ПРОБЛЕМЫ ОБУЧЕНИЯ В СФЕРЕ ГРАЖДАНСКОЙ ЗАЩИТЫ

УДК 004.8

А. Ғ. Мұсайбеков[1], И. А. Захаров[1], Ғ. Ә. Шәріпов[2]

[1]Қазақстан Республикасы ТЖМ Мәлік Ғабдуллин атындағы Азаматтық қорғау академиясы, Көкшетау, Қазақстан
[2]Д. И. Михайлик генерал-лейтенант атындағы Ресей ТЖМ Азаматтық қорғау академиясы, Химки, Ресей Федерациясы

ТАБИҒИ ТІЛДІ ӨҢДЕУГЕ АРНАЛҒАН МАШИНАЛЫҚ ОҚЫТУДЫ ЗЕРТТЕУДІҢ ЗАМАНАУИ ТЕНДЕНЦИЯЛАРЫ

Аңдатпа. Бұл мақалада табиғи тілді өңдеу контекстінде қолданылатын машиналық оқытудың өзекті тәсілдері талданады. Мәтінді морфологиялық және синтаксистік талдаудың стандартты міндеттері қарастырылады, онтологиялық модельдерді жасау үшін қолданылатын Машиналық оқыту әдістеріне ерекше назар аударылады. Бұл тәсілдің бірқатар негізгі мәселелері ерекшеленеді. Статистикалық талдау әдістері ең жақсы нәтиже беретіні атап өтілді. Талдау күрделі жарылысқа әкеледі және сөйлемді талдаудың көптеген нұсқаларын тудырады, ал негізгі қиындықтар пайда болған полисемияны шешумен байланысты.

Түйінді сөздер: машиналық оқыту әдістері, табиғи тілді өңдеу, терең талдау, білім базасы.

Кіріспе

Табиғи тілді өңдеуге арналған бағдарламалық қамтамасыз етуді әзірлеу саласындағы зерттеулер әртүрлі ғылыми бағыттарда қарқынды дамып келеді. Соңғы жылы табиғи тілді өңдеудегі басым тенденциялар статистика мен машиналық оқытудың озық әдістерін әзірлеу және қолдану саласындағы қарқынды зерттеулермен байланысты. Мұндай зерттеулердің айрықша ерекшеліктері:

- нақты бағалау критерийлері бар тәжірибелік әдістерді қолдану;
- статистикалық тәсілдерді қолдану саласын кеңейту;
- көлемді деректерді тиімді басқару;
- практикалық салаларда табиғи тілді өңдеуді қолдану.

Бұл тәсілдің бірнеше негізгі мәселелерін ажыратуға болады. Даму сапасы кең және жоғары сапалы ресурстардың болуына тікелей байланысты. Даму процесін стандарттау ерекше маңызды аспект болып табылады және қазіргі уақытта белгілі бір стандарттар бекітілді, мысалы:

- WordNet;
- PennTreeBank.

Тағы бір мәселе-қолданылатын эмпирикалық критерийлердің тиімділігін бағалау. Табиғи тілді өңдеу саласындағы сандық бағалау ақпаратты алу жүйелерінде

кеңінен қолданылатын «дәлдік» және «толықтық» ұғымдарына ұқсас. Бағалау адам талдаушысы қол жеткізген нәтижелерді және белгілі бір тапсырманы орындау кезінде компьютерлік бағдарламаның тиімділігін салыстыру негізінде қалыптасады [1]. Табиғи тілді өңдеу саласындағы салыстырмалы бағалауды қолдану салалары үздіксіз кеңейіп келе жатқанын атап өткен жөн.

Табиғи тілді өңдеу саласында статистикалық әдістерді қолданудың артуы адамның ойлауы мен тілінің негізінде жатқан терең механизмдерді зерттеу және модельдеу әдістерінен белгілі бір ауысуға әкеледі. Табиғи тілді өңдеуде статистикалық әдістерді қолдану әртүрлі мәселелерді шешуде белгілі бір нәтижелерге қол жеткізе алады. Дегенмен, интроспективті әдістерді қоса алғанда, әртүрлі әдістерді біріктіретін гибридті модельдерді қолдану перспективалы болып табылады.

