МРНТИ 20.23.17
УДК 004.421

https://doi.org/10.51889/2020-3.1728-7901.39

Д.Р. Рахимова[1], У.Ж. Кенес[1]

[1]Әл-Фараби Қазақ Ұлттық университеті, Алматы қ., Қазақстан

ҚАЗАҚ ТІЛІНЕ АРНАЛҒАН СҰРАҚ-ЖАУАП ЖҮЙЕСІН ЗЕРТТЕУ ЖӘНЕ ӘЗІРЛЕУ

Аңдатпа

Бұл зерттеуде қазақ тіліне арналған сұрақтарға жауаптардың жабық пәндік жүйесі мәселелерін талдау модулі үшін әзірленген әдістер сипатталады және бағаланады. Сұрақтарды талдау, оған не қойылатынын және оған қалай жауап беру керектігін анықтау үшін қажетті ақпаратты алу үшін сұрақтарды талдау сапаны бақылау жүйесінің ең маңызды компоненттерінің бірі болып табылады. Сондықтан, біз сұрақтарды талдауда екі негізгі проблеманың жаңа әдістерін ұсынамыз, атап айтқанда ережелер негізінде және жүйелік жіктеу тәсілін интеграциялауға негізделген жасырын марков моделіне негізделген, фокусты экстракциялау және сұрақ жіктеуіштері, олардың екеуі де мәселедегі сөздер арасындағы тәуелділік қарым-қатынасын пайдаланады. Сондай-ақ осы шешімдерді базалық модельдермен салыстыру келтіріледі. Бұл зерттеу, сондай-ақ осы саладағы одан әрі зерттеу үшін алтын стандартты жинақталған және аннотацияланған деректерді қолмен ұсынады.

Түйін сөздер: сұрақ-жауап жүйесі, табиғи тілді өңдеу, фокусты экстракциялау, ақпараты іздеу.

Аннотация

Д.Р. Рахимова[1], У.Ж. Кенес[1]

[1]Казахский Национальный университет имени аль-Фараби, г. Алматы, Казакстан

ИССЛЕДОВАНИЕ И РАЗРАБОТКА ВОПРОСНО-ОТВЕТНЫХ СИСТЕМ ДЛЯ КАЗАХСКОГО ЯЗЫКА

В данном исследовании описываются и оцениваются методы, разработанные для модуля анализа задач замкнутой предметной системы ответов на вопросы по казахскому языку. Анализ вопросов является одним из наиболее важных компонентов системы контроля качества для получения информации, необходимой для определения того, что задают и как на него ответить. Поэтому мы предлагаем новые методы анализа проблем в двух основных проблемах, а именно на основе правил и скрытой модели Маркова, основанной на интеграции систематической классификации, выделения фокуса и классификаторов вопросов, которые зависят от отношения между словами в проблеме. использует соотношение. Также сравниваются эти решения с базовыми моделями. Это исследование также вручную представляет агрегированные и аннотированные данные золотого стандарта для дальнейших исследований в этой области.

Ключевые слова: система вопросов и ответов, обработка естественного языка, извлечение фокуса, поиск информации.

Abstract

RESEARCH AND DEVELOPMENT OF QUESTION-ANSWER SYSTEMS FOR THE KAZAKH LANGUAGE

Rakhimova D.R.[1], Kenes U.Z.[1]

[1]Kazakh National University named after al-Farabi, Almaty, Kazakhstan

This study describes and evaluates the methods developed for the module for the analysis of problems of the closed subject system of answers to questions for the Kazakh language. Question analysis is one of the most important components of a quality control system to obtain the information needed to determine what is being asked and how to answer it. Therefore, we propose new methods of problem analysis in two main problems, namely rule-based and hidden Markov model based on the integration of systematic classification, focus extraction and question classifiers, both of which depend on the relationship between words in the problem. uses the ratio. It also compares these solutions with basic models. This study also manually presents the gold standard aggregated and annotated data for further research in this area.

Keywords: Question-answer system, Natural language processing, Focus extraction, Information retrieval.

