі келесідей болады.

$$arg\ min \frac{1}{2}||w||^2 + c\sum_i \zeta_{ijk} \qquad (3)$$
$$y_{ijk}(w, (x_{ik} - x_{jk}) - b) \geq 1 - \zeta_{ijk}$$

мұндағы $\zeta_{ijk}$-$x_{ijk}$ қате жіктелу дәрежесін білдіретін теріс емес бос айнымалы, ал $c$ – жаттығу қатесі мен маржа өлшемін теңестіретін реттеу параметрі.

Анафораны шешу барсында біз анафор мен антецедеттің функционалдық ерекшеліктерін төмендегі кестедегідей белгілеп аламыз.

**2-кесте**. Анафор мен антецеденттің функционалдық мүмкіндіктері

| № | mj, кандидаттың антецеденті және mk арасындағы байланысты сипаттайтын мүмкіндіктер | |
|---|---|---|
| 1 | Синтаксистік құрлым | A және R Синтаксистік құрылымы бірдей немесе жоқ<br>Y иә,<br>N жоқ |
| 2 | Өзгертілетін шектеу түрі | Өзгертілген шекті түрі A және R бірдей немесе жоқ<br>Y иә,<br>N жоқ |
| 3 | Қашықтық | A және R бір сөйлемде<br>1 иә,<br>2 интервал,<br>3 жоқ |
| 4 | Жекеше немесе көпше | A және R бірдей жекеше немесе көпше<br>Y иә,<br>N әртүрлі,<br>U жоқ |
| 5 | Сәйкестік туралы ақпарат | A және R толық сәйкестігі<br>Y иә,<br>H әртүрлі |
| 6 | Сөздердің ұқсастығы туралы ақпарат | A және R сәйкес<br>Y иә,<br>H әртүрлі |
| 7 | Қысқартылған сөздер | A және R қысқартылған сөздер<br>Y иә,<br>H әртүрлі |

*Модельдерді біріктіру* – Жоғарыда сипатталған кластерлік рейтинг анафориялық ескертпены анықтау үшін пайдалануға болады, бірақ оны ескертудің анафориялық немесе анафориялық емес екенін анықтау үшін пайдалану мүмкін емес. Себебі қарапайым: барлық оқу даналары анафориялық ескертулерден жасалған. Демек, анафорияны анықтауды және корференциялық шешімді бірлесіп үйрену үшін біз анафориялық және анафориялық емес ескертулерден алынған мысалдарды пайдалана отырып үйретуіміз керек.

Атап айтқанда, рейтинг моделді оқыту кезінде біз әрбір белсенді ескертуге (1)

функциялары бар қосымша дананы жасау арқылы жаңа кластерді іске қосу мүмкіндігін береміз, ерекше белсенді ескертуді сипаттайтын функциялар 1-кестеде көрсетілген.

Екі тапсырманы бірге шешудің басты артықшылығы – рейтингтік модельге бір уақытта белсенді ескертпенің барлық мүмкін нұсқаларын (яғни, оны шешу керек пе, егер солай болса, алдыңғы кластердің қайсысы жақсы) бағалауға мүмкіндік береді.

Мәтіндерді оқыту барсында рейтинг моделі оқылған текст мәтінін солдан оңға қарай өңдейді, әр бір белсенді ескертпе mk үшін алдыңғы үміткерлермен жұптастырамыз, егер mk үшін белсенді жұп табылмаса, тек белсенді ескертуді сипаттайтын мүмкіндіктерді қамтитын қосымша сынақ данасын жасап аламыз.

Егер сынақ данасына рейтингі ең төмен дәреже мәнін тағайындаса, онда mk анафорлық емес және одан әрі қарастырылмайды, кері жағдайда, mk ең жоғары дәрежеге ие кластермен байланыстырылады, байланыстар mk-ға ең жақын антецедентті таңдау арқылы үзіледі. Бұрынғыдай, mk алдында тұрған барлық ішінара кластерлер бірінші k – 1 ескерт-пелер бойынша рейтинг модель болжамдары негізінде қадамдық түрде қалыптасады.

