Универзитет „Св. Кирил и Методиј“ – Скопје

Факултет за информатички науки и компјутерско инженерство

ДОКТОРСКА ДИСЕРТАЦИЈА

СИСТЕМ ЗА ОДГОВАРАЊЕ ПРАШАЊА СО ПОВЕЌЕКРАТЕН ИЗБОР ЗА ТЕСТ-КОЛЕКЦИИ НА МАКЕДОНСКИ И АНГЛИСКИ ЈАЗИК

ЈАСМИНА С. ЈОВАНОВСКА

Скопје 2017

Тема:	Систем за одговарање прашања со повеќекратен избор за тест-колекции на македонски и англиски јазик
Автор:	Јасмина С. Јовановска
Научна област:	Обработка на природните јазици
Ментор:	проф. д-р Катерина Здравкова, редовен професор на Факултетот за информатички науки и компјутерско инженерство, Скопје

Членови на комисијата:	проф. д-р Катерина Здравкова, редовен професор на Факултетот за информатички науки и компјутерско инженерство, Скопје проф. д-р Жанета Попеска, редовен професор на Факултетот за информатички науки и компјутерско инженерство, Скопје д-р Андреа Кулаков, вонреден професор на Факултетот за информатички науки и компјутерско инженерство, Скопје д-р Елизабета Бандиловска, вонреден професор на Педагошкиот факултет „Св. Климент Охридски“ - Скопје д-р Соња Гиевска Крлиу, вонреден професор на Факултетот за информатички науки и компјутерско инженерство, Скопје

Датум на одбрана: _____

БЛАГОДАРНОСТ

Подготовката на докторската дисертација бараше голем напор и заложба изминатиот период. Во текот на работата од непроценлива помош и вредност ми беа советите и сугестиите на мојата почитувана менторка, проф. д-р Катерина Здравкова, за што и изразувам најискрена благодарност.

Длабоко признание оддавам на Ивана Божинова за нејзиниот исцрпувачки ангажман околу практичната реализација на сите идеи и размисли во ова истражување.

Истовремено искажувам и искрено признание за мојата фамилија која покажа огромно трпение и ме мотивираше да се посветам на обврските околу дисертацијата. Им изразувам благодарност на сите кои беа со мене во овој благороден зафат. Нека биде за општо добро.

Јасмина Јовановска

СИСТЕМ ЗА ОДГОВАРАЊЕ ПРАШАЊА СО ПОВЕЌЕКРАТЕН ИЗБОР ЗА ТЕСТ-КОЛЕКЦИИ НА МАКЕДОНСКИ И АНГЛИСКИ ЈАЗИК

РЕЗИМЕ:

Технологијата за одговарање прашања (*Question Answering – QA*) претставува мошне активно поле за истражување, која денес се стреми да ги задоволи различните побарувања на реалните корисници. За таа цел, градењето на модерен систем за одговарање прашања подразбира здружување на различни инженерски решенија од повеќе области, меѓу кои најзначајни се: прибирањето информации, обработката на природните јазици, вештачката интелигенција и машинското учење. Комбинирањето на техниките од овие области, во насока на креирање оптимален систем за одговарање прашања е исклучително предизвикувачка задача. Дел од потешкотиите во овој процес произлегуваат и од спецификите на природниот јазик на кој е поставено прашањето, како и јазикот на кој се напишани документите каде се пребарува одговорот. За негова успешна реализација, неопходно е и постоење на анотиран корпус за извлекување на знаењата од областа на која се однесува системот, располагање со лексички бази на податоци, различни алатки и пристапи кои се специфични за конкретниот природен јазик.

Целта на оваа докторска дисертација е градењето на систем за одговарање прашања со повеќекратен избор на македонски јазик. Заради непостоење на тест-колекција на македонски јазик, истражувањето, пред сè, ја потенцира неопходноста од креирање на таква колекција која мора ги задоволи стандардните протоколи за превалидација и поствалидација. Само таква колекција може да даде реална слика за постигнатите резултати од примената на веќе постојните, но и новите методи за утврдување на точниот одговор на дадено прашање, поставено на природен јазик. Сите испитувања во ова истражување се направени врз преземена тест-колекција од областа филозофија, претпроцесирана за да ги задоволи стандардните протоколи за превалидација.

Во насока на градење успешен систем за одговарање прашања на македонски јазик, истражувањето се фокусира на откривање на морфолошките белези на македонскиот јазик кои имаат силно влијание во процесот на прибирањето информации (одговарањето прашања). Поточно, се прави обид да се утврди важноста на информацијата за припадност на зборовите во одредена зборовна група и како оваа информација може да се искористи за подобрување на резултатите во пребарувањето. Исто така, се прави и обид да се оцени квалитетот на прибраните резултати, доколку се пребарува користејќи ги само зборовите од прашалникот, сите зборформи на зборовите од прашалникот, како и сите зборови кои имаат ист основен збор (*stem*) како и збор од прашалникот (а кои се појавуваат во речникот на колекцијата).

Истражувањето ја потенцира и важноста на близината меѓу зборовите од прашањето и зборовите од точниот одговор, во процесот на креирање стратегии за селекција на еден од четирите понудени одговори за дадено прашање. За таа цел, во дизајнот на *QA*-системот е вклучена *Hanning*-прозорската функција.

Конечните резултати потврдуваат дека исклучително позитивно влијание во процесот на одговарање прашања на македонски јазик има вклучувањето на зборформите на зборовите од прашањето (прашалникот), како и земањето предвид на близината меѓу зборовите од прашањето и понудениот (генерираниот) одговор. Со оглед на фактот дека не постои лематизатор за македонскиот јазик, истражувањето применува статистички пристап за групирање на зборовите од речникот генериран од тест-колекцијата „Филозофија“, кои припаѓаат на иста лексема. За таа цел, дефинирана е нова метрика за сличност на стрингови базирана на триаголниот прозорец. Примената на групите генерирани со оваа метрика во процесот на прибирањето информации (одговарањето прашања), дава подобри резултати отколку примената на точните групи од зборови (креирани рачно). Што се однесува до информацијата за зборовната група, резултатите потврдуваат дека оваа карактеристика не е доминантна (но има влијание) во процесот на прибирањето информации (одговарањето прашања).

Дизајнираниот систем е тестиран на две дополнителни колекции од областа информатички технологии, напишани на два различни јазика: македонски и англиски. Деталната анализа на постигнувањата потврдува дека системот успешно може да се користи и за одговарање прашања на македонски јазик од други области. Испитувањето потврдува успешност и при одговарање на прашањата од англиската тест-колекција. Заклучоците изведени од севкупните резултати отвораат простор за натамошни идни подобрувања на *QA*-системот, преку вклучување на синтаксичките белези на македонскиот јазик.

КЛУЧНИ ЗБОРОВИ:

Прибирање информации, одговарање прашања, обработка на природните јазици, зборовна група, зборформа, метрики за сличност на стрингови, прозорски функции.

Jasmina Jovanovska

MULTIPLE-CHOICE QUESTION ANSWERING SYSTEM FOR MACEDONIAN AND ENGLISH TEST-COLLECTIONS

ABSTRACT:

Question Answering (QA) is still a very active field for research, aiming to meet different demands of the actual users. The design of a modern question answering system requires incorporation of different engineering solutions from several areas. The most important are: information retrieval, natural language processing, artificial intelligence and machine learning. This process of combining techniques in order to build the most advantageous QA system is a very challenging task. Some of the difficulties arise from the specific properties of the natural language used to pose a particular question and the incorporated language/languages, where the answer is searched. Therefore, the successful realization of the QA system design for particular natural language requires an annotated corpus for knowledge extraction, lexical databases, as well as different tools and approaches specified for that language.

The main objective of this doctoral thesis is the development of a multiple-choice question answering system for Macedonian and English test-collections. In absence of a test-collection for Macedonian language, this research emphasizes the necessity of creating a test-collection satisfying the standard prevalidation and postvalidation protocols, which are crucial to increase questions versatility, reliability and validity. Such a collection can be used to draw reliable conclusions from the results achieved using the existing, as well as, the new methods included in the process of question answering. For all the experiments in this dissertation, a test-collection from the field of philosophy is used. It fully satisfies the standard prevalidation protocols.

In order to build a successful QA system for Macedonian language, the emphasis of this research is on the identification of morphological features of Macedonian language, which have a strong influence in the process of retrieving information and question answering. In particular, the first attempt is to determine the importance of the word class information of the query words, and how this information can be used to improve the retrieved results. In addition, three different strategies are tested for information retrieval: using only the query words, all the word forms of a query word, as well as, all the words that have the same stem with a query word (from the collection dictionary). This research also emphasizes the importance of word proximity in the process of finding the correct answer to a particular question. Therefore, the designed QA system includes a window function.

The final results confirm the positive influence of incorporating word forms in the process of answering questions posed in Macedonian language. Moreover, the implementation of the Hanning window function further improves system accuracy. Considering the fact of absence of lemmatizer for Macedonian, this research uses a statistical approach for grouping words from the collection dictionary, which belong to a same lexeme.

