**ӘОЖ 004.9**

**МРНТИ 20.19.27**

DOI:<https://orcid.org/0000-0001-8863-9447>

**Г.Қалман 1, М.А.Самбетбаева1,2 , Е.С.Жұмабай3**

1Л.Н. Гумилев атындағы Еуразиялық ұлттық университеті, Нур-Султан, Казахстан;

1,2Ақпараттық және есептеуіш технологиялар институты, Алматы, Казахстан;

*3 «Астана» халықаралық университеті, Нұр-Сұлтан, Казақстан;*

E-mail: [guljamal14@gmail.com](mailto:guljamal14@gmail.com)

**Қазақ тіліндегі есімдік анафорасын шешу алгоритімі**

**Андатпа**

*Негізгі мәселе:* Компьютерлік лингвистиканың алдында тұрған өзекті міндеттердің бірі - мәтіндік құжаттарда әртүрлі нысандар туралы: адамдар, ұйымдар, оқиғалар, орындар және т.б., сондай-ақ олардың арасындағы байланыстар туралы мәліметтерді бөліп көрсету болып табылады. Әрбір ақпараттық объект(нысан) белгілі бір ұғымға/пәндік аймақтың қатынасына сәйкес келеді және белгілі бір құрылымға ие. Табиғи тілді өңдеудегі бұл мәселе референция ұғымымен байланысты болады. Біз мақалада қазақ тіліндегі референциялық қатынасты шешу жолдарын қарастырамыз.

*Мақсаты:*

*Қ*азақ тіліндегі есімдік анафорасын шешудің жолдарын қарастыру;

Зерттеу барысында есімдік анафорасының морфологиялық, синтаксистік және семантикалық белгілерін жіктеу алгоритмдерін, Support Vector Machine және шешім ағаш әдісін пайдалана отырып «антецедент-анафор» жұбын табу.

Біз оқыту және тестілеу деректер жинағы ретінде Tengrinews.kz- тен жаңалықтар топтамасы және Ғ.Мұстафин әңгімелерінен үзінділерді қолдану арқылы әр түрлі мәтін типтеріндегі «антецедент-анафор» жұбын табамыз және сөздер арасындағы қашықтықты есептейміз. Сондай-ақ семантикалық мүмкіндіктердің, атап айтқанда, семантикалық рөлдердің қазақ тіліндегі анафораның шешілуі өнімділігіне қалай әсер ететінін бағалаймыз.

*Әдістері:* Қазақ тілі есімдіктерінің ерекшеліктерін ескеруде әр түрлі мәтіндерді жинақтап формальды талдау жасау әдісін қолдана отырып, жіктеу алгоритмдерін, Support Vector Machine және шешім ағаш әдісі қолданылады.

*Нәтижелер және олардың маңыздылығы*: Қазақ тіліндегі анафоралық есімдіктердің ішінде ең көп кездесетін жіктеу, сілтеу есімдіктері және өздік есімдіктері, зерттеу барысында жинақталған мәтіндерден «антецедент-анафор» жұбының санын білу арқылы анафоралық қатынастың нақты көрсеткіштері саналды, «антецедент-анафор» жұптарының саны есептеліп, график түрінде екінші бөлімде толықтай көрсетілді.

Бұл зерттеу жұмысы қазақ тілінің машиналық аударма, ақпаратты іздеу, ақпаратты алу және т.б. жүйелерде қолданылуы мен түрлі деңгейде талдауларға зор мүмкіндік береді.

*Түйін сөздер*: анафора, машиналық оқыту, қолдау векторы, шешім ағаштары, семантикалық рөлдер.

**Кіріспе**

Анафораны шешу табиғи тілді өңдеудің негізгі мәселелерінің бірі болып табылады. Анафора және корференцияны шешу әдістері машиналық аударма, ақпаратты іздеу, ақпаратты алу және т.б. жүйелерде қолданылады. Анафораны шешу мәселесі ағылшын және басқа еуропалық тілдер үшін кеңінен зерттелген.

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.