Соңында, біз анафораны және кореференцяны шешу бойынша бірлескен оқыту моделіміз бүтін сызықтық бағдарламалауды (ILP) қолдану арқылы шештік, мұнда анафоралық классификатор мен кореференция классификаторы бір-бірінен тәуелсіз оқытылады, содан кейін ILP алгоритм ретінде қолданылады. бұл екі тапсырманы бір-біріне тәуелсіз емес, бірлесіп зерттеуге мүмкіндік береді.

*Эксперимент және нәтижелер.* Біз зерттеу жұмыс барсында қазақ тіліндегі корпусты пайдаландық (https://qazcorpus.kz) [15]. Корпус сайтында қазақ тілінің электронды мәтіндік қоры жинақталған. Корпустағы мәтін көлемі 31 миллион. Мәтіндер қазақ тілінің 5 стиль түрінен (көркем стиль, ғылыми стиль, публицистикалық стиль, ісқағаз стилі, сөйлеу стилі) жинақталған. Біз зерттеуімізге ғылыми жанырдағы мәтіндерді таңдап алдық. Осы бағыттағы зерттелген мәтіндер саны 26, ал сөздердің саны жағынан 350 000 сөз болды. Зерттеу жұмысы кореференцияны және анафораны жеке-жеке тауып шешу болғандықтан, біз нәтижелерді салыстырмалы түрде аламыз және соңында екі модельді біріктіру сынағын аламыз.

Сынақ нәтижелерін санау барысында дәстүрлі есептеу метрикасын қолданамыз. Төменде формула көрсетілген.

$$Precision = \frac{TP}{(TP+FP)} \quad (4)$$

$$Recall = \frac{TP}{(TP+FN)} \quad (5)$$

$$F_1 = \frac{2*Recall*Precision}{(Recall+Rrecision)} \quad (6)$$

Барлық нәтижелер F-өлшемі бойынша сұрыпталған.

Төменде кестеде анафораны шешу үшін қолданылған модель нәтижесі көрсетілген. Мұнда рейтинг моделі оқытылады бұл жерде анафорға үміткер антецедеттерге рейтинг қою арқылы өзіне ең ықтимал жұпты табады, бұл үміткерлер объектіні еске түсіру моделі шығарып берген референті жұптардан таңдап алынады.

**3-кесте**. Анафораны шешу сынақ жинағы бойынша салыстыру нәтижелері

| мәтін | Precision | Recall | $F_1$ |
|---|---|---|---|
| 1-мәтін | 0,786 | 0,725 | 0,705 |
| 2-мәтін | 0,742 | 0,732 | 0,710 |
| 3-мәтін | 0,735 | 0,789 | 0,792 |
| 4-мәтін | 0,758 | 0,463 | 0,752 |
| 5-мәтін | 0,723 | 0,738 | 0,750 |

Кореференцияны шешу моделінің сынақ нәтижесі төменде көрсетілген. Бұл моделде біз 1-кестеде көрсетілгендей кандидаттар үшін арнайы функциялары жасап алдық бұл бізге сөйлемде кездескен референтті жұптарды тез табуға көмектеседі. Сынақ нәтижесі 4-кестеде көрсетілген.

**4-кесте**. Кореференцияны шешу моделінің сынақ жинағы бойынша салыстыру нәтижелері

| мәтін | Precision | Recall | $F_1$ |
|---|---|---|---|
| 1-мәтін | 0,686 | 0,625 | 0,6805 |
| 2-мәтін | 0,572 | 0,632 | 0,630 |
| 3-мәтін | 0,685 | 0,789 | 0,701 |
| 4-мәтін | 0,638 | 0,693 | 0,737 |
| 5-мәтін | 0,723 | 0,738 | 0,743 |

Моделдерді бірге қолдану барысындағы нәтиже төменде 5-кестеде көрсетілген.