For that purpose, a new metric for string similarity based on Triangular window is defined. Including the groups generated by this metric in the process of information retrieval (question answering) results with better system accuracy then using the manually created groups of word forms. Concerning the importance of the word class information, the final results confirm that this word characteristic is not dominant (though is influential) in the process of information retrieval (as well as question answering).

The designed QA system is tested on two additional test-collections from the field of information technology, written in two different languages: Macedonian and English. The detailed analysis of the overall achievements confirms that the system can be successfully used for answering question posed in Macedonian language from other fields. Furthermore, the results for the English collection show that it can be effectively used for answering question posed in English language, too. The conclusions pave the way to future system improvements including the syntactic features of Macedonian language.

KEY WORDS:

Information retrieval, question answering, natural language processing, word class, word form, string similarity metrics, window functions.

СОДРЖИНА:

ТАБЕЛИ:

СЛИКИ:

КРАТЕНКИ:

CLEF – Cross-Language Evaluation Forum
CNN – Convolutional Neural Networks
CRF – Conditional Random Fields
DD – Density Distribution
ELM – Extreme Learning Mashine
ESG – English Slot Grammar
HMM – Hidden Markov Model
IE – Information Extraction
IR – Information Retrieval
KLD language model – Kullback-Leibler Divergence language model
LAT – Lexical Answer Types
LCA – Local Context Analysis
LSI – Latent Semantic Indexing
MAP – Mean Average Precision
MLN – Markov Logic Network
MRR – Mean Reciprocal Rank
NIST – National Institute of Standards and Technology
NLIDB – Natural Language Interfaces to Databases
NLP – Natural Language Processing
NTCIR – NII Test Collections for IR Systems
PAS builder – Predicate-Argument Structure builder
PLM – Positional Language Model
PLM – Proximity Language Model
QA – Question Answering
QASs – Question Answering Systems
QC – Question Classification
SVM – Support Vector Machines
TAC – Text Analysis Conference
TREC – Text REtrieval Conference

1. Вовед

„Учи од вчера, живеј денес, надевај се за утре. Она што е важно, е никогаш да не престанеш да трагаш.“

Albert Einstein

„Одговарањето прашања“ (*Question Answering – QA*) го привлекува интересот на човекот низ вековите. Особено забележлив е интересот на старите Грци кои со познатото Сократово испрашување потврдуваат дека само длабоко, систематско и сеопфатно испрашување може да ја открие вистината или веродостојноста на нештата. Многу подоцна, потрагата по вистината и знаењето го натераа човекот да се обиде да пронајде начини за автоматско справување со изобилието достапни информации. Во таа насока се и дел од напорите на првите научници од областа вештачка интелигенција, кои се обидуваат да дизајнираат компјутерска програма способна да одговара прашања.

Од своите почетоци, па сè до денес, **системите за одговарање прашања** (*Question Answering Systems – QASs*) се развиваат во повеќе насоки:

- **Од системи со затворен домен до системи со отворен домен**. Првите *QA*-системи, како добро познатите *Baseball* [1] и *Lunar* [2], претставуваат само интерфејс кон структурираните бази на податоци. Прашањата од корисниците ги анализираат користејќи техники за **обработка на природните јазици** (*Natural Language Processing – NLP*), со цел да создадат канонична форма која потоа се користи за конструкција на стандарден прашалник кон базата на податоци. Карактеристично за овие системи е нивното ограничено знаење, односно можноста да одговорат на прашања кои се однесуваат на конкретен домен. Сепак, развојот на Интернетот овозможи огромно количество текстуални документи да им бидат достапни на корисниците. Тоа ја наметна потребата од развој на системи кои се отворени скоро за секакви прашања.
- **Од системи базирани на текст до системи оспособени за говор**. Денес *QA*-системите може да располагаат со интерфејс за говор, пришто прашањата се запишуваат користејќи софтвер за препознавање на говорот (*speech recognition software*), додека одговорите се изразуваат преку софтвер за синтеза на говорот (*speech synthesis software*). Ова особено е популарно кај новите мобилни уреди.
- **Од системи независни од контекстот до контекстуални системи**. Првите *QA*-системи одговарале прашања кои исклучиво можеле да бидат разбрани независно од претходниот контекст (сами за себе). Денешните *QA*-системи имаат можност (сепак ограничена) за разбирање на искази кои се однесуваат на претходно поставени прашања и нивните одговори.

Најрепрезентативна демонстрација за изненадувачкиот напредок во *QA*-областа претставува развојот на *IBM Watson*-системот за одговарање прашања (секција 2.3.4.1) [3]. Овој систем успева во реално време да ги победи најдобрите натпреварувачи на познатиот американски телевизиски квиз *Jeopardy*. Неговиот успех го наметна

прашањето дали *IBM Watson* навистина е способен да размислува и што останува понатаму да се истражува во областите вклучени во неговиот развој, како: **вештачката интелигенција** (*Artificial Intelligence – AI*), **прибирањето информации** (*Information Retrieval – IR*) и **обработката на природните јазици** (*Natural Language Processing - NLP*). Сепак, како обучен донесувач на одлуки, *IBM Watson* има способност за расудување, но далеку од нивото на интелектуалните процеси во човековиот мозок. Затоа со сигурност може да се каже дека овој систем не може да ги запре (и навистина не ги попречи) идните истражувања во *QA*-областа, исто како што пребарувачот *Google* не ги запре истражувањата во областа на прибирањето информации.

Кај мобилните уреди, одговарањето прашања денес широко се применува како надополнување на пребарувањето. Неколку такви мошне познати комерцијални системи се: *Siri*, *Evi* и *Cortana*. Пред сè, тие се фокусирани на корисникот, односно го користат искуството од персонифицираните пребарувања за да одговорат на неговото барање.

- *Siri* претставува вграден „интелигентен асистент“ кој им овозможува на корисниците на *Apple* уредите (*iPhone*, *iWatch*, *iPad* и *iPod*) да дадат говорни команди, со цел да ракуваат со својот уред и неговите апликации. Командите може да се однесуваат на праќање пораки, поставување потсетници, ракување со *iTunes*, и слично. Прашањата поставени од корисниците, *Siri* ги одговара со синтеза на постоечките податоци во уредот (како податоците од контактите, календарот, и слично), но и надворешни податоци (на пример, давање препораки за филмови, ресторани и слично, врз основа на веќе дадените оценки од други корисници). Компанијата *Apple*[1] потврди дека за пребарување на веб, *Siri* го користи пребарувачот *Google*, со кој од неодамна е заменет пребарувачот *Bing*.
- *Evi* е систем за одговарање прашања кој исто така поддржува говор, достапен за *iPhone* и *Android* апликациите. Одговара прашања користејќи ги базите со знаење и семантичката технологија за пребарување на *True Knowledge Ltd*. Интересно е тоа што успева да одговори прашања кои бараат прилично длабоко резонирање, како „Кој бил претседател кога кралицата Елизабета II била тинејџер?“
- *Cortana* е „виртуелен асистент“ развиен од *Microsoft*, кој како и *Siri*, одговара прашања поставени на природен јазик и може да изврши различни организациски задачи за корисникот. На *Windows* паметните телефони, *Cortana* може да иницира телефонски повици, да испраќа и чита текстуални пораки и да одговора прашања користејќи го пребарувачот *Bing*.

Она што во иднина може да се очекува од системите за одговарање прашања е тие да бидат способни за одговарање разновидни прашања, извлекување одговори од различни извори, обезбедување повисока оправданост за одговорите и можност да ги задоволат потребите на широк спектар корисници со различно искуство, возраст, ниво

[1] http://www.telegraph.co.uk/technology/2017/09/25/google-replaces-bing-become-apples-default-siri-search-engine/

на експертиза и интерес. Сепак, утврдувањето на иднината на овие системи е исклучително предизвикувачка задача заради интердисциплинарната природа на *QA*-областа (поглавје 2). Предвидувањето подразбира предвидување на иднината на неколку дисциплини, кои би можеле меѓусебно да влијаат на непредвидлив начин.

1.1. Мотивација и цели на истражувањето

Одговарањето прашања (*QA*) е во силна интеракција со обработката на природните јазици (*NLP*) и вештачката интелигенција (*AI*). Типичен пример за оваа поврзаност претставуваат истражувањата кои денес активно се вршат за конструирање роботи, способни да го заменат човекот во производствената дејност, а во блиска иднина можеби и во некои пософистицирани задачи. Во таа насока, интересен пример е неодамна претставениот робот *Todai*[2] од страна на Националниот институт за информатика во Токио (*National Institute of Informatics*), развиен за да го положи приемниот испит на Универзитетот во Токио. Резултатите потврдуваат дека *Todai* постигнува подобар резултат од 80% од идните студенти, и тоа за областите: математика, англиски јазик, природни науки и пишување есеј со околу 600 зборови. Сепак, како што потенцираат и самите креатори на *Todai*, човекот сè уште е ненадминлив во задачи како препознавањето облици, реализирањето креативни проекти, решавањето проблеми, читањето и разбирањето.