Анафора мен корференцияны шешу міндеті 1960 жылдардан бастап белсенді зерттеліп келеді, дегенмен, шешілмеген мәселелері әлі де бар. Бұл мәселені шешудің негізгі тәсілдерін Р.Митковтың [1,2] және басқа да [3,4] зерттеулерінен көреміз. Ағылшын тіліне арналған анафораларды автоматты түрде шешу саласындағы зерттеулер 70 жылдары басталды. Виноградов, Уилкс, Хоббстың [Митков, 1999] алғашқы әдістері мен жүйелері негізінен синтаксистік ақпаратқа негізделген ережелермен жұмыс істеді; сонымен қатар энциклопедиялық білім де кеңінен қолданылды. 80 жылдары бұрын бөлек қолданылған әртүрлі белгілерді біріктіру тенденциясы пайда болды. Э. Рич пен С. Луперфойдың еңбектерінде Дж. Карбонелла, Р. Митков гендер мен сандарды, синтаксистік және семантикалық қатынастарды үйлестіретін алгоритмдерді сипаттады.

Орыс тіліне арналған анафораның шешу эксперименталды түрде аз зерттелген. [5,6] автор орыс тіліндегі анафора құбылысының теориялық аспектілерін талқылайды, және анафораның табиғатын көрсететін тілдік белгілер қатарын сипаттайды. Төлпегіннің [7,8] еңбектерінде машиналық оқыту әдістерін қолдана отырып, орыс мәтіндеріндегі есімдік анафораны шешудің статистикалық моделін құру алгоритмін ұсынады. Жұмыста [9] авторлар әлеуметтік-саяси мәтіндердің үйлесімділік ережелерін талдау үшін пайдаланатын әртүрлі сөйлемдер мен жағдайларда анафориялық қатынастарды анықтау принциптерін егжей-тегжейлі сипаттайды.

Қазіргі тәсілдер аннотацияланған корпустарды қолдана отырып, автоматты оқытуға негізделген. Олар дәстүрлі лингвистикалық әдістерді статистикалық әдістермен біріктіреді және морфологиялық, синтаксистік, семантикалық және тезаурус жиынтығы сияқты әр түрлі оқыту түрлерін қолданады.

Бұл жұмыста біз тек есімдік анафора шешуін қарастырамыз және әр түрлі мәтіндердегі нәтижелерді салыстырамыз. Оқыту және тестілеу деректер жинағы ретінде біз Tengrinews.kz- тен жаңалықтар топтамасын және Ғ.Мұстафин әңгімелерінен үзінділерді қолдандық.

**Материалдар мен әдістер**

Мақаланы жазу барысында формальды талдау жасау әдісі, жіктеу алгоритмдерін, Support Vector Machine және шешім ағаш әдісі қолданылады.

**Есімдік анафорасы**

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 сілтеу жіктеу есімдігінің анафоралық қызметін көре аламыз.

Мысалы: Труба түбіндегі жапырық тас үй – **мехцех**. **Бұл** - әншейін келешегіне қарай қойылған ат, əйтпесе нобайы түзу бір механизм жоқ.

Бұл синтаксистік күрделі бірліктің бірінші сөйлемі мен екінші сөйлемін байланыстырып тұрған – бірінші сөйлемдегі **мехцех** сөзінің екінші сөйлемде **бұл** есімдігімен қайталанып тұруы анафоралық қатынас болады.

**Елжас** өткен айда ***Бразилияда*** болды. **Ол *сол*** елден саған сыйлық алып келіпті.

Мысалда бірінші сөйлемде **Елжас**ты екінші сөйлемде жіктеу есімдігінің үшінші жақ формасында **ол** арқылы қайталанып тұруы және **Бразилияның** сол сілтеу есімдігімен қайталанып тұруы анафоралық қатынас болады.

**Есімдік анафорасын шешу**

Анафораны шешу – «антецедент-анафор» дұрыс жұптарын анықтау міндетін біз өз зерттеулерімізде тек жіктеу, сілтеу, өздік есімдіктермен қарастырамыз.