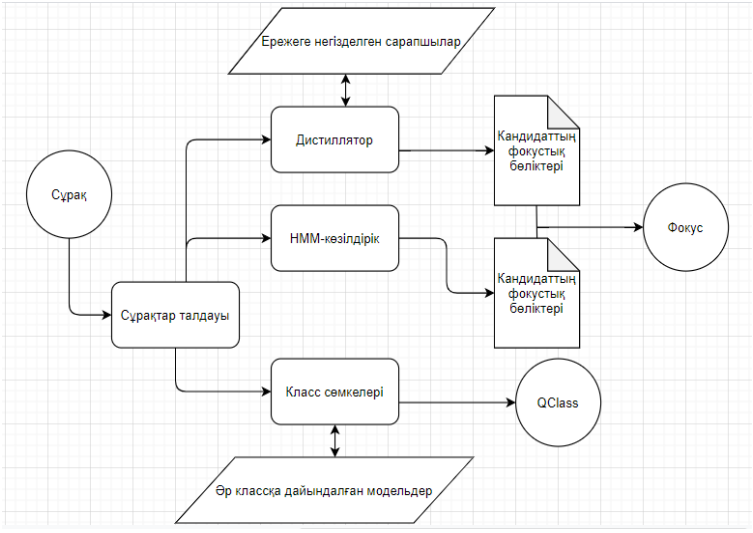
1 Кіріспе

Сұрақтарға жауап беру жүйесі табиғи тілдерде қалыптастырылған сұрақтарға автоматты түрде генерацияланатын жауаптар алуға бағытталған. Соңғы онжылдықта табиғи тілді өңдеу және ақпаратты іздеу әдістерін түбегейлі жетілдіру белгілі сапаны бақылау жүйелерін әзірлеуге әкелді, олардың кейбіреулері AnswerMachine және WolframAlpha сияқты көпшілікке қолдануға қол жетімді. Тіпті адам қарсыластарымен теледидар шоуында бәсекелесе алатын сапаны бақылау жүйесін әзірлеу мүмкін болды [1]. Алайда, сапаны бақылаудың толық әрекетке қабілетті жүйесін құру, негізінен

мәселелерді талдау, ақпаратты іздеу, кросс-лингвистика және жауап генерациясы, сондай-ақ қайта жазу, ақылға қонымды импликация немесе сілтемелерді шешу сияқты неғұрлым төмен деңгейдегі кейбір қосымша берулер сияқты мәселелерді шешу қажет болатын көптеген күрделі берулердің салдарынан қиындықтармен ұштасады. Олардың кейбірі шешілген деп есептелсе де, проблемалардың көпшілігі одан әрі зерттеулер үшін әлі де ашық [2, 3]. Бұл зерттеуде біз бірінші модульді әзірлеу мен бағалауды ұсынамыз, атап айтқанда, біздің жүйеміздің конвейерінде география пәні аймағының прототипінде қолдануға арналған мәселелерді талдау. Сұрақтарды талдаудың негізгі міндеті дұрыс жауапты түпкілікті қалыптастыру үшін келесі модульдерде пайдаланылатын берілген сұрақтан пайдалы ақпарат алу болып табылады. Қазақ тілі морфологиялық бай және деривациялық құрылымы бар агглютинативті тіл. Сол себепті біз морфологиялық талдау және бір мәнді жоюды, сондай-ақ NLP конвейерін пайдалана отырып, тәуелділікті талдауды орындай отырып, алдын ала сұрақтарды өңдейміз [4, 5, 6]. Тәуелділіктерді талдау осы сөйлемдегі сөздер арасындағы тәуелділік қарым-қатынасын жасайды. Тәуелділік талдағышы қолданатын тегтер тәуелділік ағашының қазақ банкінде анықталады, ол субъект, объект, ұсыныс, модификатор, классификатор, иеленуші және т. б. сияқты тегтерді қамтиды [5, 7]. Біз сапаны бақылаудың тұйық жүйесі үшін НММ негізінде тізбектерді жіктеу әдісімен ережелерге негізделген әдісті интеграциялауға негізделген сұрақтарды жіктеуге және фокусты анықтауға жаңа көзқарасты ұсынамыз.

2 Сұрақтарды талдау модулі

Сұрақтарды талдау модулі 1-суретте көрсетілген үш параллель қосалқы модульден тұрады, Дистиллятор, НММ-көзілдірік және класс ережелері (ClassRules). Алғашқы екі модуль сұрақтың фокустарын шығаруға арналған, ал үшінші модуль - сұрақтың алдын-ала анықталған сыныптар тобына (QClass) жіктелуін анықтауға арналған. Фокус берілген сұрақтың нақты не сұрайтындығын және оның қандай түрі екенін көрсетеді. 1-бөлімдегі мысалда назардың негізгі бөлігі осы бөліктердің жиынтығы болып табылады: «қаланың аты» (нақты қаланың аты), өйткені сұрақ атауды сұрайды. Сондықтан «қала аты» сөз тіркесін «қаланың ады» сөзінен синтаксистік тұрғыдан да жасауға болады, өйткені бізде сұрақ бөліктерінде морфологиялық тамырлар бар. Себебі «қала» – түбір, ал «ның» – «қаланың атауы» деген мағына беретін жалғау. Бұл сұраққа арналған QClass – бұл ENTITY (2-кестені қараңыз).