Ақпаратты алу саласындағы зерттеудің перспективалы бағыттарының бірі «машиналық оқыту» саласы болып табылады. Машиналық оқыту әдістерін жүзеге асыратын жүйелер оқыту процесін автоматтандыру арқылы жаңа білім алуға бағытталған [2-3]. Білім қорын құруда эмпирикалық деректерді тиімді пайдалану жаңа білімді автоматты түрде алу әдістерін сәтті қолдануға мүмкіндік береді. Бұл жағдай тілді оқыту саласындағы ғылыми зерттеулердің өзектілігін көрсетеді, олардың нәтижелері табиғи тілді өңдеуге арналған практикалық қосымшаларда сәтті қолданылуы мүмкін. Табиғи тілді өңдеу саласында одан әрі дамыту үшін машиналық оқыту әдістері саласындағы зерттеулердің маңыздылығын көрсететін бірнеше факторлар бар 1-кесте.

1 кесте – Табиғи тілдің дамуы

Себептері	Зерттеу
Тапсырмалардың күрделілігі	Тіл-бұл заңдылықтар, заңсыздықтар, алып тастау аймақтары және басқа құбылыстар арасындағы күрделі өзара әрекеттесуді қамтитын жоғары ұйымдастырылған объект. Тілді түсінудің тиімдірек тәсілі үшін салыстырмалы түрде қарапайым семантикалық аймақтарды сипаттайтын жеке тілдер үшін модельдер әзірлеуді бастауға болады.
Нақты қолданбалар	Қазіргі уақытта машиналық аударма, реферат және т.б. сияқты табиғи тілді өңдеуге арналған қосымшалар саласында айтарлықтай өсім бар. Машиналық оқыту әдістері табиғи тілді өңдеу саласындағы әртүрлі негізгі мәселелерді шешудің ажырамас бөлігі болып табылады.
Үлкен деректер ресурстарының болуы	Стандарттау және көптеген негізгі ресурстардың қол жетімділігі машиналық оқыту әдістерінің ажырамас ресурстық негізі болып табылады.

Машиналық оқыту әдістерінің жіктелуі

Машиналық оқыту әдістері-бұл деректердегі заңдылықтарды талдауға және шығаруға арналған құралдар. Бұл әдістер сипаттамалық кеңістіктердегі сипаттамалармен ұсынылған объектілерді жіктеу мәселелерін шешу үшін кеңінен қолданылады. Оқытудың негізгі мақсаты-оқыту үлгісін қалыптастыруда қолданылғанға ұқсас жаңа объектілерді жіктеуге мүмкіндік беретін қажетті және жеткілікті ережелерді анықтау. Әрбір оқу мысалы, әдетте, белгілі бір сыныпқа жататындығын көрсететін белгімен бірге жүреді. Бұл тұрғыда жіктеуіш жасалады

1-сурет, ол ұсынылған объектінің класын болжайды. Үздіксіз белгілермен жұмыс істеу жағдайында мұндай процесс регрессия деп аталады [4].

Жіктеуіш:

- көптеген санаттар бар $\Omega = \{К_1, К_2, \ldots К_n\}$;
- көптеген нысандар бар $\Sigma = \{О_1, О_2, \ldots О_m\}$;
- белгісіз мақсатты функция $F: \Omega * \Sigma \rightarrow \{0, 1\}$.

F * классификаторын мүмкіндігінше F-ге жақын құру қажет.

Сурет 1 – Жіктеуішті құру

Машиналық оқытудың негізгі аспектілері:

- сипаттамалық саланы талдау және құру;
- объектілер мен объектілер кластары арасындағы айырмашылықтар немесе ұқсастықтар туралы гипотезаларды зерттеу;
- оқу деректер жинағын құрастыру;
- тексеру үлгісін құру;
- оқу алгоритмін дұрыс таңдау.

Егер белгілердің конфигурациясы оқу тапсырмасына негізінен мазмұн тұрғысынан әсер етсе, онда оқытудың тиімділігі, жылдамдығы мен дәлдігі оқыту мен бақылау үлгілерінің қалыптасуына байланысты. Оқу процесінің бағытын дұрыс таңдалған мысалдар арқылы басқаруға болады [5]. Қадамдық оқыту процедураларын қолдану және мысалдар тізбегін әзірлеу сонымен қатар қажетті оқыту мысалдарының санын азайтуға мүмкіндік береді.

Табиғи тілді өңдеуде жіктеулерді құруға мысал келтірейік.

Мәтіндік құжаттарды жіктеу міндеті.

