ӘОЖ 004.9
МРНТИ 20.19.27

DOI: https://doi.org/10.37788/2022-2/120-125

Г. Қалман[1*], М.А. Самбетбаева[2], Е.С. Жұмабай[3]
[1]Л.Н. Гумилев атындағыЕуразиялықұлттықуниверситеті, Казақстан
[2]Ақпараттықжәнеесептеуіштехнологиялар институты,Казақстан
[3]«Астана» халықаралықуниверситеті, Нұр-Сұлтан, Казақстан
*(e-mail: 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 біздің есімдік анафора шешуінқарастырамыз және әр түрлі мәтіндердегі нәтижелерді салыстырамыз. Оқыту және тестілеу деректер жинағы ретінде бізTengrinews.kz- тен жаңалықтар топтамасын жәнеҒ.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, әйтпесе 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 Елжасты 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;

– 10-нан 30 сөзге дейін; вектор];

– 30 сөзден артық; вектор [10].

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) [10] және шешім ағаш әдісін [11] 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.

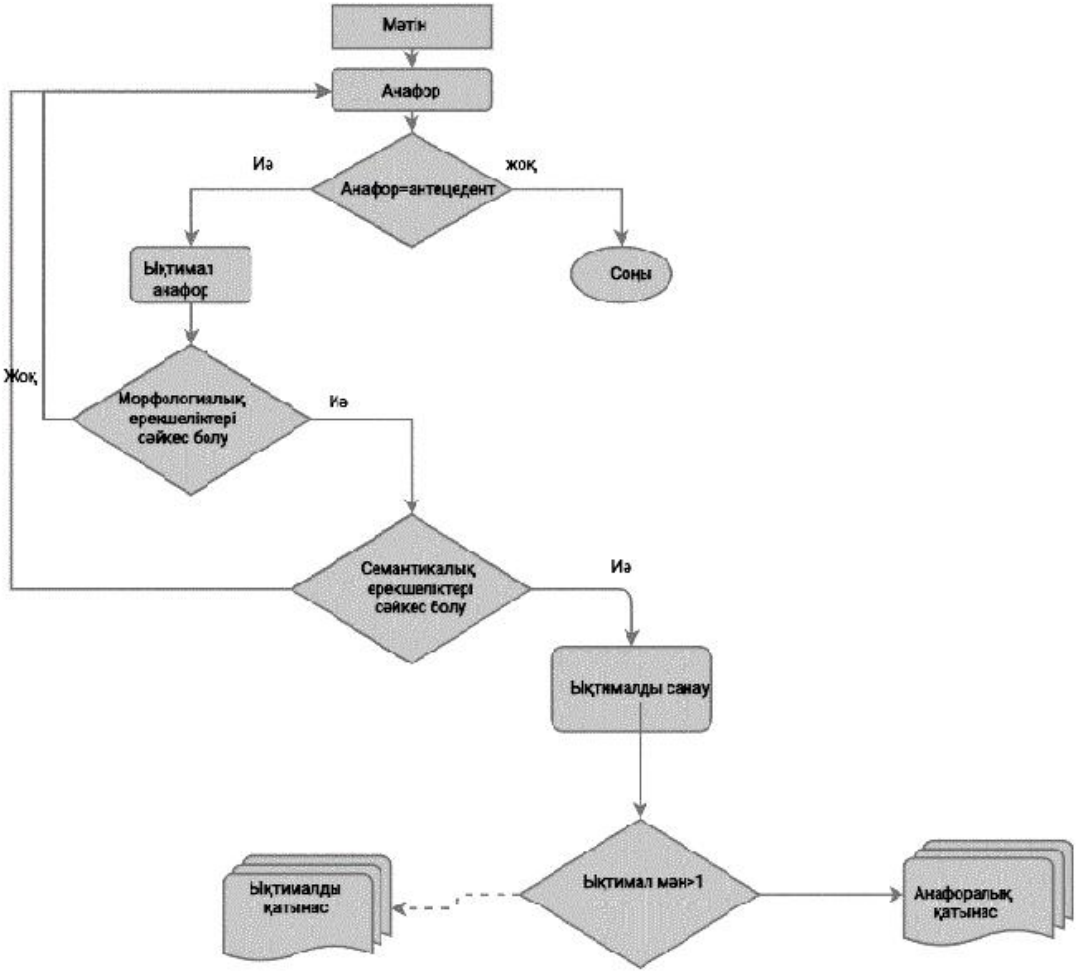
3. Оларды гипотетикалық антецеденттер жиынына қосу.