Сурет 1. Сұрақтарды Талдау Модулі "Алғаш рет жерді шарлаған теңізші кім?"

Келесі мысалда назар аударыңыз – «теңізші кім» және Qclass – бұл HUMAN.INDIVIDUAL. Сұрақ адамның аты-жөнін сұрайтын фокустың негіздемесі және адамның теңізші екендігі белгілі. Біз сұрақта мәннің айырмалық қасиеттерін түсіреміз (мысалы, бірінші теңізші), өйткені осы сәтте бізді белгілі бір нысан түрін көрсететін "бар" және "бөлік" қарым-қатынастар қызықтырады. Қалған қасиеттерді жүйенің кейінгі модульдері тиісті білім бірліктерін де, үміткердің жауаптарын да семантикалық кесу үшін пайдаланады.

3 Әдістеме

Фокусты алу үшін бізде географиялық домендегі нақты сұрақтардың барлық түрлері үшін тәуелділік ағаштарына қатысты арнайы ережелері, жылдам басқарылатын фокус экстракторы, ал

дистиллятор және вариацияны қолданатын HMM классификаторы, HMM-көзілдірігі бар. Ветерби алгоритмінің [8] белгілі бір дәрежеде оны дистиллерге қарағанда әлдеқайда либералды етіп көрсетеді.

Қарастырылып отырған сөздер арасындағы тәуелділік қатынастарына әсер ететін бір жалпы белгілерден басқа, олардың негізгі мәселеге деген көзқарастары (яғни, фокусты алу үшін) шешудің әртүрлі деңгейлеріндегі мүлдем басқа принциптерге негізделген. Бұл ерекшелік біздің әдіснамамыз үшін өте маңызды, өйткені ол қазақ тілі сияқты бай туынды құрылымы бар тілдерді тиімді басқаруға қажетті түсінік береді. Осы кезде осы модельдердің үйлесуі үшін нәзік тепе-теңдік қажет. Осы мақсатта біз дистиллятордың да, HMM-көзілдіріктің де жеке мәліметтерімен жаттығулар жиынтығындағы жеке сенімділіктерін ескереміз.

3.1 Фокусты экстракциялау

Дистиллятор. Біз таңдаған география доменімізде көптеген сұрақтарға ортақ (предикатқа негізделген) сұрақ қоюдың белгілі бір заңдылықтары бар екенін байқадық. Біз әрбір осындай үлгіні (сұрақ түрін) анықтадық және әр сұрақтың тәуелділікті талдаудан фокусты алу ережелерін (сарапшыларды) қолмен анықтадық. Бұл ережелер жиынтығын «дистиллятор» деп атаймыз. Қазіргі уақытта бізде жеті ереже бойынша сарапшы бар, сонымен қатар бір жалпы ережені қолдана отырып, сирек кездесетін істерді өңдейтін жалпы сарапшы бар. Жалпы сарапшыны қосудың негізгі себебі - мәліметтер тапшылығы. Алайда, біз мұны міндетті емес етіп таңдағымыз келеді, өйткені белгілі бір жалпы сарапшының және бірқатар сарапшылардың болуы мәліметтер жиынтығының мөлшеріне байланысты аз немесе көбейтілген еске түсірудің орнына айыппұлдың дәлдігіне әкелуі мүмкін. Барлық сарапшылар және олардың жиынтығындағы мәліметтер жиыны 1-кестеде келтірілген. Ережелерде берілген сұрақтың тәуелділік тармағында шарлау туралы нұсқаулар бар. Мысалы, «не» сарапшысына қатысты ереже, ал «берілген» сарапшыға арналған ереже, сондай-ақ жалпы ереже келесідей (2-сурет) келтірілген:

не: (what is…)

- Сұрақ бойынша сөйлем (SENTENCE) алыңыз.
- Субъекттен кері байланыс (traceback) алыңыз және тек иесі (POSSESSOR) және классификатор (CLASSIFIER) жинаңыз.

берілген: (…is given…)

- Сұрақ бойынша сөйлемнен (SENTENCE) субъектті алыңыз.
- Сөйлемнің (SENTENCE) бірінші дәрежелі DATIVE.ADJUNCT ұстап және бақылап, тек бірінші дәрежелі модификаторды (MODIFIER) жинаңыз.