Нысандар: Мәтіндік құжаттар D_j, j = 1,…, n;

Тапсырма: Мәтіндерді жіктеу ережелерін алыңыз;

Белгісі: Түйінді сөздер T_j, i = 1,…, n;

Wj белгілерінің векторы, сипаттайтын құжат D_j;

- вектор компоненті υ_{ij} O {1,0}, мұндағы 1 D_j мәтінінде T_i кілт сөзінің болуын білдіреді, 0 оның болмауы; мысалы, W_j = [011110], яғни T_1, T_6 D_j-де жоқ; T_2, T_3, T_4, T_5 бар;
- вектор компоненті υ_{ij} O {1,0} D_j құжатындағы T_i кілт сөзінің пайда болу жиілігін көрсетеді, мысалы W_j = [0.00 1.00 0.10 0.75 0.90 1.00].

Табиғи тілді өңдеу саласында екі негізгі бағыт ерекшеленеді: мәтіннен ақпарат алу және мәтіндерден білім алу. Ақпаратты алу кезінде мәтіндерде бар нақты мәліметтер, мысалы, кілт сөздер, күндер, ұйымдардың атаулары, есімдері және т.б. бөлінеді. Бұл процесті мәтіндерден білімді тереңірек алу алдындағы ажырамас кезең ретінде қарастыруға болады.

Бұл кезеңде машиналық оқыту әдістеріне негізделген мәтіндерді санаттау мен жіктеудің әртүрлі жүйелері қолданылады. Атап айтқанда, мәтіндер мен олардың фрагменттері үшін жіктеу белгілерін алдын-ала дайындау қажеттілігімен бірге

анықтамалық векторлар әдісі мен логикалық әдістер қолданылады. Бұл әдістер мәтіндік ақпаратты тиімді өңдеуге көмектеседі және кейінірек мәтіндерден күрделі білімді алуға негіз жасайды.

Табиғи тілді өңдеуде машиналық оқыту әдістерін қолдану

Табиғи тілді өңдеу шеңберінде машиналық оқытудың екі негізгі тәсілі ерекшеленеді: «жалқау» және «ашкөз» оқыту әдістері. Бұл әдістердің басты айырмашылығы – «жалқау» тәсілмен алынған ақпарат жалпыланбайды, ал «ашкөз» тәсілмен ақпарат артық және маңызды емес элементтерді қайта құрылымдау және жою арқылы жалпыланады. «Жалқау» оқыту тәсілі эксперименттерден алынған дерексіз ережелер негізінде емес, когнитивті есептерді шешу ұқсастықтарына сәйкес қорытындылар құру арқылы жүзеге асады деген ұсынысқа негізделген [6]. Бұл әдіс жасанды интеллектті зерттеудің әртүрлі салаларында кеңінен қолданылады, келесі әдістерге негіз болады: ұқсастыққа негізделген қорытынды; мысалға негізделген қорытынды; аналогияға негізделген қорытынды; прецедентке негізделген қорытынды. Бұл әдіс фонология, морфология, сөйлеуді тану және синтаксисті талдау мәселелерінде қолданылады.

«Ашкөздік» оқыту тәсілінің негізгі әдістері шешім ағаштарына негізделген оқыту, индуктивті қорытынды, нейрондық желілерді оқыту және индуктивті логикалық бағдарламалау болып табылады. Шешім ағаштарына негізделген оқыту идеясы мысалдар арасындағы ұқсастықты шешім ағаштарын автоматты түрде қалыптастыру үшін пайдалануға болады [7]. Бұл ағаштар жалпылау мен түсініктеме беру үшін негіз бола алады.

Индуктивті қорытындының мақсаты оқыту мысалдары немесе шешім ағаштары негізінде түсіндірілетін ережелердің шектеулі жиынтығын әзірлеу болып табылады. Индуктивті логикалық бағдарламалау алгоритмдері алдыңғы жағдайларға негізделген бірінші ретті логикалық гипотезаларды қалыптастыру үшін қолданылады.

Қорытынды

Табиғи тілді өңдеудегі шешілетін міндеттер саласын талдау кезінде әртүрлі тәсілдердің тиімділігі туралы қорытынды жасауға болады. Әдісті таңдау жүйенің мақсаттарына байланысты. Егер дәлдік негізгі мақсат болса, «жалқау» оқыту әдісі қолайлы болады. Белгілерді өлшеу әдістерімен және ықтималдық ережелерімен біріктірілген бұл әдістің алгоритмдері лингвистикалық есептердің кең ауқымы үшін тамаша нәтижелерге қол жеткізеді. Егер машиналық оқытудың негізгі мақсаты тексерілетін және түсіндірмелі деректерді жалпылау болса, «ашкөз» оқыту әдістеріне артықшылық беру керек.