Три значајни кампањи кои го стимулираа развојот на *QA*-системите низ годините, се следниве:

- ***TREC QA*-работилницата**[3] (*Text REtrieval Conference QA Track*). Во периодот од 1999 до 2007 година, *TREC QA*-работилницата ја промовира задачата за одговарање прашања, при што главниот акцент е ставен на кратките, фактовидни прашања. Според сегашните стандарди, во тој период *TREC QA*-работилницата користи мал корпус документи и ограничен број категории на прашања (најголем дел од прашањата се фактовидни, но подоцна е вклучено и одговарање на прашањата со набројување и дефиницискиите прашања). По повеќегодишна пауза, *TREC* за првпат ја организира *LiveQA*-работилницата во 2015-тата година. Од таа година, работилницата се фокусира на одговарање прашања во реално време и тоа прашања кои се поставени од реални корисници (секција 2.4.2).
- ***CLEF QA*-работилницата**[4] (*Cross-Language Evaluation Forum QA Track*). Оваа работилница започнува да се реализира во 2000-та година. Нејзината цел низ годините е промовирање и развој на системите за одговарање прашања способни да пронајдат информации низ документи напишани на неколку европски јазици. За оваа цел работилницата обезбедува инфраструктура за тестирање на *QA*-системите, креирање на форум за размена на идеи, методи и методологии, како и креирање на соодветни тест-колекции.

[2] http://21robot.org
[3] http://trec.nist.gov/data/qa.html
[4] www.clef-initiative.eu/

- **Квиз шоуто *Jeopardy***[5] (1964 - 1975, 1984 - до денес). Ова американско шоу опфаќа прашања од повеќе области (наречени категории на прашања), при што за одговарање на дел од нив неопходно е да се изведат заклучоци, да се открие играта на зборови или искажувањето ирониja. Пример на прашање кое припаѓа во категоријата „наука“, е следново: „Погоден од електрони, фосфорот испушта електромагнетна енергија во оваа форма.“[6]

Овие настани обезбедија вредни извори неопходни за истражувањата во *QA*-областа и мотивираа развој на различни техники. Притоа, може да се потенцира дека *CLEF QA*-работилницата се истакнува со својата специфичност, бидејќи креирањето на систем кој успева да пребарува низ документи напишани на различни природни јазици е неверојатно предизвикувачка задача. Причината за тоа е потребата системот да ги земе предвид специфичните карактеристики на јазикот на кој е напишан документот подложен на пребарување.

Главната мотивација да се пристапи кон истражување во областа одговарање прашања (секако, освен нејзината занимливост), е желбата да се откријат спецификите на македонскиот јазик кои силно влијаат врз процесот на прибирањето информации. Фокусот е ставен на утврдувањето на морфолошките белези на македонскиот јазик кои можат да го подобрат процесот на прибирање информации од документи напишани на македонски јазик (следствено на тоа и процесот на одговарање прашања). Испитувањата се вршат врз тест-колекција на македонски јазик од областа филозофија. Истражувањето, пред сè, прави споредба на резултатите добиени од прибирањето информации од оваа колекција, доколку во овој процес се користат:

- само зборовите од прашалникот,
- сите зборформи на збор од прашалникот кои се вклучени во речникот генериран од колекцијата документи, и
- сите зборови кои имаат ист **основен збор**[7]**,** таканаречен **стем** (*stem*), со збор од прашалникот, и се вклучени во речникот генериран од колекцијата документи.

За македонскиот јазик не постои ниту **лематизатор**[8] (*lemmatizer*), ниту **стемер**[9] (*stemmer*) кои би можеле да се искористат за групирање на зборовите од одреден корпус документи, во согласност со тоа дали припаѓаат во иста лексема (т.е. се претставени со иста лема, ист заглавен збор во речникот на конкретниот јазик), односно дали имаат ист **основен збор**. Од друга страна, рачното утврдување на овие групи од зборови е исклучително макотрпна задача, која побарува и стручна експертиза. Токму затоа, во ова истражување се дефинира нова метрика за автоматско групирање на зборовите претставени со иста лема (заглавен збор во речникот на еден јазик) и го оценува влијанието на вака дефинираните групи во процесот на прибирање информации.

[5] https://www.jeopardy.com/games/
[6] Одговор: Светлина (или фотони)
[7] Збор од кој се изведуваат други зборови со исто или блиско значење (секција 3.3).
[8] Алгоритам за утврдување на речничката форма на зборот.
[9] Алгоритам за сведување на зборот на **основен збор**.

Истражувањето ја испитува и важноста од познавањето на зборовната група на зборовите кои се користат во прибирањето, заради перцепцијата дека истата може до одреден степен да укаже на присуство или отсуство на информативна содржина во одреден јазик. Интуицијата е дека задавањето различна важност на различните зборовни групи во македонскиот јазик, влијае врз квалитетот на прибраните резултати. За таа цел, истражувањето утврдува која е оптималната разлика во важноста на зборовните групи, и тоа: именките, глаголите, придавките, предлозите и броевите.

Следното прашање на опсервација е неопходноста од проширување на прашалникот, во случај кога содржи збор кој не е застапен во речникот на колекцијата (наречен **непознат збор**). За оваа цел, искористена е новодефинираната метрика спомната претходно, односно испитана е можноста дали оваа метрика може да се вклучи во процесот на успешно пронаоѓање на морфолошки **сличен** збор од речникот (односно, збор кој се сведува на ист основен збор (*stem*) како и непознатиот збор).

На крај, истражувањето се фокусира и на важноста од близината (во документите) меѓу низата зборови која го индицира точниот одговор и клучните зборови извлечени од прашањето, кај системите за одговарање прашања. Во таа насока, последниот модул за селекција на точниот одговор на дадено прашање поставено на македонски јазик, имплементира прозорска функција. Целта од вклучувањето ваква функција е задавање поголеми тежини на термините (зборовите) од прашањето и одговорот, кои се наоѓаат блиску меѓу себе во одредени текстуални сегменти од документите.

Анализите направени врз добиените резултати се искористени за давање одредени препораки кои би требало да се следат во иднина, заради успешно прибирање информации од документи напишани на македонски јазик (како и одговарање прашања поставени на македонски јазик).

1.2. Истражувачки прашања и методологија на истражување

Главните прашања разработени во ова истражување се следниве:

1. Кои се морфолошките белези на македонскиот јазик кои го подобруваат процесот на прибирање информации од документи напишани на македонски јазик? Поточно, колкава е важноста од вклучувањето на зборформите на збор од прашалникот (како и зборовите со кои споделува ист основен збор (*stem*)) во процесот на прибирање информации, и дали познавањето на зборовната група на зборот може да ја подобри точноста на системот (и на кој начин)?
2. Кои се предусловите за дефинирање метрика за сличност на стрингови неопходна за автоматско групирање на зборовите од речникот (генериран од одредена колекција документи на македонски јазик), во согласност со тоа дали припаѓаат на иста лексема? Дали оваа метрика може успешно да се користи и за утврдување на **сличен** збор за збор од прашалникот кој не се содржи во речникот? Односно, во која мера селектира збор од речникот, кој се сведува на ист основен збор (*stem*) како и непознатиот збор?

3. Дали земањето предвид на близината меѓу зборовите од прашањето и одговорот може да ја подобри точноста на системот за одговарање прашања на македонски јазик?
4. Дали резултатите добиени од трите претходно наведени истражувачки прашања, можат успешно да се применат и врз тест-колекција за одговарање прашања на македонски јазик од друга област, како и тест-колекција на англиски јазик (доколку истите се позитивни)?

Општата претпоставка е дека примената на наведените морфолошки белези на македонскиот јазик и вклучувањето на близината меѓу термините, ќе дадат позитивни, прифатливи и соодветни резултати во процесот на одговарање прашања на македонски јазик. Следната појдовна претпоставка е дека оваа проблематика не е проучувана кога станува збор за прибирање информации од содржини на македонски јазик. При прецизирањето на предметот на проучување, определен е и обемот на содржините врз кој се извршува истражувањето. Исто така, опишани се и две тематски слични колекции на македонски и англиски јазик, врз кои ќе се применат постигнувањата во истражувањето.

За да се одговори на горенаведените прашања и претпоставки, се користи следнава методологија на истражување:

1. Дизајн на систем за одговарање прашања во кој е имплементиран моделот на векторски простор со различни метрики за сличност на векторите, искористен за прибирање пасуси, каде **најверојатно** се наоѓа одговорот на поставеното прашање на македонски јазик. За да се утврди значајноста на зборовните групи од македонскиот јазик, се применува методот на експериментирање (тестирани се различни вредности на тежините за различните зборовни групи на термините од документите и прашањата).
2. Имплементација на инвертираниот индекс кој овозможува да се земе предвид близината на зборовите од прашалникот во документите.
3. Примена на методите на анализа и синтеза, за давање одредени препораки кои би требало да се следат во иднина, заради успешно прибирање информации од документи напишани на македонски јазик.