жалпы:

- Сұрақ бойынша сөйлемнен (SENTENCE) субъектті алыңыз.
- Субъекттан бастап бақылап, иесі (POSSESSOR) және / немесе классификатор (CLASSIFIER) бірінші дәрежесін олардың иесі (POSSESSOR) және / немесе классификаторымен (CLASSIFIER) бірге алыңыз. Әрбір ережеге негізделген сарапшының сараптамасына қатысты сұрақтардан дұрыс фокусты бөліп алу үшін оның жұмыс нәтижелеріне негізделген сенімділік деңгейі бар. Бұл балл кейінірек HMM-көзілдірікпен үйлескенде сарапшының пікірінің сенімділігін көрсету үшін қолданылады. Сұрақтың фокустық бөліктерімен қатар, сенімділік көрсеткіштері туралы дистиллятор да, HMM-көзілдіріктер де үштік түрінде хабарлайды:

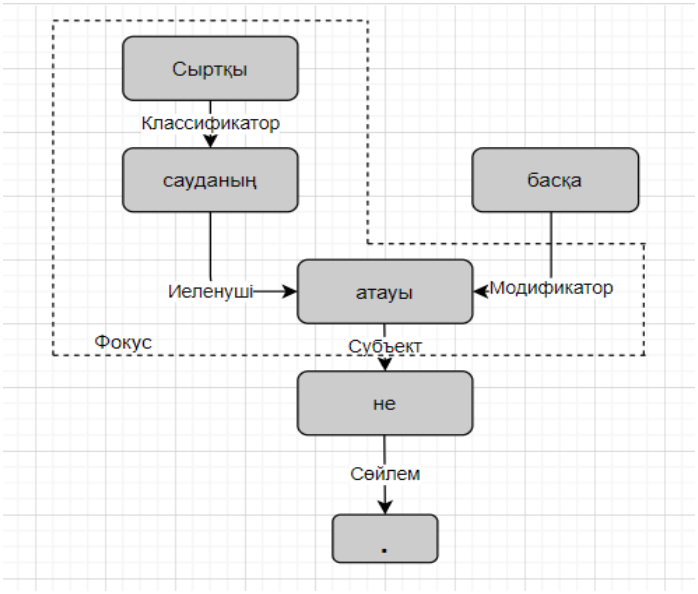
$$\langle fpt, fpd, fpc \rangle_n$$

мұнда $n \in \{1..|Q|\}$, *fpt (focus part text)* фокус бөлігінің мәтіні, *fpd (focus part dependency tag)* фокус тәуелділік тегінің бөлігі және *fpc (focus part confidence score)* фокустық бөліктің сенімділігін білдіреді.

$|Q|$ – сұрағындағы сөздер санын білдіреді.

Кесте 1. Сарапшылар және олардың оқыту деректеріндегі сұрақтар жиілігі

Сараптамалық түрі	*Жиілігі (%)*
жалпы	*25.6*
қайсы	*19.5*
деген не	*15.0*
деп аталады	*9.6*
қанша	*9.6*
беріледі	*7.2*
қайсысы	*7.2*
неше	*6.3*



Сурет 2. Сарапшы "не" сұрағының фокусы "сыртқы сауданың атауы" екенін көрсетіп тұр

Екі модель де фокустың әрбір бөлігі үшін осындай триплеттерді шығарады. Алайда, алынған фокустың әрбір бөлігі үшін құпия деректерді ұсыну тәсілінде ережелер мен статистикалық модельдер арасында айтарлықтай айырмашылық бар. 3.1-бөлімде егжей-тегжейлі түсіндіргендей, HMM-көзілдірік сұрақтың жекелеген бөліктерінде жұмыс істейді, ал дистиллятор под-ағаштарды сұрақтың тәуелділігі ағаштан алады. Сондықтан дистиллятордың шешімі фокустағы әрбір бөлік үшін жеке ықтималдықтарды ескеру үшін жеткіліксіз. Осылайша, дистиллятор фокус ретінде бөлшектерді жинайды, сондай-ақ фокуста барлық бөлшектерді дистиллятор тұрғысынан тең етіп жасай отырып, барлық бөлшектердің f pc баллдарымен салыстырылатын жауапты сарапшы хабарлаған бір сенім баллы (total confidence score).