Антецедент анаформен саны мен септелуі бойынша сәйкес болуы керек. Сөздегі анафор мен антецедент арасындағы қашықтық мәтінге байланысты алдын ала анықталған мәннен аспауы керек. Біз анафора мәселесін жіктеу мәселесі ретінде қарастырамыз және машиналық оқыту әдістерін қолдану арқылы шешеміз. Жіктеу үшін келесі белгілер қолданылды:

Морфологиялық және синтаксистік ерекшеліктері:

1) анафордың тегі, саны, септелуі

2) антецеденттің саны, септелуі;

3) жанды және жансыз (зат есім болған жағдайда) анафора мен антецедентті салыстыру;

4) анафор мен антецедент арасындағы сөйлемдер саны;

5) анафор мен антецедент арасындағы сөздердің саны;

6) анафор мен антецедент арасындағы зат есімдердің саны;

Семантикалық ерекшеліктері:

7) анафордың семантикалық рөлдері;

8) антецеденттің семантикалық рөлдері;

Морфологиялық талдау барысында анафор антецедентпен сәйкес келуі, яғни 1-3 жіктеулерінде саны, тегі, септелуі, жанды және жансыз (зат есім болған жағдайда) сәйкес болуы, 4-6 жіктеулерінде ерекшеліктері анафор мен антецедент арасындағы қашықтық туралы әртүрлі масштабта ақпарат береді.

Сөздегі қашықтық. Әрбір үміткер үшін сөздердегі есімдікке дейінгі қашықтық есептеледі. Осы қашықтыққа байланысты вектор бірліктермен толтырылады. Оларды векторға бекіту үшін үш үзіліс бөлінеді:

* 10 сөзден бастап санау; вектор [1, 0, 0];
* 10-нан 30 сөзге дейін; вектор [0, 1, 0];
* 30 сөзден артық; вектор [0, 0, 1].

Үміткерге векторлық формадағы сипаттамасы бар бір ғана вектор сәйкес келуі мүмкін.

**Оқу деректер жинағын құру алгоритмі**

1. Мәтіндер жиынынан «антецедент-анафор» жұбын табыу.
2. Анафор мен антецедент арасындағы барлық есімдік пен зат есімді табу. Олардың саны мен септелуіне анафора сәйкес келуі керек. Іздеу аймағы алдын ала анықталған сөздер санымен шектеледі.
3. 2-қадамда табылған барлық зат есімдер мен есімдіктер дұрыс емес гипотетикалық антецеденттер.
4. Егер дұрыс антецедент іздеу аймағында болмаса, ол оқу жинағына қосылмайды.
5. Әрбір өңделген мысал үшін 1-4 қадамдарды орындалады.