**5-кесте.** Модельдері салыстыру нәтижесі

| мәтін | Рейтинг модель (анафораны шешуде) | Объектіні еске түсіру моделі(кореференцияны шешуде) | $F_1$ |
|---|---|---|---|
| 1-мәтін | 0,705 | 0,6805 | 0,7105 |
| 2-мәтін | 0,710 | 0,630 | 0,6804 |
| 3-мәтін | 0,792 | 0,701 | 0,7501 |
| 4-мәтін | 0,752 | 0,737 | 0,7204 |
| 5-мәтін | 0,750 | 0,743 | 0,7310 |

*Қорытынды***.** Осы зерттеу жұмыста кореференция мен анафораны шешуды кластерлік әдістерді қолдана отырып екі модельді біріктіру арқылы бір мезетте екі тапсырманы шешу әдісі ұсынылды.

Жұмыс барысында екі модель оқытылып зерттелді. Рейтинг моделіне (mention-ranking model) кластерлік рейтинг моделін және SVM пакетінен SVMlight алгоритмін оқыту арқылы жүзеге асырылды, ал кореференцияны шешуде объектіні еске түсіру моделін (entity-mention model) және SVM алгоритмін қолдандық.

Жұмыс барысында әр модель жеке-жеке сыналды, мұнда тәжрибе көрсеткендей анафораны табу нәтижесі жоғары болды, тұжырымдаумыз бойынша модель сөйлемдегі анафор және антецеденттің граматикалық белгілерін жақсы талдау жасай алады. Екі моделді бірге қолданған кездегі нәтижені бағалау метрикасы арқылы санағанда 65% құрады, бұл нәтиже кластерлік әдістің тиімділігін көрсетті.

Әдебиеттер тізімі


1. V. Ng, C. Cardie. Improving machine learning approaches to coreference resolution. In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL). – 2002. – Pp. 122-129.
2. Lu, J., & Ng, V. Learning Antecedent Structures for Event Coreference Resolution. 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA). – 2017.
3. Park, C., Choi, K.-H., Lee, C., & Lim, S. Korean Coreference Resolution with Guided Mention Pair Model Using Deep Learning. ETRI Journal. – 2016. – 38(6). – Pp. 1207-1217.
4. Auliarachman, T., & Purwarianti, A. Coreference Resolution System for Indonesian Text with Mention Pair Method and Singleton Exclusion using Convolutional Neural Network. 2019 International

Conference of Advanced Informatics: Concepts, Theory and Applications (ICAICTA). – 2019.
5. Peng, H.; Khashabi, D.; and Roth, D. Solving hard coreference problems. NAACL HLT. – 2015.
6. Martschat, S., and Strube, M. Latent structures for coreference resolution. TACL. – 2015.
7. Қалман Г., Самбетбаева М.А., Жұмабай Е.С. Қазақ тіліндегі есімдік анафорасын шешу алгоритмі, Инновациялық Еуразия университетінің Хабаршысы. – 2022. – № 2. ISSN 2709-3077. – 126 б.
8. Gulzhamal, K., Есмағанбет М.Ғ., Жаманкарин М.М., Габдулина А.И., Плескачев Д.В. Кластерлеу әдісін қолданып кореференциян шешу. Известия НАН РК. Серия физико-математическая. – 2023. – ББ. 121-135. https://doi.org/10.32014/2023.2518-1726.173
9. Қалман, Г., Самбетбаева, М., Жұмабай, Е., Кусаинова, У., Айткенова, М. Решение анафоры в казахском языке на основе алгоритма k-en nearest neighbor. Вестник КазАТК. – 2023. – 124(1). – ББ. 433-441. https://doi.org/10.52167/1609-1817-2023-124-1-433-441
10. Yerzhan Zhumabay,Gulzhamal Kalman; Madina Sambetbayeva; Aigerim Yerimbetova; Assem Ayapbergenova; Almagul Bizhanova. Building a model for resolving referential relations in a multi-lingual system. Eastern-European Journal of Enterprise Technologies, – 2022. – Pp. 27-35. https://doi.org/10.15587/1729-4061.2022.255786
11. Gulzhamal, K., Самбетбаева, М., Актаева, Д., Илюбаев, А. Машиналық оқыту әдістеріне негізделген анафораны шешу моделі. Известия НАН РК. Серия физико-математическая. – 2022. – 4. – ББ. 56-67. https://doi.org/10.32014/2022.2518-1726.156
12. Roth, D. Reasoning with classifiers. In Proceedings of the 13th European Conference on Machine Learning (ECML). – 2017. – Pp. 506-510.
13. Joachims, T. Optimizing search engines using clickthrough data. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). – 2018. – Pp. 133-142.
14. Bjorkelund, A., and Kuhn, J. Learning structured perceptrons for coreference resolution with latent antecedents and non-local features. ACL. – 2017.
15. Қазақ тілінің ұлттық корпусы (qazcorpus.kz)