1.3. Структура на трудот

Овој истражувачки труд е поделен на шест поглавја и содржи шестнаесет додатоци.

Првото поглавје ја потенцира актуелноста на темата „Одговарање прашања“ и дава преглед на мотивацијата од која произлезе целокупното презентирано истражување. Во него се наведени истражувачките прашања и методологијата која се користи за да се одговорат истите.

Во второто поглавје е даден детален преглед на системите за одговарање прашања развивани низ годините. На почетокот се опишани клучните компоненти на еден општ *QA*-систем и наведени се нивните карактеристики. Потоа, дадена е

класификацијата на *QA*-системите врз основа на повеќе критериуми. Централниот дел на ова поглавје се однесува на значајните пристапи кои се применуваат за анализа на природните јазици. Поглавјето завршува со презентација на мерките за проценка на системите за одговарање прашања.

Третото поглавје ја опишува потребата од вклучување на *NLP*-техниките во насока на успешното одговарање прашања. Истото содржи детален преглед на референтната литература и врвните решенија кои се применети за одговарање прашања на англиски јазик, како и на други јазици. Акцентот е ставен врз морфолошките карактеристики на јазиците и нивното влијание во процесот на прибирање информации и селекција на точниот одговор на поставеното прашање. При тоа, детално е опишано кои морфолошки белези се разгледуваат во ова истражување, кога станува збор за македонскиот јазик. Поглавјето содржи презентација на нова метрика за групирање на зборовите кои припаѓаат во иста лексема, која е искористена и за проширување на прашалникот со **сличен** збор од речникот, во случај да содржи **непознат** збор. На крајот, потенцирана е потребата *IR* (*QA*) системот да ја земе во предвид и близината на термините од прашалникот во документите, со цел да се дојде до подобри резултати.

Четвртото поглавје започнува со презентација на стандардите кои треба да ги задоволи една колекција од прашања со повеќекратен избор, со цел нејзината примена да даде веродостојни резултати. Потоа, во него е даден опис на тест-колекцијата за одговарање прашања на македонски јазик од областа филозофија, искористена за тестирање и анализа во ова истражување. Во продолжение се презентирани севкупните резултати добиени од дизајнираниот *QA*-систем, од резултатите добиени со вклучување на морфолошките белези на македонскиот јазик во прибирањето, па сè до резултатите добиени со инкорпорирање на близината на термините.

Петтото поглавје претставува компаративна анализа на резултатите добиени од примената на дизајнираниот систем врз две дополнителни тест-колекции со прашања со повеќекратен избор. Колекциите се од областа информатички технологии, напишани на два различни јазика: македонскиот и англискиот. Поглавјето дава забелешки кои треба да се земат во предвид за подобрување на точноста на системот врз наведените, но и идни тест-колекции.

На крај, во шестото поглавје се дадени заклучоците од направените анализи во областа на истражување, т.е. во областа прибирање информации и одговарање прашања поставени на македонски јазик, а воедно се дадени и насоките за понатамошна работа.

Додатокот А содржи таксономии кои се дефинирани во различни истражувања, за различни тест-колекции.

Додатокот Б содржи примери на прашања од македонската тест-колекција „Филозофија“, кои не ги задоволуваат стандардните протоколи за превалидација.

Во додатокот В се дадени примери на прашања од македонската тест-колекција „Филозофија“ од секоја категорија, во согласност со новодефинираната таксономија.

Во додатокот Г се дадени примери на соодветни и несоодветни кластери, генерирани со вклучување на метриката за сличност на стрингови, базирана на *Dice*-коефициентот (со праг 0.5). Целта е примена на оваа метрика за автоматско утврдување на зборoформите од речникот зборови генериран од учебникот „Филозофија“ (именуван Речник_1).

Додатокот Д содржи примери на прашања од македонската тест-колекција „Филозофија“, за кои системот не прибира точен пасус, заради незастапеност на одредени клучни зборови од прашањата во Речникот_1.

Додатокот Ѓ содржи примери на непознати зборови (зборови кои ги нема во Речникот_1, генериран од тест-колекцијата „Филозофија“) за кои точно/погрешно е утврден сличен збор, во насока на проширување на прашалникот за пребарување. Прикажаните резултати се однесуваат на примена на метриките за сличност на низи последователни знаци (стрингови) базирани на коефициентите *Dice*, *Positional Dice* и *Jaccard* (за вредност на прагот 0.35), како и метриката базирана на триаголниот прозорец (за вредноста на прагот 1.5).

Додатокот Е ги прикажува резултатите добиени во процесот на прибирање пасуси, со примена на точните кластери од зборoформи (дефинирани рачно) и кластерите генерирани автоматски со метриката базирана на триаголниот прозорец, за праговите 0.5, 1.5 и 2.5, заедно со проширување на прашалникот со метриките базирани на коефициентите *Dice*, *Positional Dice* и *Jaccard* (за повеќе вредности на прагот).

Во додатокот Ж се дадени постигнувањата на *Hanning*-прозорската функција, применета врз петте најдобро рангирани пасуси од *IR*-фазата. Пасусите се добиени со вклучување на кластерите од зборoформи генерирани автоматски со метриката за сличност базирана на триаголниот прозорец (за праговите 0.5 и 2.5) и проширување на прашалникот со истата метрика (за прагот 1.5). Притоа, при имплементација на *Hanning*-прозорската функција не се искористени зборoформите на зборовите од четирите прашалници и не е направено проширување на прашалникот во случај на непознат збор.

Примери на прашања од македонската тест-колекција од областа „Информатички технологии“ (именувана МакИнфо) од секоја категорија, во согласност со новодефинираната таксономија, се дадени во додатокот З.

Примери на прашања од англиската тест-колекција од областа „Информатички технологии“ (именувана АнгИнфо) од секоја категорија, во согласност со новодефинираната таксономија, се дадени во додатокот S.

Додатокот И ја прикажува распределбата на англиските термини од колекцијата МакИнфо по зборовни групи.

Додатокот J ја дава распределбата по зборовни групи на зборовите од речникот генериран од колекцијата АнгИнфо, кои имаат повеќе од една ознака.

Додатокот К ја демонстрира потребата од модификација на **густината на распределба** (*density distribution - DD*) базирана на *Hanning*-прозорската функција, за тест-колекците МакИнфо и АнгИнфо.

Во додатокот Л се наведени неколку причини за намалување на точноста на дизајнираниот *QA*-системот, при одговарање на прашањата од тест-колекцијата МакИнфо.

Додатокот Љ содржи сегмент од примената на дистрибутивен метод за наоѓање зборови кои се појавуваат заедно во текстуални сегменти (*co-occurring words*), врз проширената тест-колекција МакИнфо.

Во додатокот М се прикажани делови од кодот со кои се реализирани најкарактеристичните функции на системот за одговарање прашања на македонски јазик.

2. 50 години истражувања во областа „Одговарање прашања“

„Способноста да се изрази сомнеж е движечка сила на севкупниот напредок на човештвото.“

Indira Gandhi

Прецизното дефинирање на поимот **„Одговарање прашања“** (*Question Answering - QA*) не е толку едноставно како што може да се очекува. Од многу општа перспектива, *QA* може да се дефинира како автоматски процес способен да разбере прашање формулирано на природен јазик и да одговори точно со потребните информации [4].

Сепак, оваа навидум едноставна дефиниција станува мошне комплексна, доколку во детали се анализира кои карактеристики и функционалности треба да ги поседува еден „идеален“ *QA*-систем. Имено, системот треба да биде способен да ја определи потребата од информација изразена во прашањето, да ја лоцира таа информација, да ја извлече, а потоа да генерира одговор и да го претстави во согласност со барањата изразени во прашањето [4]. Освен тоа, системот треба да овозможи и соодветна интеракција со корисниците, преку правилно толкување на нивните прашања, чии одговори ги пребарува во неструктурирани документи напишани на природен јазик.

Во насока на креирање на идеален *QA*-систем, а посебно заради самата комплексност на овој процес, научната заедница се концентрира на разрешување на поедноставни и полесно решливи проблеми, како што се:

- селектирањето на документи за одредена потреба од информација, познато како прибирање информации (*Information Retrieval – IR*),
- прибирањето на специфична информација од структурирани податоци за кориснички прашалници (*Natural Language Interfaces to Databases – NLIDB*), и
- извлекувањето на специфична информација од неструктурирани документи (*Information Extraction - IE*).

Од успешното разрешување на овие проблеми зависи и успешноста на идеалниот *QA*-систем.

Интересот за *QA*-областа произлезе од претпоставката дека корисниците би преферирале прецизни одговори на нивните прашања, отколку да ги испитаат сите достапни документи соодветни на нивната потреба од информација. Како одговор на овој растечки интерес, два документа прават обид да ги организираат *QA*-истражувањата, а тоа се: „Визионерски исказ за водење на истражувањата во *QA* и областа автоматско извлекување на резиме од текстуален документ“ [5] и „Предизвици, задачи и програмски структури за водење на истражувањата во *QA*“ [6].