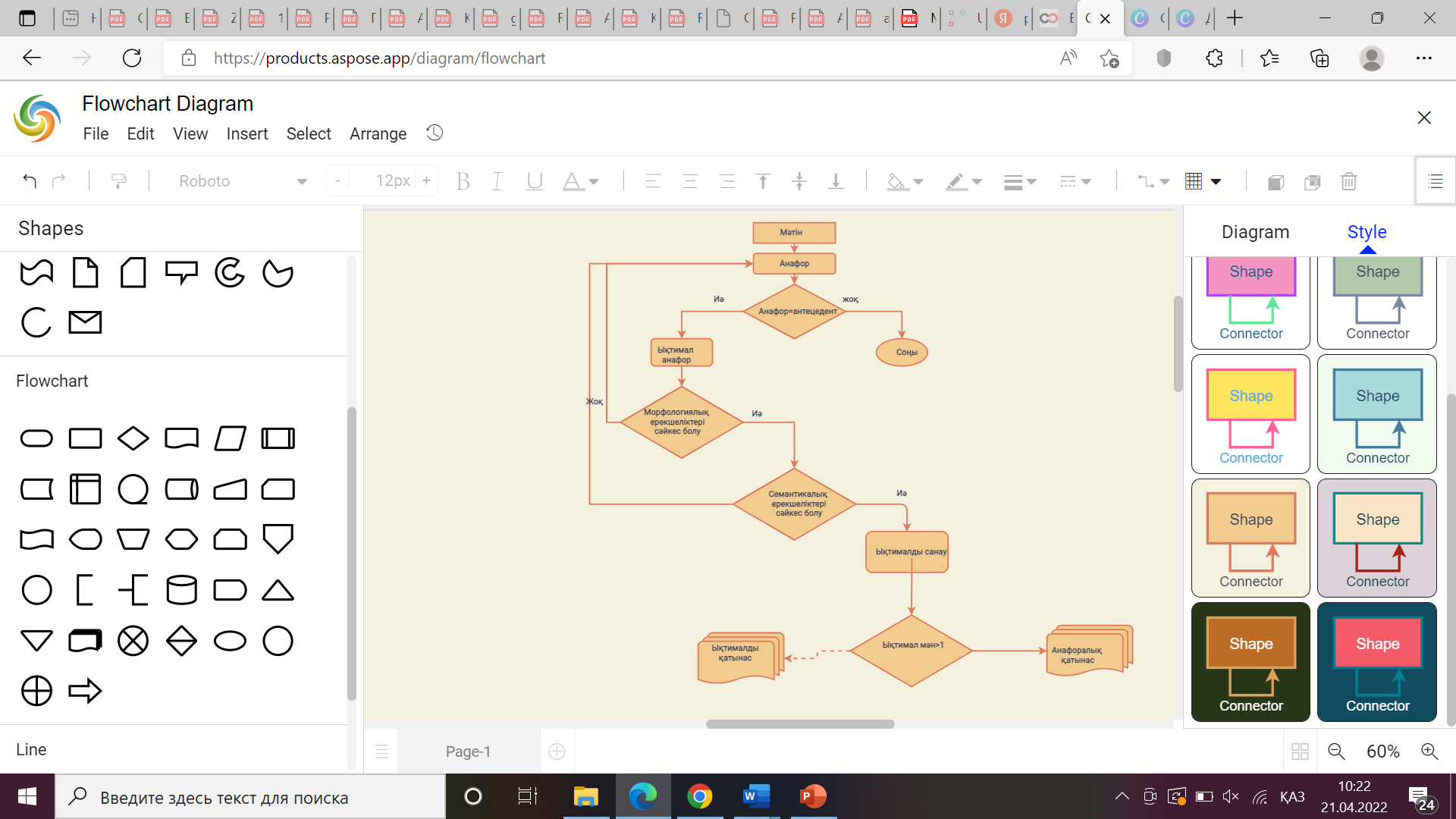
Дұрыс/дұрыс емес жұптарды оқыту және жіктеу үшін біз REPtree векторлық машина әдісін (SVM) [Чанг және Лин, 2014] және шешім ағаш әдісін [Вайкато университеті, 2014] қолдандық.

**Анафораны шешу алгоритмі**

1. Антецеденті табылмаған бірінші анафораны табу. Егер анафора табылмаса, алгоритм аяқталады.
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.
3. Оларды гипотетикалық антецеденттер жиынына қосыу.
4. Гипотетикалық антицеденттер жиынтығындағы әр есімдікке, антецеденттің семантикалық рөлдерін сәйкестендіру.
5. Жіктеу әдісін қолдана отырып, әрбір гипотетикалық антецеденттің дұрыс антецедент болу ықтималдығын есептеу.

6. Ықтималдығы жоғары антецедентті таңдап, оны тиісті анафорамен байланыстыру. 1-қадамға өту.

Гипотетикалық антецедентті іздеу аймағы 2-қадаммен шектеледі, себебі анафора әдетте ең жақын гипотетикалық антецедентті білдіреді. Бұл мән біздің эксперименттерімізде есептелді.