References

1. V. Ng, C. Cardie. Improving machine learning approaches to coreference resolution. In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL). – 2002. – Pp. 122-129.
2. Lu, J., & Ng, V. Learning Antecedent Structures for Event Coreference Resolution. 2017 16th IEEE International Conference on Machine Learning and Applicaions (ICMLA). – 2017.
3. Park, C., Choi, K.-H., Lee, C., & Lim, S. Korean Coreference Resolution with Guided Mention Pair Model Using Deep Learning. ETRI Journal. – 2016. – 38(6). – Pp. 1207-1217.
4. Auliarachman, T., & Purwarianti, A. Coreference Resolution System for Indonesian Text with Mention Pair Method and Singleton Exclusion using Convolutional Neural Network. 2019 International Conference of Advanced Informatics: Concepts, Theory and Applications (ICAICTA). – 2019.
5. Peng, H.; Khashabi, D.; and Roth, D. Solving hard coreference problems. NAACL HLT. – 2015.
6. Martschat, S., and Strube, M. Latent structures for coreference resolution. TACL. – 2015.
7. G. Kalman, M.A. Sambetbayeva, Y.S. Zhumabay. Algorithm for solving the anaphora of a pronoun in the Kazakh language, bulletin of the innovative university of Eurasia, – 2022. – № 2. – ISSN 2709- 3077. – Pp. 126.
8. G. Kalman, M.G. Esmaganbet M.M. Zhamankarin,A.I. Gabdulina, D.V.Pleskachev coreference solution using the clustering method. news of nas rk. Series physico-mathematical. – 2023. № (1), – Pp. 121-135. https://doi.org/10.32014/2023.2518-1726.173.
9. G. Kalman, M.A. Sambetbayeva, Y.S. Zhumabay, U. Kussainova., M. Aitkenova. Solution of anaphora in the kazakh language based on the algorithm k k-en nearest neighbor. Bulletin of KazATC. – 2023. № 124(1). – Pp. 433-441. https://doi.org/10.52167/1609-1817-2023-124-1-433-441.
10. Yerzhan Zhumabay,Gulzhamal Kalman; Madina Sambetbayeva; Aigerim Yerimbetova; Assem Ayapbergenova; Almagul Bizhanova. Building a model for resolving referential relations in a multi-lingual system. Eastern-European Journal of Enterprise Technologies. – 2022. – № 2. – Pp. 27-35. https://doi.org/10.15587/1729-4061.2022.255786.
11. Gulzhamal K., Sambetbaeva M., Aktaeva D., Ilyubaev A. anaphora resolution model based on machine learning methods. news of nas rk. Series physico-mathematical. – 2022. – № (4). – Pp. 56-67. https://doi. org/10.32014/2022.2518-1726.156.
12. Roth, D. Reasoning with classifiers. In Proceedings of the 13th European Conference on Machine

Learning (ECML). – 2017. – Pp. 506-510.
13. Joachims, T. Optimizing search engines using clickthrough data. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). – 2018. – Pp. 133-142.
14. Bjorkelund, A., and Kuhn, J. Learning structured perceptrons for coreference resolution with latent antecedents and non-local features. ACL. – 2017.
15. Қазақ тілінің ұлттық корпусы (qazcorpus.kz)