Првиот документ ги дефинира дострелот и идните можности на *QA*-системите, со цел да се задоволат очекувањата и барањата на широк спектар потенцијални

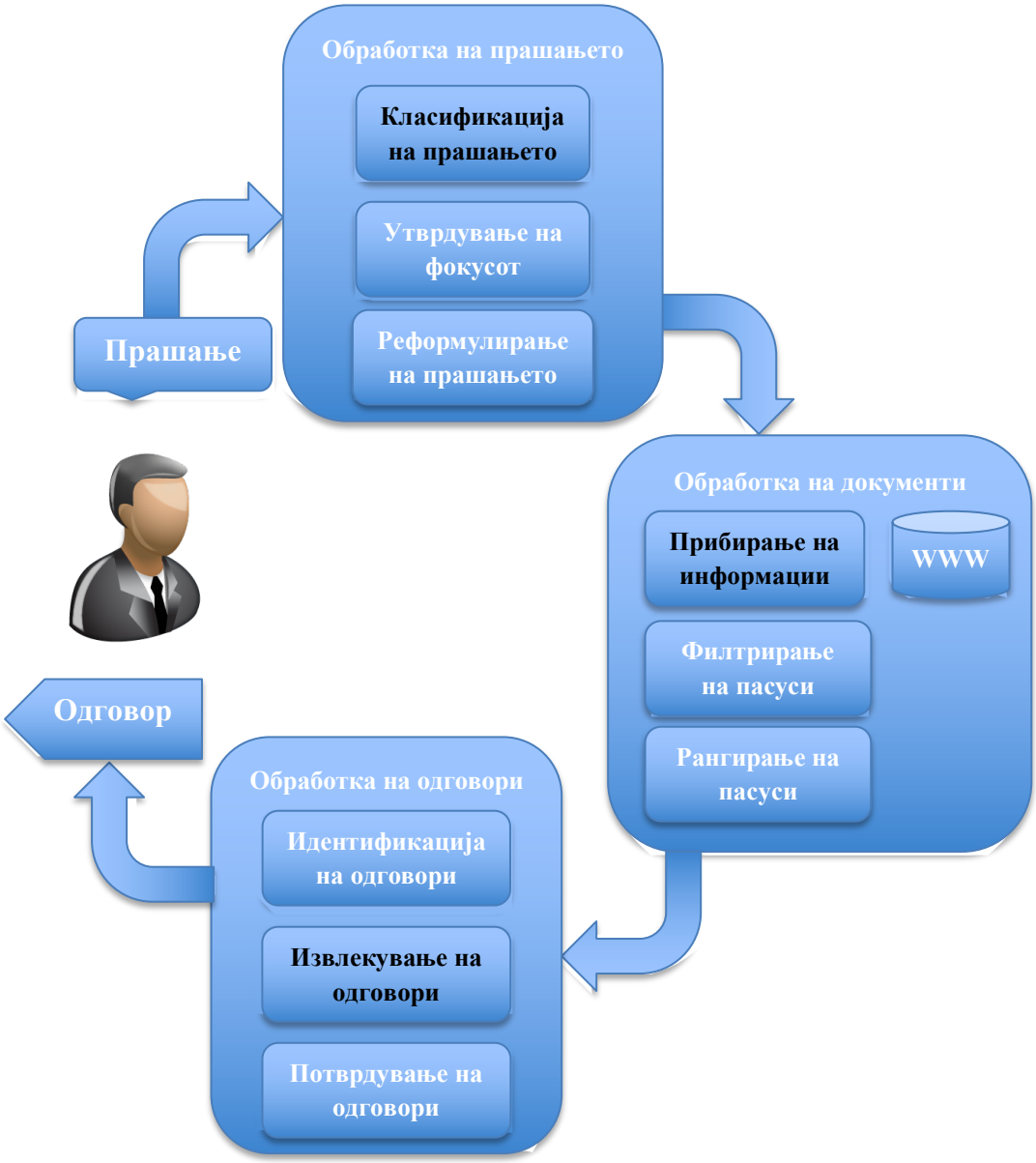
корисници, од обични корисници до професионални аналитичари. Освен одговарање прашања кои побаруваат конкретна информација изразена во единствена фраза во рамките на одреден документ, *QA*-системот треба да биде способен да прифати прашања чии одговори се засноваат на носењето одлуки. Со цел да се генерираат такви одговори, системот треба да врши синтеза на информациите добиени од различни извори и да ги претстави на корисникот во соодветна форма. Уште повеќе, системот треба да користи моќни мултимедиски алатки за навигација кои овозможуваат проверка на сите можни извори на информации и кои ја олеснуваат интеракцијата со корисникот.

Вториот документ презентира строго дефинирана истражувачка рамка чија цел е во претстојните пет години да овозможи реализација на визијата за истражување предвидена во првиот документ. Тој детално ги опишува предизвиците во истражувањето со кои *QA*-заедницата треба да се соочи во блиска иднина, како: интерактивното одговарање прашања, интегрирањето на информациите од различни извори, обработката на повеќејазичните информации, генерирањето на одговори и нивната конечна презентација. Оваа агенда дефинира ефективно разрешување на сите предизвици во *QA*-областа за само пет години. Сепак, се покажува како исклучително амбициозна бидејќи голем дел од тие предизвици остануваат нерешени до денес.

За да се долови наведената комплексност на поимот автоматско одговарање на прашања, ова поглавје дава детална анализа на клучните компоненти опфатени во современите *QA*-системи, ги претставува критериумите за нивна класификација, основните принципи и техники кои се интегрирани во нив и дава анализа на метриките кои се користат за нивна проценка.

2.1. Компоненти на системите за одговарање прашања

Од доцните 1960-ти години па сè до денес, развиени се бројни системи за одговарање прашања (*QASs*) [7]. Тие опфаќаат широк спектар на различни техники и архитектури, заради што практично е невозможно да се опфатат сите варијанти во единствена архитектура. Сепак, постојат неколку карактеристики кои се заеднички за *QA*-системите, што овозможуваат да се даде една општа архитектура на прототип систем за одговарање прашања. Општо земено, *QA*-системот се состои од три главни модули кои играат исклучително важна улога во процесот на одговарање прашања, а тоа се: **обработка на прашањата**, **обработка на документите** и **обработка на одговорите** [8]. На сликата 1 е прикажана општата архитектура на *QA*-системот и начинот на кој модулите заемно дејствуваат. Секој модул има централна компонента, покрај другите дополнителни компоненти.



Слика 1. Архитектура на општ систем за одговарање прашања

2.1.1. Модул за обработка на прашањето

Задачата на модулот за обработка на даденото прашање поставено на природен јазик е негова анализа и креирање на репрезентација за пребарување. За таа цел, овој модул е неопходно да направи:

- **класификација на прашањето** (уште познато и како предвидување на **очекуваниот тип на одговор**), обично врз основа на таксономија од веќе познати прашања за системот,
- **утврдување на фокусот на прашањето**, кој ја претставува главната информација која се побарува за да се одговори прашањето на корисникот, и
- **реформулирање на прашањето**, со цел да се трансформира во прашалници за пребарување.

2.1.1.1. Класификација на прашањето

Задача на класификаторот на прашањата е да додели една или повеќе ознаки на секое прашање поставено на природен јазик, во зависност од стратегијата за класификација. На пример, за прашањето „Кој е основач на скептицизмот?“, класификаторот треба да додели ознака „личност“, бидејќи одговорот на прашањето е од типот „личност“. Имајќи го предвид фактот дека се предвидува типот на одговор, класификацијата на прашањето уште се нарекува и **предвидување на типот на одговор** (*answer type prediction*). Множеството од претходно дефинираните категории, кои се сметаат за класи на прашањето, се нарекува **таксономија на прашањата** или **таксономија на типот на одговорите**.

Зошто е неопходна класификацијата на прашањата? Точната класификација на прашањето се покажува како најзначаен фактор за успешно идентификување на точниот одговор во огромни колекции документи. Исто така, успешната класификација овозможува системот да искористи различни стратегии за обработка на прашањето, што од друга страна придонесува за зголемување на севкупната ефективност на системот [9]. Според статистиките од *Moldovan et al.* [10], 36.4% од неуспешноста на QA-системот се должи на погрешната анализа на прашањето, додека 28.2% се резултат на погрешната класификација на прашањето.

***Пристапи за класификација на прашањата*.** Постојат два клучни пристапи за класификација на прашањата: **пристап базиран на правила** и **пристап базиран на учење**. Компромисот меѓу заемно спротивставените пристапи во кои првиот зависи од строго дефинирани правила, а вториот се обидува да предвиди врски меѓу елементите, е воспоставувањето на **хибридниот пристап** кој ги комбинира двата претходно наведени [11].