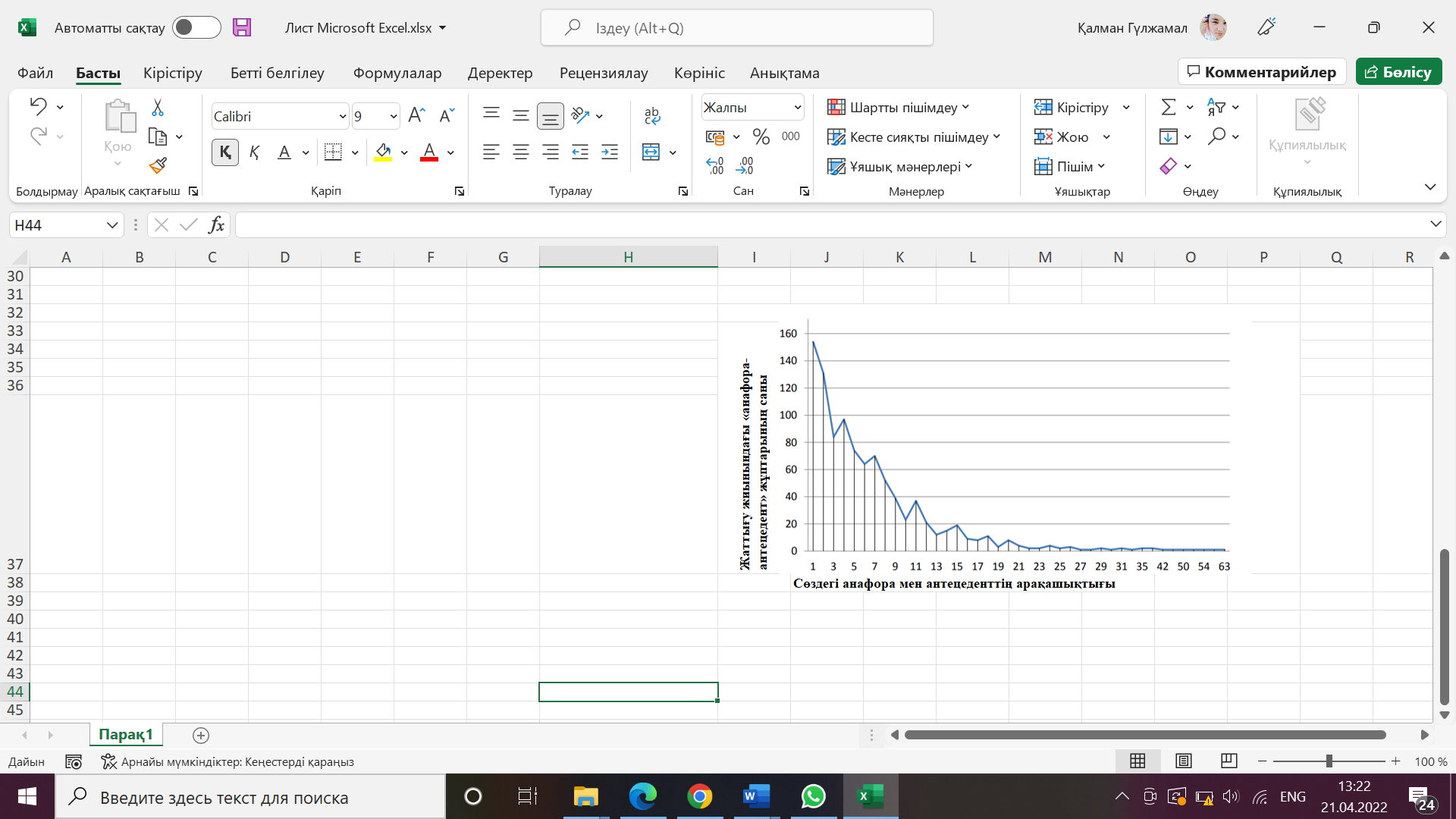
1-сурет-Анафораны шешу алгоритмінің блок-схемасы

Анафоралық қатынасты шешуде анафораның морфологиялық және семантикалық ерекшеліктері (1-суретте) антецедентпен сәйкес келуі анафоралық қатынасты дұрыс табуда маңызды факторлардың бірі болып саналады.