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

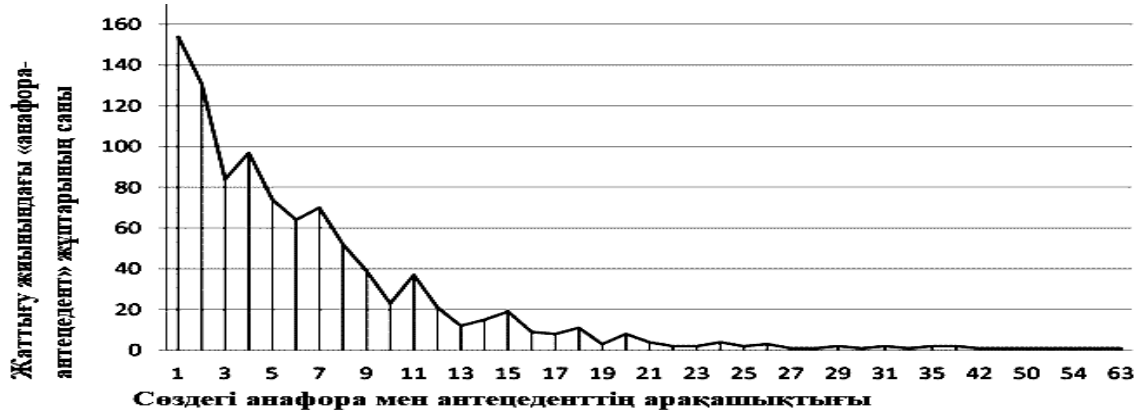
Анафоралық қатынасты шешуде анафораның морфологиялық және семантикалық ерекшеліктері (1-суретте) антецедентпен сәйкес келуі анафоралық қатынасты дұрыс табуда маңыздыфакторлардың бірі болып саналады.

Нәтижелер

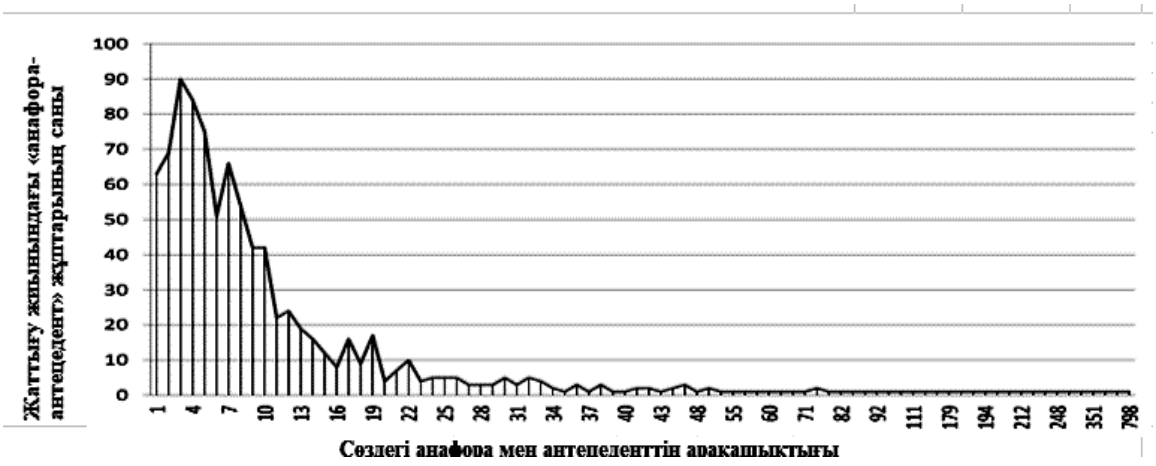
Зерттеу жұмысы аясында біз 100-ден аса әр түрлі тақырыптағы мәтіндерге талдау жасадық, зерттеу нәтижелерін сандық бағалау кезінде біз келесі мәліметтерді алдық.Бірінші мәтін жиынтығында

жалпы 17 мәтінге зерттеу жасалып, 46 «антецедент-анафор» жұбы табылса, екінші мәтіндер жиынында 20мәтін қаралып, 67«антецедент-анафор» жұбы табылды.

Алдын ала жүргізілген тәжірибелердің нәтижелері гипотетикалық антецедентті іздеу аймағын шектейтін сөздердегі қашықтық ең маңызды мүмкіндіктердің бірі екенін көрсетті.Әрбірдеректер 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



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.

Талқылау

Зерттеу барысында есімдік анафорасын шешу жұмыстанырна теориялық зерттеулер жүргізілді.[1-9] жұмыстарында ағылшын және орыс тіліндегі есімдік анафорасын шешу алгоритмдері мен әдістеріне, сондай ақ анафораны шешудің қазіргі кезде көп қолданылатын аннотацияланған корпустарды қолдана отырып, автоматты оқытуға негізделген әдістеріне толық теориялық зерттеу жүргізілді.