- **Пристапот базиран на правила** (*rule-based approach*) се обидува прашањето да го класифицира користејќи рачно подготвени правила [12]. Главниот недостаток на овој пристап е потребата од дефинирање на големо множество правила, со што би се овозможило утврдување на типот на одговор за различни прашања. Исто така, истражувањата покажуваат дека креираните правила може да дадат одлични резултати на одредно множество податоци, но да бидат слабо применливи на ново множество податоци [13]. Една од причините за ова е можноста две прашања да припаѓаат во иста категорија, но да имаат различни синтаксички форми заради што е неопходно да се креираат различни правила за нивна точна класификација. Затоа е исклучително тешко рачно да се креира успешен класификатор со ограничен број правила.
- **Пристапот базиран на учење** (*learning-based approach*) ја извршува класификацијата преку извлекување на одредени карактеристики од прашањата. Користејќи ги тие карактеристики, класификаторот се обучува на множество прашања кои се рачно означени, а потоа се користи за предвидување на категоријата (класата) на нови прашања.
- **Хибридните пристапи** (*hybrid approaches*) се покажуваат исклучително успешни при класификација на прашањата. Така, *Silva et al.* [11] прво го

класифицираат прашањето користејќи одредени рачно дефинирани правила, а потоа информациите добиени од тие правила ги користат како карактеристики во пристапот базиран на учење.

Таксономија на прашањата. Многу истражувачи имаат предложено различни таксономии за класификација на прашањата. Врз основа на деталната анализа направена на 40 таксономии, *Hao et al.* [14] ги забележуваат следниве четири различни вида таксономии, кои даваат јасна слика за нивните различни карактеристики:

- **Таксономија базирана на прашалниот збор од прашањето** (*taxonomy based on interrogative type*). Така, *Mudgal et al.* [15] предлагаат „6W+1H“[10] таксономија која се состои од 7 груби категории („кој“, „каде“, „што“, „кога“, „чиј“, „зошто“ и „како“), како и 27 фини категории, вклучувајќи ги: „личност“, „организација“, „време“, „локација“, итн. (табела А1). *Moldovan et al.* [16] дефинираат множество од девет категории, секоја со одреден број на поткатегории и соодветните типови на одговори (табела А2).
- **Таксономија базирана на стилот (начинот) на искажување на прашањето** (*taxonomy based on question description style*). Ваквите таксономии не се повикуваат на прашалниот збор („кој“, „како“, „зошто“, итн) и обично се состојат од едно ниво категории (груби категории). Таков е примерот со таксономијата предложена од *Graesser et al.* [17] која содржи 18 категории, како: „верификација“, „квантификација“, „споредба“, итн. (табела А3).
- **Таксономија базирана на семантичката интерпретација на очекуваниот тип на одговор** (*taxonomy based on the semantic interpretation of question target*). Оваа таксономија може да содржи повеќе нивоа. На пример, *Xin et al.* [18] дизајнираат таксономија во две нивоа, поточно таксономија со 6 груби и 50 фини категории (табела А4).
- **Таксономија за ограничен домен** (*taxonomy for restricted domain*). Тука се вбројува таксономијата дадена од *Benamara* [19] за туристички домен која вклучува 9 категории (табела А5). За медицински домен, *Roberts et al.* [20] предлагаат нова таксономија со 14 главни категории, чија цел е класификација на прашања за заболувања (како „анатомија“, „дијагноза“, „компликација“, итн.) (табела А6).

Испитувајќи ја корелацијата меѓу 28 таксономии за *QA*-системи на општ домен, *Hao et al*. [14] откриваат дека меѓу нив постои големо споделување на категориите. Така, петте најзастапени категории во таксономиите се: „локација“ (77%), „број“ (63%), „време“ (63%), „личност“ (60%) и „организација“ (57%). Она што исто така треба да се забележи е специфичноста на таксономиите во зависност од природениот јазик за кој се креирани. И покрај популарноста на некои таксономии направени за англискиот јазик, задолжителна е нивна ревизија пред истите да бидат применети врз други јазици. За таа цел, ова истражување дефинира нова таксономија согласна на спецификите на македонскиот јазик и тест-колекцијата на македонски јазик, опишана во секцијата 4.2.

[10] who, where, what, when, which, why, how

2.1.1.2. Утврдување на фокусот на прашањето

Класификацијата на прашањето понекогаш се покажува како недоволна за наоѓање на неговиот одговор. На пример, прашањата кои започнуваат со прашалните зборови „кој“ („која“, „кое“, „кои“) може да бидат прилично нејасни во поглед на информацијата која ја побаруваат. За разрешување на оваа двосмисленост, неопходно е и утврдување на концептот наречен **фокус** на прашањето (*question focus*). *Moldovan et al.* [16] го дефинираат **фокусот** на прашањето како збор или низа од зборови кои навестуваат каква информација се бара во самото прашање. На пример, во прашањето „Кој просветител бил наречен Лав од Фарнеј?“, фокусот е „просветител“. Доколку е познат фокусот, тогаш системот е во можност полесно да го идентификува очекуваниот тип на одговор, кој во овој пример е „личност“.

Идентификувањето на фокусот се прави преку рачно дефинирани правила кои ја користат класификацијата на прашањето, или преку примена на статистички пристап. Статистичките пристапи побаруваат множество за обука кое се состои од прашања со познат фокус, што може да биде исклучително скапо во однос на потрошеното време и вложениот труд.

2.1.1.3. Креирање прашалници за пребарување

Штом ќе го определи фокусот и типот на очекуваниот одговор, системот преминува кон креирање на листа од клучни зборови со цел да генерира прашалник(ци) за прибирање информации. Процесот на извлекување на клучните зборови може да се изврши преку стандардните техники, како: препознавање на именуваните ентитети (*named entity recognition*), дефинирање на листа од стоп зборови, користење на означувачи на зборовните групи (*part-of-speech taggers*), но и користење на множество од хеуристики [16]. Често се применува и метод за проширување на множеството од клучни зборови од прашањето, преку вклучување на онлајн лексички извори, како што е на пример *WordNet*[11] онтологијата. Множествата синоними од *WordNet* може да се искористат за проширување на множеството од клучни зборови со семантички поврзани зборови, кои може да се појават во документите што го содржат точниот одговор на прашањето.

Дури и во ситуации на достапност на он-лајн лексикон за одреден природен јазик, методите кои го користат лексиконот за утврдување на релациите меѓу зборовите имаат бројни ограничувања. Ограничувањето најчесто е резултат на незастапеноста на одредени зборови во самиот лексикон, особено зборови кои се однесуваат на специфични домени. Токму затоа, новите истражувања сè почесто ги применуваат **дистрибутивните методи** (*distributional methods*), кои претставуваат статистички методи за утврдување на степенот на поврзаност на два збора [8]. Она што тие го претпоставуваат е дека значењето на зборот е тесно поврзано со дистрибуцијата на зборовите кои се појавуваат околу него. Разгледувањето на контекстот овозможува проширување на прашалникот со зборови слични, поврзани или синоними на даден клучен збор. *Curran* [21] испитува повеќе дистрибутивни алгоритми и забележува дека

[11] http://wordnet.princeton.edu/

за утврдување на сличноста на два збора, најдобри резултати се добиваат доколку се примени $t-test$ за задавање тежини на компонентите од векторот кој го претставува зборот и примена на *Dice*-методот или *Jaccard*-методот за мерење на сличноста на векторите.

Во случај да се извлечат (генерираат) многу клучни зборови од прашањето, истите може да бидат сортирани во согласност со нивната важност, односно може да се изберат најважните N зборови со кои ќе се креира прашалникот за пребарување. N е вредност која се нагодува во зависност од добиените резултати вклучувајќи различен број клучни зборови во прибирањето информации. Нивниот број може да биде контролиран и од компонентата за прибирање пасуси (секција 2.1.2), која заради (не)квалитетот на добиените пасуси може да утврди дека е потребно да се искористи друго множество од клучни зборови.

2.1.2. Модул за прибирање информации

Системите за одговарање прашања инкорпорираат еден или неколку системи за прибирање информации (*IR systems*) со цел детерминирање на сегменти од огромните колекции документи, кои во наредните фази се предмет на анализа. Скоро секогаш *QA*-системите на општ домен го користат Интернетот како еден од изворите за прибирање информации. Резултатите кои се добиваат од *IR*-системите (најчесто документи) се филтрираат за да се отстранат оние делови (пасуси) кои не содржат доволен број клучни зборови од прашањето кое треба да се одговори. Со ова се овозможува генерирање на пасуси кои овој модул ги рангира врз основа на тоа колкава е нивната веродостојност. Доколку постојат премногу или премалку пасуси, модулот побарува генерирање на нови прашалници (со повеќе или помалку клучни зборови) и прибирање на нови документи, со што ќе се обезбеди разумен број пасуси кои се предаваат на модулот за обработка на одговорите. Мотивацијата за извлекување на најважните аспекти од документите (идентификувани како пасуси) е неопходна со цел мал дел од содржината да се проследи на детална анализа, со што ќе се забрза самиот систем. Јасно е дека времето на одговор на *QA*-системите е исклучително важно заради интерактивната природа на одговарањето прашања.