**Нәтижелер**

Зерттеу жұмысы аясында біз 100- ден аса әр түрлі тақырыптағы мәтіндерге талдау жасадық, зерттеу нәтижелерін сандық бағалау кезінде біз келесі мәліметтерді алдық. Бірінші мәтін жиынтығында жалпы 17 мәтінге зерттеу жасалып, 46 «антецедент-анафор» жұбы табылса, екінші мәтіндер жиынында 20 мәтін қаралып, 67 «антецедент-анафор» жұбы табылды.

Алдын ала жүргізілген тәжірибелердің нәтижелері гипотетикалық антецедентті іздеу аймағын шектейтін сөздердегі қашықтық ең маңызды мүмкіндіктердің бірі екенін көрсетті. Әрбір деректер жинағы үшін антецедент пен анафора арасындағы қашықтыққа сәйкес дұрыс «антецедент-анафор» жұптарының санының үлестірімі 2-3 суреттерде берілген.



2-сурет- Tengrinews-тегі «антецедент-анафор» жұптарының санымен арақашықтығы

үстел элементтерін қамтитын сурет

Сипаттама автоматты түрде жасалды

3-сурет- Ғ.Мұстафин әңгімелері «антецедент-анафор» жұптарының санымен арақашықтығы

Біз осы үлестірімдерді (2-3 суреттерде) пайдалана отырып, әрбір деректер жинағы үшін дұрыс «антецедент-анафор» жұптарының 90% қамтитын оңтайлы қашықтықты есептедік. Бұл оңтайлы қашықтық tengrinews үшін 14 сөзге, Ғ.Мұстафин әңгімелері үшін 25 сөзге тең.

**Қорытынды**

Мақалада қазақ тіліндегі есімдік анафорасын шешу алгоритмінің қарапайым әдістері баяндалды. Әрбір деректер жинағы үшін «антецедент-анафор» жұптарының морфологиялық және семантикалық ерекшеліктерін өзара сәйкестігі және сөздер арасындағы арақашықтық ең маңызды фактор екендігі айқындалды.

**ПАЙДАЛАНЫЛҒАН ӘДЕБИЕТТЕР ТІЗІМІ**

1 Mitkov R. (1999) Anaphora resolution: the state of the art, working paper (based on the COLING’98/ACL’98 tutorial on anaphora resolution), available at: clg.wlv. ac.uk/papers/mitkov-99a.pdf

2 Mitkov R. Anaphora resolution. Mitkov R. (ed.) The Oxford handbook of computational linguistics, ch.14, N.Y.: Oxford university press, 2003.

3 Elango P. Coreference Resolution: A Survey // Technical Report. UW-Madison. 2006. – <https://ccc.inaoep.mx/~villasen/index_archivos/cursoTATII/EntidadesNombradas/Elango-SurveyCoreferenceResolution.pdf>

4 Prokofyev R., Tonon A., Luggen M., Vouilloz L., Difallah D.E., Cudr´e-Mauroux P. SANAPHOR: Ontology-Based Coreference Resolution // In Lecture Notes in Computer Science, Volume 9366, 14th International Semantic Web Conference, Proceedings, 2015, Part I.

5 Kibrik A. A. (1996), Anaphora in Russian narrative discourse: A cognitive calculative account In B, Fox (ed.) Studies in anaphora, Amsterdam, pp. 255–304.

6. Kibrik A. A., Dobrov G. B., Khudyakova M. V., Loukachevitch N. V., Pechenyj A. (2013), A corpus-based study of referential choice: Multiplicity of factors and machine learning techniques, Text processing and cognitive technologies. Cognitive modeling in linguistics: Proceedings of the 13th International Conference, Corfu, pp. 118–126.

7. Tolpegin P. V. (2006), The new methods and algorithms of automated third person pronominal reference resolution of Russian texts, [Novye metody i algoritmy avtomaticheskogo razreshenija referentsii mestoimenij tret’ego litsa russkojazychnyh tekstov], Komkniga, Moscow.