Әдебиеттер тізімі

1. Брокман Д. Что мы думаем о машинах, которые думают // Ведущие мировые ученые об искусственном интеллекте. – М.: Альпина нон-фикшн, 2017. – 552 с.

2. Майер-Шенбергер, Кукьер К. Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим / пер. с англ. – М.: Мании, Иванов и Фербер, 2014. – 240 с.

3. Гудфеллоу Я., Бенджио И., Курвиль А. Глубокое обучение / пер. с англ. А. А. Слипкина, 2-е изд. – М.: ДМК Пресс, 2018. – 652 с.

4. Лука В. Д. Сравнение алгоритмов Machine Learning в решении задач распознавания изображений // Информатика; проблемы, методы, технологии: материалы XXIII Междунар. научно-практической конф. им. Э. К. Алгазинова. – Воронеж, 2023. – С. 593-602.

5. Смирнова В. С. Оптимизация гиперпараметров на основе объединения априорных и апостериорных знаний о задаче классификации // Научно-технический вестник информационных технологий, механики и оптики. – 2020. – Т.20, № 6. – С. 828-834.

6. Васильев С.П., Полетаева Н.Г. Применение методов машинного обучения в задачах оптимизации // Информационные системы и технологии: теория и практика: сб. науч. тр. – СПб., 2019. Вып. 11. – С. 28-40.

7. Плескачев Д.В., Кусаинова У.Б., Актаева А.У. Совершенствование методов и средств контроля знаний при помощи технологий визуализации // Вестник Кокшетауского технического института. – 2020. – № 3 (39). – С. 93-96.

References

1. Brokman D. Chto my dumaem o mashinah, kotorye dumayut // Vedushie mirovye uchenye ob iskusstvennom intellekte. – M.: Alpina non-fikshn, 2017. – 552 s.

2. Majer-Shenberger, Kuker K. Bolshie dannye. Revolyuciya, kotoraya izmenit to, kak my zhivem, rabotaem i myslim / per. s angl. – M.: Manii, Ivanov i Ferber, 2014. – 240 s.

3. Gudfellou Ya., Bendzhio I., Kurvil A. Glubokoe obuchenie / per. s angl. A. A. Slipkina, 2-e izd. – M.: DMK Press, 2018. – 652 s.

4. Luka V.D. Sravnenie algoritmov Machine Learning v reshenii zadach raspoznavaniya izobrazhenij // Informatika; problemy, metody, tehnologii: materialy XXIII Mezhdunarodnoj nauchno-prakticheskoj konferencii im. E. K. Algazinova. – Voronezh, 2023. – S. 593-602.

5. Smirnova V.S. Optimizaciya giperparametrov na osnove obedineniya apriornyh i aposteriornyh znanij o zadache klassifikacii // Nauchno-tehnicheskij vestnik informacionnyh tehnologij, mehaniki i optiki. – 2020. – T.20. №6. – S. 828-834.

6. Vasilev S. P., Poletaeva N. G. Primenenie metodov mashinnogo obucheniya v zadachah optimizacii // Informacionnye sistemy i tehnologii: teoriya i praktika: sb. nauch. тr. – SPb., 2019. Vyp. 11. – S. 28-40.

7. Pleskachev D. V., Kusainova U. B., Aktaeva A. U. Sovershenstvovanie metodov i sredstv kontrolya znanij pri pomoshi tehnologij vizualizacii // Vestnik Kokshetauskogo tehnicheskogo institute. – 2020. – № 3 (39). – S. 93-96.

А. Г. Мусайбеков[1], И. А. Захаров[1], Г. А. Шарипов[2]

[1]Академия гражданской защиты имени Малика Габдуллина МЧС Республики Казахстан, Кокшетау, Казахстан

[2]Академия гражданской защиты МЧС России имени генерал-лейтенанта Д. И. Михайлика, Химки, Российская Федерация

СОВРЕМЕННЫЕ ТЕНДЕНЦИИ ИССЛЕДОВАНИЙ В ОБЛАСТИ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА

Аннотация. В данной статье анализируются актуальные подходы к машинному обучению, применяемые в контексте обработки естественного языка. Рассматриваются стандартные задачи морфологического и синтаксического анализа текста, при этом особое внимание уделяется методам машинного обучения, используемым для разработки онтологических моделей. Выделяется ряд ключевых проблем данного подхода. Отмечается, что наилучшие результаты дают статистические методы синтаксического анализа. Синтаксический анализ приводит к сложностному взрыву и порождает множество вариантов синтаксического разбора предложения, при этом основные трудности связаны именно с разрешением возникающих многозначностей.