HMM-көзілдірік. HMM-көзілдірік фокусты алуды HMM (Hidden Markov Model) ретінде моделдейді және Витерби алгоритмін пайдалана отырып сұрақта сөздердің тізбекті жіктеуін орындайды. Тек екі жасырын күйі бар, атап айтқанда FOC (яғни бақыланатын бөлік фокустың бір бөлігі болып табылады) және NON (яғни бақыланатын бөлік фокустың бөлігі болып табылмайды), ол мәселенің әрбір бөлігін қадағалау ретінде қарастырады және бақылаудағы бөлік сұрақ фокусының бір бөлігі болып табыла ма? Алдымен біз сұрақтың тәуелділік ағашын сериализациялаймыз және сериалданған ағашты жібереміз. Ағаштың сериализациясы оның жүйелі түрде көрініс алуы болып табылады, ол негізінен қолданбалы математика, деректер қоры және желілер салаларында қолданылады [9,10]. Әрине, ағаш сериалданатын әдіс алгоритм нәтижелерінің сипаттамаларына елеулі әсер етеді. Біз бұл әсерді сериализацияға екі жалпы тәсіл арқылы зерттейміз және эмпирикалық тексердік (5-бөлімді қараңыз). Ағаш сериализациясының жалпы қабылданған тәсілдері ақпараттық-теориялық ресурстық шекаралар шеңберінде (уақыт және кеңістік терминдерінде) ағашты тиімді сериализациялауға тырысады. Екінші жағынан, бізді тек ағаш тәрізді құрылымның когеренттілігі қызықтырады.

Дистиллятор мен HMM-көзілдіріктердің үйлесімі. Естеріңізге сала кетейік, дистиллятор сарапшының бірыңғай жиынтық сенім балы бар фокустық бөліктерді шығарады. Сонымен қатар, бізде бар HMM-көзілдірік өнімдері:

$$\begin{matrix} \text{HMM} & \text{Дистиллятор} \end{matrix}$$
$$\begin{Bmatrix} \langle fpn_1, fpt_1, fpc_1 \rangle & \langle fpn_1, fpt_1, fpc \rangle \\ \langle fpn_2, fpt_2, fpc_2 \rangle & \langle fpn_2, fpt_2, fpc \rangle \\ \vdots & \vdots \\ \langle fpn_p, fpt_p, fpc_p \rangle & \langle fpn_q, fpt_q, fpc \rangle \end{Bmatrix}$$

Әр түрлі модельдермен жасалатын ықтимал фокустық бөліктерді біріктіру бөліктер бойынша жүзеге асырылады. Басқаша айтқанда, модельдер әрбір бөлігі фокустың соңғы бөліктерінің ішінде екенін бір-біріне сендіруге тырысады. Ол үшін біз fpc ұпайларын пайдаланамыз, оларды оқыту деректері бойынша жеке f модель баллдарымен өлшейміз және максимумын аламыз. Назар аударыңыз, егер деталь әлеуетті фокустық бөлшектер ретінде анықталса ғана M_1 модельдерінің бірі

(яғни, басқа M_2 моделі бұл бөлік фокустың бөлігі емес деп болжайды), содан кейін біз жоғарыда сипатталғандай M_1 сенімді балын есептеп, оны M_2 f балымен салыстырамыз. Егер M_1 сенімді балы M_2-ден артық болса, онда сөз фокустың бөлігі ретінде жіктеледі, әйтпесе ол фокустан шығарылады.

3.2 Класстарды экстракциялау (Class extraction)

Сұрақтарды жіктеу үшін біз қолмен сыныптардың екі түрін анықтадық, атап айтқанда [11, 12] бейімделген, әртүрлі семантикалық рұқсаттары бар дөрекі және жұқа сыныптар. Сұрақтың жұқа сыныбы нақты пәндік саламен күшті байланысты орнатады, ал оның дөрекі сыныбы мәні бойынша жалпылау моделіне енгізеді, ол географиядан басқа басқа салаларда қолданылатын жіктеуді жасайды.

Қазіргі уақытта бізде жеті өрескел сынып (2-кестені қараңыз), сондай-ақ жалпы 57 жұқа сынып бар. Бұл зерттеуде біз тек дөрекі сыныптарда ғана шоғырландық. Біз статистикалық тәсілдерді пайдалана отырып, жұқа сыныптарды топтастыруды жоспарлап отырмыз, бұл әрбір жұқа сыныпта толық сұрақтар санын талап етеді.

Кесте 2. География доменіне арналған дөрекі класстар

Сұрақ классы	*Жиілігі (%)*
Сипаттама	*25,2*
Сандық	*24,2*
Мәні	*19,6*
Уақытша	*12,4*
Орналасуы	*11,9*
Аббревиатура	*3,8*
Адам	*2,4*