8 Tolpegin P. V., Vetrov D. P., Kropotov D. A. (2006), Automated third person anaphora resolution algorithm on the basis of machine learning methods [Algoritm avtomatizirovannogo razreshenija anafory mestoimenij tret’ego litsa na osnove metodov mashinnogo obuchenija], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2006”, [Komp’juternaja Lingvistika i Intellektual’nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii “Dialog 2006”], Bekasovo, pp. 504–507.

9 Abramova N. N., Abramov V. E., Nekrasova E. V., Ross G. N. (2011), Statistic analysis of social and political texts coherence [Statisticheskij analiz svjaznosti tekstov po obshchestvenno-politicheskoj tematike], Proceedings of the 13th All-Russian Scientific Conference “Digital Libraries: Advanced Methods and Technologies, Digital Collections” [Trudy 13j Vserossijskoj nauchnoj konferentsii “Èlectronnye biblioteki: perspektivnye metody i tehnologii, èlektronnye kollektsii”], Voronezh, pp. 127–133.

10 Chang C.-C., Lin C.-J. (2014), LIBSVM—A Library for Support Vector Machines, available at: [www.csie.ntu.edu.tw/~cjlin/libsvm/](http://www.csie.ntu.edu.tw/~cjlin/libsvm/)

11 University of Waikato, (2014), Weka 3: Data Mining Software in Java, available at: www.cs.waikato.ac.nz/ml/weka/ 20. Zhou H., Li Y., Huang D., Zhang Y., Wu C., Yang Y. (2011), Combining syntactic and semantic features by SVM for unrestricted coreference resolution, Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task (CONLL Shared Task’11), Stroudsburg, pp. 66–70

**Қалман.Г1, Самбетбаева.М.А1,2** **Жұмабай Е.С.3**

1Евразийского национального университета им. Л. Н. Гумилева, Нур-Султан, Казахстан;

1,2Институт информационных и вычислительных технологий, Алматы, Казахстан.

3Международный университет «Астана», Нур-Султан, Казахстан;

E-mail: [guljamal14@gmail.com](mailto:guljamal14@gmail.com)

Алгоритм решения анафоры местоимения в казахском языке

Аннотация

*Основная проблема*: Одной из важнейших задач компьютерной лингвистики является выделение в текстовых документах информации о различных объектах: людях, организациях, событиях, местах и т. д., а также о связях между ними. Каждый информационный объект (объект) соответствует определенному отношению понятия/предметной области и имеет определенную структуру. Эта проблема обработки естественного языка связана с понятием референции. В данной статье мы рассмотрим способы решения анафоры языка в казахском языке.

*Цель:* рассмотреть способы решения анафоры местоимений в казахском языке;

В ходе исследования найти пару «антецедент-анафора» с помощью алгоритмов классификации морфологических, синтаксических и семантических признаков местоимения анафоры, машины опорных векторов и метода дерева решений.

В качестве набора обучающих и тестовых данных находим пару «антецедент-анафора» в разных типах текста, используя серию новостей Tengrinews.kz и отрывки из рассказов Г.Мустафина, и вычисляем расстояние между словами. Мы также оцениваем, как семантические возможности, в частности семантические роли, влияют на выполнение анафорных решений в казахском языке.

*Методы:* с учетом особенностей существительных казахского языка используются алгоритмы классификации, метод опорных векторов и метод дерева решений, с использованием метода формального анализа различных текстов.

*Результаты и их значимость:* наиболее распространенная классификация, опорные имена и имена собственные среди анафорических местоимений в казахском языке, фактические показатели анафорической связи рассчитаны по знанию количества пар «антецедент-анафора» из текстов, собранных в ходе исследования. исследование полностью показано во втором разделе.

Данная исследовательская работа включает в себя машинный перевод казахского языка, поиск информации, поиск информации и т. д. могут использоваться в системах и анализироваться на разных уровнях.