Біздің зерттеуімізде қазақ тіліндегі есімдік анафорасын шешу алгоритмі ұсынылды, мұнда REPtree векторлық машина әдісін (SVM) [10] және шешім ағаш әдісіне [11] біріктіру әрекеті жасалды.

Ұсынылып отырған алгоритм қазақ тіліндегі есімдік анафорасын шешу және референциялық қатынастрады шешу ережелерін кеңейту болып табылады. алгоритм ерекшелігі морфологиялық және синтаксистік өңдеу сатысында мәтінді талдаудың есептеу және мұнда REPtree векторлық машина әдісін (SVM) [10] және шешім ағаш әдістерін [11] біріктіруінде. Анафоралық қатынастарды шешудің эксперименталды зерттеуі жүргізілді. Эксперименталды зерттеу tengrinews жаңалық топтамаларымен Ғ.Мұстафин әңгімелеріне жүргізілді, әрбір деректер жинағы үшін дұрыс «антецедент-анафор» жұптарының 90 % қамтитын оңтайлы қашықтықты есептелді. Бұл оңтайлы қашықтық tengrinews үшін 14-сөзге, Ғ.Мұстафин әңгімелері үшін 25-сөзге тең.

Бұл алгоритм қазақ тілінде референциялық қатынастарды шешу, корпустық зерттеу және мәтінді автоматты өңдеуде және осы тақырып бойынша басқада зерттеулер жүргізуде ыңғайлы қол жетімділікті қамтамасыз етеді.

Қорытынды

Мақалада қазақ тіліндегі есімдік анафорасын шешу алгоритмінің қарапайым әдістері баяндалды. Әрбірдеректер 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 artCiteseer.
2Mitkov R. (2003).Anaphora resolution. The Oxford handbook of computational linguistics, ch.14, N.Y.: Oxford University.
3 Elango P. (2006).Coreference Resolution: A Survey. Technical Report. UW-Madison.
4 Prokofyev R., Tonon A., Luggen M., Vouilloz L., Difallah D.E., Cudr´e-Mauroux P.(2015) SANAPHOR: Ontology-Based Coreference Resolution.14th International Semantic Web Conference, part I, LNCS, vol. 9366, 458-473.
5 Kibrik A.A. (1996). Anaphora in Russian narrative discourse: A cognitive calculative account in B, Fox (ed.) Studies in anaphora.Amsterdam, 255–304.
6. KibrikA.A., DobrovG.B., KhudyakovaM.V., LoukachevitchN.V., PechenyA. (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 (pp. 118–126).Corfu.
7Толпегин П.В.Новые методы и алгоритмы автоматического разрешения референции местоимений третьего лица русскоязычных текстов. – М.: Ком-Книга, 2006. – 88 с.
8 Толпегин П.В., Ветров Д.П., Кропотов Д.А. Алгоритм автоматизированного разрешения анафоры местоимений третьего лица на основе методов машинного обучения /Диалог 2006 – Компьютерная лингвистика и интеллектуальные технологии:материалы Международной конференции (31 мая – 4 июня 2006 года). –Бекасово, 2006. – С. 504–507.
9 Абрамов В.Е., Абрамова Н.Н., Некрасова Е.В., Росс Г.Н. Статистический анализ связности текстов по общественно-политической тематике / Электронные библиотеки: перспективные методы и технологии, электронные коллекции: материалы 13 Всероссийской научной конференции. - Воронеж, 2011. - С. 127–133.
10 ChangC.C., LinC.J. (2014).LIBSVM-ALibrary for Support Vector Machines, availableat [Электронныйресурс]. – Режимдоступа: www.csie.ntu.edu.tw/~cjlin/libsvm .
11 Zhou, H., Li, Y., Huang, D., Zhang, Y., Wu, C., Yang, Y. (2011).Combining syntactic.Weka 3: Data Mining Software in Java, available at University of Waikato. Retrievedfrom: www.cs.waikato.ac.nz/ml/weka/ 20 .