2.1.2.1. Прибирање на информациите

За прибирање на информациите, многу често *QA*-системите користат некој од постоечките веб-пребарувачи (најчесто Гугл). Причината за тоа е што стандардните *IR*-пристапи (како моделот на векторски простор со косинус случноста) за мерење на сличноста меѓу документите и прашалникот не се покажуваат како најсоодветни за *IR* при одговарањето прашања. Имено, *QA*-системот обично побарува документи во кои се содржат сите клучни зборови внимателно определени во првиот модул (модулот за обработка на прашањето), додека стандардните *IR*-модели прибираат и документи во кои не се содржат сите клучни зборови. Исто така, не може да се игнорира ниту бројот на документи кои веб-пребарувачите ги имаат индексирано (на пример, Гугл во моментов има индексирано 130 трилиони[12] индивидуални страници), како и можноста

[12] https://www.google.com.au/insidesearch/howsearchworks/thestory/

да се применат Булови оператори при пребарувањето.

Системите за прибирање информации обично се проценуваат со примена на две метрики: **прецизност** (*precision*) и **отповикување** (*recall*). **Прецизноста** покажува колкав дел од прибраните документи се релевантни на прашалникот, додека **отповикувањето** покажува колкав дел од релевантните документи се прибрани од системот. Генерално, *IR*-системите имаат цел да ги оптимизираат овие две метрики. Што се однесува до системите за одговарање прашања, целта е значително различна. Бидејќи тие вршат дополнителна обработка на прибраните документи, отповикувањето на *IR*-системот имплементиран во нив е многу позначајно отколку прецизноста. *QA*-системот ги отфрла ирелевантните документи или нивни сегменти, како дел од компонентата **филтрирање на пасуси** (секција 2.1.2.2). На овој начин се зголемува прецизноста на информациите, со што се компензира пониската прецизност добиена од *IR*-системот. Од друга страна, пониското отповикување на *IR*-системот значи дека помалку е веројатно одговорот да се наоѓа во прибраните документи. Доколку одговорот не се наоѓа во нив, *QA*-системот мора да го увиди тоа (што е исклучително тешко), а потоа да направи повторна селекција на клучни зборови со кои би се повторил процесот на прибирање информации. Ако и во случај *IR*-системот не успее да прибере документ во кој се содржи точниот одговор, тогаш и покрај редефинирање на прашалникот, ни најдобриот *QA*-систем не може да го одговори поставеното прашање.

2.1.2.2. Филтрирање на пасусите

Филтрирањето на пасусите (*passage filtering*) се користи со цел да се редуцира бројот на документи кандидати (кој обично е многу голем), како и количеството на текст во секој документ. Концептот на филтрирање пасуси се заснова на принципот дека најрелевантните документи би требало да ги содржат клучните зборови од прашањето во неколку соседни пасуси, односно дека тие не треба да се распространети низ целиот документ. За таа цел се анализира локацијата на множеството од клучни зборови во секој документ. Доколку тие се наоѓаат во некое множество од N последователни пасуси, тоа множество од пасуси се прибира. Во спротивно, документот не се проследува на понатамошна анализа. Параметарот N експериментално се утврдува врз основа на проценката на перформансите на системот, испитувајќи различни растојанија меѓу клучните зборови во документите.

Друг пристап е користењето на очекуваниот тип на одговор кој е утврден во претходниот модул, при обработка на самото прашање. Доколку се заклучи дека одредени пасуси во документот не содржат ентитет од очекуваниот тип, тогаш тие пасуси се изоставуваат од понатамошната анализа.

2.1.2.3. Рангирање на пасусите

Останатите пасуси се рангираат користејќи различни хеуристики. Една од нив е задавањето тежина на пасусот, во согласност со степенот на доверба на неговиот извор (на пример, поголема тежина може да се зададе на познати податочни извори). Сепак, поновите истражувања најчесто користат надгледувано машинско учење со мало множество од карактеристики, кое лесно може да се извлече од големиот број пасуси.

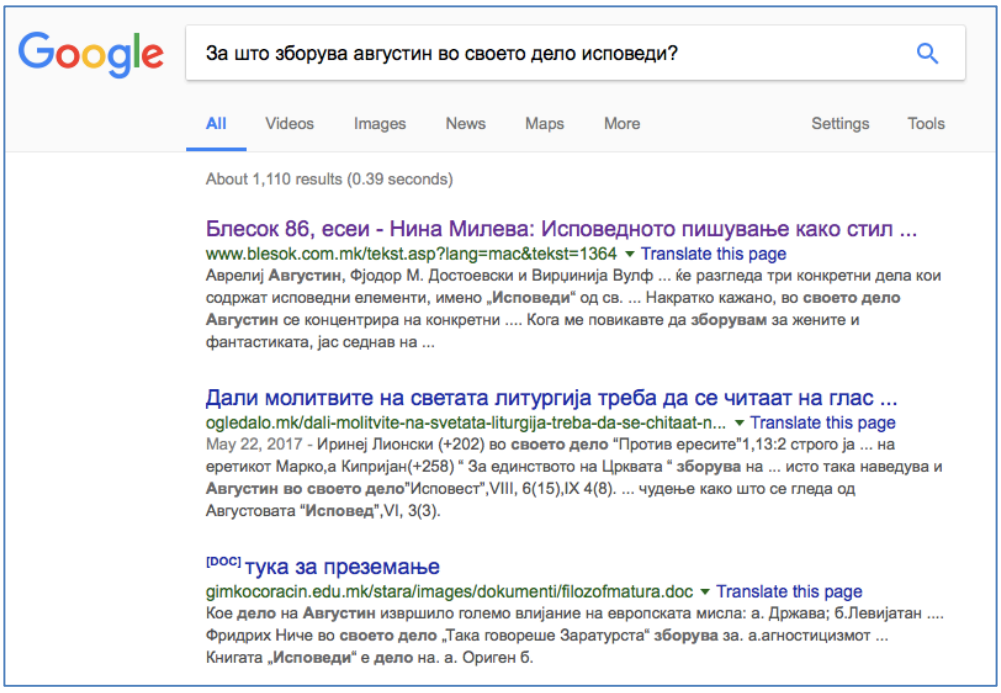
Меѓу нив, најзначајни се:

- бројот на **именувани ентитети** од очекуваниот тип во пасусот,
- бројот на **клучни зборови** во пасусот,
- **најдолгата точна низа од клучни зборови** од прашањето која се појавува во пасусот,
- **рангот на документот** од кој е извлечен пасусот,
- **близината на клучните зборови** од прашањето во пасусот, при што најчесто се користи најмалата покривка која опфаќа што повеќе клучни зборови во тој пасус [22], [23], и
- **бројот на *N* - грами** од прашањето кои се појавуваат и во пасусот, при што се преферираат пасуси со поголем број на заеднички *N* – грами со прашањето [24].

Дефинирањето на поимот **пасус** во рамки на документите зависи од истражувачите кои го дизајнираат системот. Генерално, различните пасуси кои се користат се групирани во три класи: **дискурс пасуси** (*discourse passages*), **семантички пасуси** и **пасуси базирани на прозорец**. **Дискурс пасусите** се дефинираат преку поврзана серија од искази (на пример, реченици, параграфи и секции), кај **семантичките пасуси** границите ги дефинира промената на темата, додека **пасусите базирани на прозорец** содржат одреден број зборови [25].

Она што е уште позначајно, независно од тоа каков тип на пасус ќе се примени во системот, е одлуката како најдобро да се имплементира прибирањето пасуси кај *QA*-системите. Односно, дали документите да се поделат на пасуси пред нивното индексирање (т.е. во фазата на претпроцесирањето), со што пасусот се дефинира како единица за прибирање, или тоа да се направи по рангирањето на документите во реално време. Двата начина може да доведат до значително различно рангирање на истите пасуси, што пак од друга страна може силно да влијае врз успешноста на одговарањето прашања. Во своето истражување *Roberts et al.* [26] испитуваат неколку пристапи за прибирање пасуси. Тие заклучуваат дека најдобрите резултати се добиваат доколку прво се приберат целосни документи, потоа најдобро рангираните документи се разделат на пасуси и на крај се изврши повторно прибирање, но овојпат на претходно дефинираните пасуси.

Од друга страна, *QA*-системите кои имплементираат веб-пребарувач, наместо извлекување на пасуси од прибраните документи, ги искористуваат фрагментите (*snippets*) кои веб-пребарувачите ги прикажуваат за одредена потреба од информација. На сликата 2 се прикажани фрагменти од првите три документи прибрани од пребарувачот Гугл за прашањето „За што зборува Августин во своето дело Исповеди?“, прашање кое се содржи во тест-колекцијата на македонски јазик од областа филозофија (секција 4.2).



Слика 2. Три фрагменти од прабарувачот Гугл како одговор на прашањето „За што зборува Августин во своето дело Исповеди?“

2.1.3. Модул за обработка на одговорите

Последен модул во архитектурата на еден општ систем за одговарање прашања е модулот за обработка на одговорите. Тој е задолжен за идентификување и извлекување на одговори од множството подредени пасуси, кои се проследени од претходниот модул.