*Ключевые слова:* разрешение анафоры, машинное обучение, метод опорных векторов, деревья решений, семантические роли

**G.Kalman1, M.A. Sambetbayeva1,2,** **Y.S. Zhumabay 3**

1L.N. Gumilyov Eurasian National University, Nur-Sultan, Kazakhstan.

1,2Institute of Information and Computing Technologies, Almaty, Kazakhstan.

3Astana international university, Nur-Sultan, Kazakhstan.

E-mail: [guljamal14@gmail.com](mailto:guljamal14@gmail.com)

**Algorithm for solving the anaphora of a pronoun in the Kazakh language**

Abstract

*Main problem*: One of the most important tasks of computer linguistics is the separation of information in text documents on various objects: people, organizations, events, places, etc. , and about the ties between them. Each information object (object) corresponds to a specific concept / subject area and has a definite structure. This problem of processing of the natural language relates to the concept of reference. In this article we will consider the methods of solving the anaphora of the language in the Kazakh language.

*Purpose:* review the methods of solving anaphora pronouns in the Kazakh language.

In the course of research to find a pair of "antecedent-anaphora" with the help of algorithms for the classification of morphological, syntactic and semantic signs of the place names of anaphora, machines of support vectors and the method of tree solutions.

As a set of training and test data we find a pair of "antecedent-anaphora" in different types of text, using a series of news Tengrinews.kz and excerpts from the stories of G. Mustafina, and calculate the distance between words. We also appreciate the fact that semantic possibilities, in particular semantic roles, affect the implementation of anaphoric decisions in the Kazakh language.

*Methods:* considering the specifics of the existing Kazakh language, use the algorithms of classification, the method of supporting vectors and the method of tree solutions, using the method of formal analysis of different texts.

*The results and their significance:* the most common classification, support names and proper names of anaphoric places in the Kazakh language, the actual indicators of anaphoric connections are calculated according to the meaning of the words "anaphoric" the full study is shown in the second section.

This research work includes machine translation of the Kazakh language, search of information, search of information, etc. can be used in systems and analyzed at different levels.

*Keywords:* anaphora, machine learning, support vector, tree of solutions, semantic roles.

**Авторлар туралы мәліметтер:**

**Қалман.Г –** 8D06103 «Ақпараттық жүйе» мамандығының докторанты,Л.Н. Гумилев атындағы Еуразиялық ұлттық университеті, Нұр-Сұлтан, Қазақстан Республикасы. **Қалман.Г-** докторант по специальности 8D06103 «Информационные системы» Евразийского национального университета им. Л. Н. Гумилева, Нур-Султан, Казахстан. **G.Kalman** doctoral student of specialty 8D06103 “Information system” at L.N.Gumilyov Eurasian National University, Nur-Sultan, Kazakhstan. [guljamal14@gmail.com](mailto:guljamal14@gmail.com) ,

**Самбетбаева.М.А-**PhD, Доцент Л.Н. Гумилев атындағы Еуразиялық ұлттық университеті, Нұр-Сұлтан, Ақпараттық және есептеуіш технологиялар институтының, Алматы, Қазақстан. **Самбетбаева.М.А-** PhD, ДоцентЕвразийского национального университета им. Л. Н. Гумилева, Нур-Султан, Институт информационных и вычислительных технологий, Алматы, Казахстан. **Sambetbayeva .M.A** PhD, Associate Professor of L.N. Gumilyov Eurasian National University, Nur-Sultan, Institute of Information and Computing Technologies, Almaty, Kazakhstan. E-mail: [madina\_jgtu@mail.ru](mailto:madina_jgtu@mail.ru),

**Жұмабай Е.С**. 8D06102 – IT-менеджмент мамандығының докторанты, Астана халықаралық университеті, Нұр-Сұлтан, Қазақстан Республикасы. **Жұмабай Е.С.** докторант специальности 8D06102 – Управление информационным технологиями в международном университете Астана, Нур-Султан, Казахстан. **Zhumabay Y. S** – doctoral student of specialty 8D06102 – ITmanagement at Astana international university, Nur-Sultan, Kazakhstan. E-mail: erzhan\_93kz@list.ru