Ключевые слова: методы машинного обучения, обработка естественного языка, глубинных анализ, базы знаний.

A. G. Mussaibekov[1], I. A. Zakharov[1], G. A. Sharipov[2]

[1]*Malik Gabdullin Academy of Civil Protection of the MES of the Republic of Kazakhstan, Kokshetau, Kazakhstan*
[2]*Civil Defence Academy of the EMERCOM of Russia named after lieutenant-general D. I. Mikhaylik, Khimki, Russian Federation*

CURRENT RESEARCH TRENDS IN MACHINE LEARNING FOR NATURAL LANGUAGE PROCESSING

Abstract. This article analyzes current approaches to machine learning applied in the context of natural language processing. The standard tasks of morphological and syntactic text analysis are considered, with special attention being paid to machine learning methods used to develop ontological models. A number of key problems of this approach are highlighted. It is noted that statistical methods of syntactic analysis provide the best results. Syntactic analysis leads to a complex explosion and generates many options for syntactic parsing of a sentence, while the main difficulties are related precisely to the resolution of emerging ambiguities.

Keywords: machine learning methods, natural language processing, in-depth analysis, knowledge bases.

Авторлар туралы мәлімет / Сведения об авторах / Information about the authors

Асхат Ғайнуллауұы Мұсайбеков – техника ғылымдарының кандидаты, Қазақстан Республикасы ТЖМ Мәлік Ғабдуллин атындағы Азаматтық қорғау академиясының ақпараттық жүйелер мен технологиялар жалпы техникалық пәндер кафедрасының бастығы. Қазақстан, Көкшетау, Ақан Сері к-сі, 136. E-mail: lettermus@mail.ru

Игорь Анатольевич Захаров – техника ғылымдарының кандидаты, Қазақстан Республикасы ТЖМ Мәлік Ғабдуллин атындағы Азаматтық қорғау академиясының ғылыми-зерттеу орталығының бастығы. Қазақстан, Көкшетау, Ақан Сері к-сі, 136. E-mail: emercom.87@mail.ru

Ғабит Әубәкіруұлы Шәріпов – PhD докторы, қауымдастырылған профессор (доцент), Ресей ТЖМ Азаматтық қорғау академиясының адъюнкті. Ресей, Мәскеу облысы, Химки, Новогорск ш/а. E-mail: emersharipovg@mail.ru

Мусайбеков Асхат Гайнуллаулы – кандидат технических наук, начальник кафедры общетехнических дисциплин информационных систем и технологий Академии гражданской защиты им. М. Габдуллина МЧС Республики Казахстан. Казахстан, Кокшетау, ул. Акан Серэ, 136. E-mail: lettermus@mail.ru

Захаров Игорь Анатольевич – кандидат технических наук, начальник научно-исследовательского центра Академии гражданской защиты им. М. Габдуллина МЧС Республики Казахстан. Казахстан, Кокшетау, ул. Акан Серэ, 136. E-mail: emercom.87@mail.ru

Шарипов Габит Аубакирович – доктор PhD, ассоциированный профессор (доцент), адъюнкт Академии гражданской защиты МЧС России. Россия, Московская обл., Химки, мкр. Новогорск. E-mail: emersharipovg@mail.ru

Askhat G. Mussaibekov – Candidate of Technical Sciences, Head of the Department of General Technical Disciplines of Information Systems and Technologies of the Malik Gabdullin Academy of Civil Protection of the MES of the Republic of Kazakhstan. Kazakhstan, Kokshetau, 136 Akan Sere street. E-mail: lettermus@mail.ru

Igor A. Zakharov – Candidate of Technical Sciences, head of the research center of the Malik Gabdullin Academy of Civil Protection of the MES of the Republic of Kazakhstan. Kazakhstan, Kokshetau, 136 Akan Sere street. E-mail: emercom.87@mail.ru

Gabit A. Sharipov – Doctor PhD, Associate Professor of the Academy of Civil Protection of the Ministry of Emergency Situations of Russia. Russian Federation, Moscow region, Khimki, md. Novogorsk. E-mail: emersharipovg@mail.ru