Бұл сұрақты өрескел кластардың біріне жіктеу үшін біз осы классқа ғана тән әр класс үшін жалпы тіркестер жиынтығын жасадық. Мысалы, NUMERIC класы үшін бізде екі сөйлем бар: «қанша» және «неше». Классификатор берілген заңдылықтарды берілген сұрақта іздейді және сәйкесінше жіктейді. Біз ережелерге негізделген тәсілмен салыстыру үшін базалық модель ретінде tf-idf негізінде өлшенген «сөздер қабы» стратегиясын пайдаланатын статистикалық жіктеуішті қосымша іске асырамыз. Базалық модельде c сыныбы үшін w сөзінің салмағы келесідей есептеледі:

$$tf\_idf_{w,c} = tf_{w,c} \times idf_w$$

мұнда $tf_{w,c}$ сөз c классында болған кездегі санын көрсетеді, және idf_w төменде көрсетілгендей есептеледі:

$$idf_w = \log \frac{\text{класс саны}}{w \text{ бар класстар саны}}$$

Содан кейін, берілген Q сұраққа біз оны tf-idf ұпайларының суммасын көбейтетін классқа береміз: $argmax_c \sum_{w \in Q} tf\_idf_{w,c}$.

4 Бағалау және нәтижелер

Біз тап болған басты проблемалардың бірі проблеманың нақты қатаңдығы мен біздің шешімдеріміздің нақты тиімділігін көрсету үшін қолайлы базалық желінің болмауы (алдыңғы зерттеулерден және т.б.) болды. Сондықтан біз фокустың бөлігі ретінде белгілі бір жақындығы үшін сұрақтың негізгі сөзімен көршілес сөздерді сәйкестендіретін фокусты алу үшін базалық модельді іске асырдық. Жақындық моделі сәл нашар, бірақ tf.idf моделімен ұқсас нәтижелер. Біз нақты салыстыру үшін ең жақсы нәтижелермен (яғни tf.idf) тек бастапқы деректерді ғана таңдадық. Айта кетейік, бастапқы модельдер қарапайым түрде жасалуы керек, өйткені қазақ тілінде статистикалық сұрақтарды талдауда алдын ала зерттеу жүргізілмеген. Сондықтан, мәселенің төменгі шекараларын белгілеу үшін негіздер қарапайым түрде сақталады. Сонымен қатар, сұрақтар жіктелуі үшін tf-idf негізіндегі статистикалық базалық модель енгізілген, ол сөздер жиынтығы стратегиясын қолданады. Барлық нәтижелер 3 және 4 кестеде келтірілген негізгі модельдермен салыстыру түрінде баяндалған.

Біздің модельдеріміздің негізінде бағаланатын деректер осы зерттеу курсында дайындалған болғандықтан, біз гиген тұжырымдамасының айналасында біздің бағалау стратегиямызды құрудамыз,онда біз екі іргелі қағидатты қамтамасыз етеміз. Біріншіден, кез келген нүктеде және әрбір модель үшін балдар модель бұрын қиылыспаған сұрақтар үшін алынған нәтижеден алынады.

Екіншіден, модельдерді ақылға қонымды салыстыру үшін бір ұпайлар әр түрлі параметрлермен әр түрлі модельдер үшін әр баға итерациясында бір сұрақтарды пайдалана отырып есептеледі. Дистилляторды бағалау үшін, ережелерге негізделген сарапшылар біз басында болған алғашқы 107 мәселені ғана пайдалана отырып әзірленеді.

Фокусты алу үшін түпкілікті нәтижелер (яғни дәлдік, кері қайтарып алу және f-балл) жекелеген нәтижелерді макросреднациялау арқылы алынады. Дистиллятордың әмбебап сарапшыны қосу және өшіру мүмкіндігі бар, ал HMM-Glasses тәуелділік ағашының сериализациясын калибрлейтін алға, артқа және алға-артқа режимдері бар.

Кесте 3. Фокусты алудың барлық модельдерін бағалау нәтижелері

Модель	***Дәлдігі***	***Кері қайтарып алу (recall)***	***F-Бағалау***
Базалық (tf.idf модель)	***0,769***	***0,197***	***0,290***
Дистиллятор (Generic Enabled)	*0,714*	*0,751*	*0,732*
Дистиллятор (Generic Disabled)	*0,816*	*0,623*	*0,706*
HMM-Glasses (Backward Mode)	*0,839*	*0,443*	*0,580*
HMM-Glasses (Forward Mode)	*0,847*	*0,495*	*0,625*
HMM-Glasses (Forward and Backward Mode)	*0,821*	*0,515*	*0,633*
Combined (Generic Enabled, Backward)	*0,734*	*0,841*	*0,784*
Combined (Generic Enabled, Forward)	*0,732*	*0,846*	*0,785*
Combined (Generic Enabled, Forward & Backward)	***0,721***	***0,851***	***0,781***
Combined (Generic Disabled, Backward)	*0,821*	*0,759*	*0,789*
Combined (Generic Disabled, Forward)	*0,818*	*0,765*	*0,791*
Combined (Generic Disabled, Forward & Backward)	***0,802***	***0,776***	***0,788***