REFERENCES

1 Mitkov, R. (1999). Anaphora resolution: the state of the art. Citeseer
2 Mitkov, R. (2003).Anaphora resolution.The Oxford handbook of computational linguistics, ch.14, N.Y.: Oxford University.
3 Elango, P. (2006). Coreference Resolution: A Survey. Technical Report. UW-Madison.
4 Prokofyev, R., Tonon, A., Luggen, M., Vouilloz, L., Difallah, D.E., Cudr´e-Mauroux P. (2015.) SANAPHOR: Ontology-Based Coreference Resolution.14th International Semantic Web Conference, part I, LNCS, vol. 9366, 458-473.
5 Kibrik, A.A. (1996). Anaphora in Russian narrative discourse: A cognitive calculative account in B, Fox (ed.) Studies in anaphora. Amsterdam, 255–304.
6. Kibrik, A.A., Dobrov, G.B., Khudyakova, M.V., Loukachevitch, N.V., Pecheny, 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 from the 13th International Conference (pp. 118–126).Corfu.
7 Tolpegin, P.V. (2006).Novye metody i algoritmy avtomaticheskogo razreshenija referentsii mestoimenij tret'ego litsarusskojazychnyh tekstov [The new methods and algorithms of automated third person pronominal reference resolution of Russian texts]Moscow: Kom-kniga [in Russian].
8 Tolpegin, P.V., Vetrov, D.P., Kropotov, D.A. (2006). Algoritm avtomatizirovannogo razreshenija anaforymestoimenij tret'ego litsa na osnove metodov mashinnogo obuchenija [Automated third person anaphora resolution algorithm based on machine learning methods].Proceedings from the Dialog 2006 - Computational Linguistics and Intellectual Technologies*:* Mezhdunarodnaya konferentsai Dialog 2006 – Kompyuternaya lingvistika i intellektualnyye tekhnologii (31 maya – 4 iyunya 2006 hoda)-International Conference(pp. 504–507).Bekasovo[in Russian].

9 Abramova, N.N., Abramov, V.E., Nekrasova, E.V., Ross, G.N. (2011). Statisticheskij analiz svjaznostitekstovpoobshchestvenno-politicheskoj tematike [Statistic analysis of social and political texts coherence].]. Proceedings from Digital Libraries: Advanced Methods and Technologies, Digital Collections' 13:V serossiyskay nauchnaya konferentsia Elektronnyye biblioteki: perspektivnyye metody i tekhnologii. elektronnyye kollektsii (pp. 127–133). Voronezh [in Russian].
10 Chang C.C., Lin C.J. (2014). LIBSVM-ALibraryforSupport Vector Machines, availableat [Электронныйресурс]. – Режимдоступа: www.csie.ntu.edu.tw/~cjlin/libsvm .
11 Zhou, H., Li, Y., Huang, D., Zhang, Y., Wu, C., Yang, Y. (2011).Combining syntactic.Weka 3: Data Mining Software in Java, available at University of Waikato. Retrievedfrom: www.cs.waikato.ac.nz/ml/weka/ 20 .

Г. Қалман[1], М.А. Самбетбаева[2], Е.С. Жұмабай [3]
[1]Евразийского национального университета им. Л. Н. Гумилева, Казахстан
[2]Институт информационных и вычислительных технологий, Казахстан
[3]Международный университет «Астана», Казахстан

Алгоритм решения анафоры местоимения в казахском языке

В статье рассматриваются два метода разрешения анафоры казахских текстов на основе данных. Эти методы основаны на машинном обучении с аннотированными корпусами и не используют никакой дополнительной информации, кроме лингвистических признаков.

Первый метод использует метод опорных векторов в качестве алгоритмов обучения и классификации, второй метод использует индуктор дерева решений. Авторы оценивают производительность методов с несколькими наборами функций и корпусов. Наборы признаков включали морфологические, синтаксические и семантические признаки. В этой статье оцениваютсятакже семантические особенности, а именно семантические роли, которые влияют на разрешение анафоры в казахском языке. Эксперименты показали, что точность SVM выше на экспериментальных данных практически для всех случаев. Показано, что семантические признаки повышают эффективность методов анафорного разрешения казахских текстов. Былорассчитано оптимальное расстояние между анафорой и гипотетическим антецедентом и использовано в применяемых методах.

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

G. Kalman[1], M.A. Sambetbayeva[2], Y.S. Zhumabay[3]
[1]L.N. Gumilyov Eurasian National University, Kazakhstan
[2]Institute of Information and Computing Technologies, Kazakhstan
[3]Astana International University, Kazakhstan

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

The paper considers two data-driven methods for anaphora resolution of Kazakh texts. These methods are based on machine learning with annotated corpora and using no additional information except linguistic features.

The first method uses Support Vector Machine as learning and classifying algorithms, the second method uses Decision Tree inducer. We evaluate the performance of the methods with several feature sets and corpora. Feature sets included morphological, syntactic, and semantic features. In this paper

We also evaluate how semantic features, namely semantic roles, impact the performance of anaphora resolution in Kazakh language. Experiments showed that precision of SVM is higher on experimental data for almost all cases. It was shown that semantic features enhance the performance of the methods for anaphora resolution of Kazakh texts. We have also calculated the optimal distance between the anaphor and the hypothetic antecedent and used it in our methods.

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

Қолжазбаның редакцияға келіп түскен күні: 28.04.2022 ж.