Кесте 4. QClass классификациясының нәтижелері. Жоғарғы бөлім - tf-idf негізіндегі модель, төменгі бөлім - ережеге негізделген модель

Класстар	***Дәлдігі***	***Кері қайтарып алу (recall)***	***F-Бағалау***
Сипаттамасы	*0,662*	*0,908*	*0,764*
Уақытша	*0,767*	*0,618*	*0,670*
Сандық	*0,801*	*0,758*	*0,776*
Мәні	*0,100*	*0,025*	*0,040*
Қысқарту	*0,933*	*0,766*	*0,823*
Орналасу	*0,759*	*0,212*	*0,312*
Адам	*0,600*	*0,600*	*0,600*
Tf.Idf жалпы	***0,660***	***0,555***	***0,569***
Сипаттамасы	*0,874*	*0,732*	*0,797*
Уақытша	*1,000*	*1,000*	*1,000*
Сандық	*0,995*	*0,911*	*0,951*
Мәні	*0,603*	*0,817*	*0,694*
Қысқарту	*0,871*	*0,894*	*0,883*
Орналасу	*0,944*	*0,880*	*0,911*
Адам	*0,869*	*0,833*	*0,851*
Rule-based жалпы	***0,879***	***0,867***	***0,869***

Осы параметрлердің барлық әр түрлі комбинациялары әр модель үшін жеке-жеке, сондай-ақ комбинацияда, жинақтау процесінің әр итерациясында жеке бағаланады. Фокусты алу және сұрақтардың жіктелу нәтижелері тиісінше 3 және 4-кестелерде келтірілген.

4.1 Фокусты алу нәтижелері

Дистилляторды жеке бағалау нәтижесінде салыстыру дәлдігі және төменгі қайтару баллдары (аралас модельдермен салыстырғанда) алынды. Дистилляторды бағалаудың маңызды нәтижесі жалпы сарапшының әрекеті болып табылады. Нәтижелер жалпы сарапшының үлгіні (яғни кері қайтаруды)

үлкейту кезінде алынған нәтижелердің дәлдігін төмендететінін көрсетеді (яғни дәлдік). Алайда, екі нәтиже өтемейді, өйткені алынған нәтижелер көрсеткендей, жалпы сарапшы қосылған дистиллятордың f-баллы жалпы сарапшы өшірілгенге қарағанда жоғары.

Сериялау әдістерінің әсерін жеке бағалау алға және артқа өту режимдерінде f-баллдарды ескере отырып, кері режимге қарағанда сәл жақсы екенін көрсетеді. Кері режим, шамасы, ол қосылған кез келген модельдің кері әсерін арттырады, алайда f-баллдар бұл кері қайтарып алуды арттыру пайдалы емес екенін көрсетеді, өйткені ол қосылған кезде аралас модельдердің өнімділігін төмендетеді.

4.2 Класс ережелері нәтижелері

Нәтижелер көрсетіп отырғандай, пәндік саладағы білімді пайдалану статистикалық базалық модель жақындай алмаған елеулі табысқа әкелді. Дегенмен, қолмен жасалған ережелер жиынтығы - бұл доменді өзгерту кезінде үлкен проблема. Сондықтан, осы домендерге қатысты тіркестерді автоматты түрде үйренетін статистикалық оқуды әрі қарай дамыту жоспарлануда, өйткені әр класс үшін көптеген инстанциялар қажет. Бұл тапшылық болашақ оқуға арналған жақсы класстардың анықтамасын қалдыруға себеп болып табылады. 4-кестеде tf-idf негізіндегі жіктеудің нәтижелерімен бірге ережеге негізделген классификатордың өрескел класты сәйкестендірудің макро-дәлдігі, кері қайтарып алу және f-баллдары көрсетілген.

5 Қорытынды

Осы зерттеуде біз қазақ тілі сияқты агглютинативті тілге арналған жабық домендік сұрақтарға жауап беру жүйесінде қолданылатын сұрақ талдауға ережелік және статистикалық тәсілдердің жаңа үйлесімін ұсындық. Сұрақтарды талдау фокусты бөлу және сұрақтарды жіктеуден тұрады. Фокусты алу үшін бізде қазақ тілінде жиі кездесетін сұрақтарға арналған ережелерге негізделген бірнеше сарапшылар бар. Сонымен қатар, біз НММ негізіндегі романның дәйектілік классификациясының тәсілін сипаттадық, сонымен қатар әр модельдің жеке сенімділік көрсеткіштері бойынша ережелер мен статистикалық модельдердің нәтижелерін біріктірдік. Сұрақтарды жіктеу үшін біз ережелерге негізделген классификаторды қолдандық, ол әр классқа сәйкес емес сөз тіркестерін қолданады. Біз екі мәселе үшін де базалық модельдерді қолдандық және салыстыру туралы осында баяндадық. Ұсынылған әдістемеден басқа біз репродуктивтілікке және кейінгі зерттеулерге арналған қолмен жазылған сұрақтар жиынтығын ұсынамыз.

Пайдаланылған әдебиеттер тізімі:

1 Ferrucci D.A.: Introduction to “this is watson”. IBM Journal of Research and Development 56, 1–15 (2012)

2 Gupta P., Gupta V.: A survey of text question answering techniques. International Journal of Computer Applications 53, 1–8 (2012)

3 Allam A.M.N., Haggag, M.H.: The question answering systems: A survey. International Journal of Research and Reviews in Information Sciences (IJRRIS) 2 (2012)

4 Şahin M., Sulubacak U., Eryiğit, G.: Redefinition of turkish morphology using flag diacritics. In: Proceedings of The Tenth Symposium on Natural Language Processing (SNLP 2013) (2013)

5 Eryiğit G.: The impact of automatic morphological analysis & disambiguation on dependency parsing of turkish. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC), Istanbul, Turkey (2012)

6 Nivre J., Hall, J., Nilsson, J., Chanev, A., Eryiğit, G., Kübler, S., Marinov, S., Marsi, E.: Maltparser: A language-independent system for data-driven dependency parsing. Natural Language Engineering Journal 13, 99–135 (2007)

7 Eryiğit G., Nivre, J., Oflazer, K.: Dependency parsing of turkish. Computational Linguistics 34, 357–389 (2008)

8 Viterbi A.: Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Transactions on Information Theory 13 (1967)

9 Wen L., Amagasa, T., Kitagawa, H.: An approach for XML similarity join using tree serialization. In: Haritsa, J.R., Kotagiri, R., Pudi, V. (eds.) DASFAA 2008. LNCS, vol. 4947, pp. 562–570. Springer, Heidelberg (2008)

10 Munro J.I., Raman, V.: Succinct representation of balanced parentheses and static trees. SIAM J. Comput. 31, 762 –776 (2002)

11 Li, X., Roth, D.: Learning question classifiers: the role of semantic information. Natural Language Engineering 12, 229–249 (2006)

12 Metzler D., Croft, B.W.: Analysis of statistical question classification for fact-based questions. Information Retrieval 8, 481–504 (2005)

References

1 Ferrucci D.A. (2012) Introduction to "this is watson". IBM Journal of Research and Development 56, 1–15. (In English)

2 Gupta P., Gupta V. (2012) A survey of text question answering techniques. International Journal of Computer Applications 53, 1–8. (In English)

3 Allam A.M.N., Haggag, M.H. (2012) The question answering systems: A survey. International Journal of Research and Reviews in Information Sciences (IJRRIS). (In English)

4 Şahin M., Sulubacak U., Eryiğit, G. (2013) Redefinition of turkish morphology using flag diacritics. In: Proceedings of The Tenth Symposium on Natural Language Processing (SNLP 2013). (In English)

5 Eryiğit G. (2012) The impact of automatic morphological analysis & disambiguation on dependency parsing of turkish. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC), Istanbul, Turkey. (In English)

6 Nivre J., Hall, J., Nilsson, J., Chanev, A., Eryiğit, G., Kübler, S., Marinov, S., Marsi, E. (2007) Maltparser, A language-independent system for data-driven dependency parsing. Natural Language Engineering Journal 13, 99–135. (In English)

7 Eryiğit G., Nivre, J., Oflazer, K. (2008) Dependency parsing of turkish. Computational Linguistics 34, 357–389. (In English)

8 Viterbi A. (1967) Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Transactions on Information Theory 13 (In English)

9 Wen L., Amagasa, T., Kitagawa, H. (2008) An approach for XML similarity join using tree serialization. In: Haritsa, J.R., Kotagiri, R., Pudi, V. (eds.) DASFAA 2008. LNCS, vol. 4947, pp. 562–570. Springer, Heidelberg. (In English)

10 Munro J.I., Raman, V. (2002) Succinct representation of balanced parentheses and static trees. SIAM J. Comput. 31, 762 –77. (In English)

11 Li, X., Roth, D. (2006) Learning question classifiers: the role of semantic information. Natural Language Engineering 12, 229–249. (In English)

12 Metzler D., Croft, B.W. (2005) Analysis of statistical question classification for fact-based questions. Information Retrieval 8, 481–504. (In English)