2.1.3.1. Идентификување на одговорите

За водење на овој процес клучен е очекуваниот тип на одговор (најверојатно) утврден при обработката на прашањето. Со цел да се идентификуваат делови од пасусите кои содржат зборови од очекуваниот тип, често се користат техники за површно разложување (*shallow parsing*), со кои може да се препознаат именуваните ентитети (на пример, имиња на личности и организации, датуми, итн.). Исто така, користењето на означувач на зборовните групи (*part of speech tagger*) може да помогне да се препознаат кандидатите за одговори во рамки на проследените пасуси.

Сепак, одговорите на одредени прашања, како прашањата кои побаруваат дефинирање на поими, немаат конкретен именуван ентитет. За овие прашања се користат шаблони од изрази напишани рачно, со цел да се идентификува одговорот. Креираните шаблони се исто така применливи и во ситуации кога пасусот содржи повеќе примери од истиот претходно утврден именуван ентитет. Шаблоните се специфични за секоја категорија прашања и освен нивно рачно креирање, може да бидат научени и автоматски, користејќи методи за извлекување на релации (*relation extraction methods*).

2.1.3.2. Извлекување на одговорите

Штом се идентификувани кандидатите за одговори, се применуваат одредени методи со цел да се изврши нивното рангирање. Таквите методи се бројни и обично вклучуваат комбинација од неколку пристапи, како:

- **Определување на сличноста помеѓу прашањето и кандидатот за одговор**. Системот повисоко ги рангира кандидатите кои се во контекст сличен како и прашањето. Сличноста меѓу две реченици може да се определи доколку се искористи бројот на заеднички зборови или да се искористат покомплексни мерки за сличност базирани на синтаксички и семантички информации. Со цел да се земе предвид и можноста за варијации во изразувањето, одредени системи користат и дополнителни лингвистички извори, како *WordNet* [27], кои помагаат да се определат релациите меѓу зборовите (на пример, дали два збора се синоними).
- **Определување на популарноста на кандидатот за одговор**. Во балансиран корпус[13], разумно е да се претпостави дека бројот на појавувања на одреден стринг како можен одговор на дадено прашање, е директно поврзан со веројатноста тој стринг да биде одговорот кој се бара. Затоа, наједноставен метод за рангирање на кандидатите за одговор е да се искористи бројот на појавувања на кандидатите во корпусот. Ова ја повлекува и неопходноста од утврдување дали два стринга во суштина се однесуваат на истиот ентитет, односно имплементација на задачата **соединување на одговори** (*answer merging*). Ваквата задача за детектирање на варијации на одговорот е тесно поврзана со задачата за **разрешување на кореференците** (*coreference resolution*) од областа на извлекувањето информации.
- Денес, новите истражувања ја користат и достапноста на огромното количество електронски податоци, особено преку вебот [28]. Таквото количество податоци создава ситуација на **редундантност на податоци** (*data redundancy*), при што одговорот на одредено прашање може да се појави многу пати во различен контекст. Методите за редундантност на податоци се базираат на претпоставката дека како се зголемува бројот на варијации за потврда на одговорот, така се зголемува и веројатноста за наоѓање на нивната оправданост. Дури и во ситуации каде се употребува фиксно множество од документи (како *TREC*-конференциите до 2007 година), Интернетот се користи за да се реализира паралелен процес на пребарување на одговорот на дадено прашање. Овој пристап овозможува прибирање на дополнителни информации со цел да се прошири оригиналното прашање [29] или добивање на редундантни податоци за кандидатите за одговор, со што се овозможува валидација на одговорите кои системот ги извлекол од документ-колекцијата [30]. Оваа тенденција сè повеќе се зголемува бидејќи направените експерименти потврдуваат дека користењето

[13] **Балансираниот корпус** обично опфаќа широк опсег текстуални категории кои би требало да бидат карактеристични за јазикот кој се разгледува, како и за неговата разноликост [31].

на вебот на овој начин доведува до извонредно подобрување на перформансите на системот.

- **Утврдување дали кандидатот за одговор се совпаѓа со шаблон соодветен на прашањето**. Креирањето на множество шаблони (рачно или со користење на техники за машинско учење) е еден од начините за одговарање специфични прашања. Шаблоните се рачно анотирани со вредност на доверливост. Системот може да одбира меѓу повеќе кандидати за одговор во зависност од нивото на доверливост на шаблонот, соодветен на прашањето со кој се совпаѓа.

Поновите системи за одговарање прашања, во оваа фаза сè повеќе ги користат методите базирани на учење. Односно, ги рангираат потенцијалните одговори со примена на класификатори со одредени карактеристики, меѓу кои [8]:

- **Совпаѓањето со очекуваниот тип одговор**: Оваа карактеристика има потврдна вредност доколку се утврди дека кандидатот за одговор содржи фраза со точниот тип на одговор.
- **Совпаѓањето со шаблон**: Карактеристика која го содржи идентитетот на шаблонот кој соодветствува на кандидатот за одговор.
- **Бројот на клучните зборови кои се совпаѓаат**: Го содржи бројот на клучните зборови од прашањето кои се појавуваат во кандидатот за одговор.
- **Растојанието меѓу зборовите**: Се дефинира како растојание меѓу кандидатот за одговор и клучните зборови од прашањето (измерено во број на зборови).
- **Факторот на новост**: Карактеристика со потврдна вредност доколку барем еден збор од кандидатот за одговор е нов, односно не се појавува во прашалникот.
- **Поставеноста на интерпункцијата**: Карактеристика со потврдна вредност доколку кандидатот за одговор го следи интерпункциски знак.
- **Низите од зборови од прашањето**: Карактеристика еднаква на должината на најдолгата низа од зборови од прашањето која се јавува во кандидатот за одговор, итн.

Во оваа фаза системот може да побара примена на нов план за генерирање на нови пасуси.

2.1.3.3. Валидација на одговорите

Валидацијата на одговорот е пристап кој се користи со цел да се направи подлабока проверка на кандидатите за одговори, пред истите да му бидат презентирани на корисникот. Причината за ова е можноста кандидатот за одговор да е од очекуваниот тип, но сепак да е комплетно погрешен. На пример, негативните броеви не смее да се дозволи да бидат одговор на прашања за растојание или години. За разрешување на овие ситуации се применуваат повеќе пристапи, како: вклучување на специфични извори со знаење, редундантност на податоците, методи за валидација базирани на логика, итн. [32].

2.2. Класификација на системите за одговарање прашања

Големиот број системи за одговарање прашања кои денес се достапни, опфаќаат различни домени, извори на податоци, типови прашања кои ги одговараат, формати на одговорите, итн. Врз основа на тоа, во литературата се забележуваат повеќе клучни критериуми кои се однесуваат на нивната класификацијата, и тоа [33]:

- доменот врз основа на кој се развива *QA*-системот;
- типовите кориснички прашања кои системот ги одговара;
- типот на анализа која се извршува врз корисничките прашања, како и врз документите кои се пребаруваат;
- типот на податочните извори, како: структурирани (базите на податоци), неструктурирани (извештаи, книги, статии) и полуструктурирани (*XML*);
- карактеристиките на податочните извори, како: големината, јазикот, хетерогеноста, дијалектот на податоците (формален или неформален јазик), итн.;
- техниките кои се користат за прибирање одговори (како, техники од податочното рударење (*Data Mining*), прибирањето информации (*IR*), обработката на природните јазици (*NLP*)), и
- формата на одговорите генерирани од *QA*-системот, како: реченица (најчесто за прашања кои побаруваат факт како одговор), пасус (за прашања кои побаруваат опис или дефиниција) или резиме (добивање на скратена верзија од текст).

Во продолжение подетално се разработени клучните критериумите за класификација на *QA*-системите, наведени се карактеристиките на системите од секоја класа, како и предизвиците со кои тие се соочуваат.

2.2.1. Класификација на *QA*-системите врз основа на доменот

Задачата за генерирање одговор на прашање поставено од корисникот е поврзана со специфичноста на информацијата која прашањето ја побарува. Имено, одредени корисници имаат потреба од општи информации за различни теми, додека некои од нив побаруваат конкретна информација од одреден домен. Токму поради тоа, како критериум за класификација на *QA*-системите се наметнува доменот на кој му припаѓаат прашањата и документите.

QA-системи на општ домен. Главна карактеристика на овие системи е тоа што одговараат прашања кои припаѓаат на различни домени. Со цел системот да биде во можност да одговора прашања на различни теми, неопходно е да располага со различни колекции од документи. Во методологијата за генерирање одговори, генерално овие системи користат општа онтологија и општо знаење. Тоа од друга страна резултира со квалитет на одговори кој не е многу висок, заради што и најчесто се користат од обичните корисници [4]. Со оглед на фактот дека домен експертите побаруваат специфични информации, далеку посоодветни за нив се *QA*-системите со ограничен домен.

QA-системи со ограничен домен. Трите клучни карактеристики кои еден ограничен домен го прават добар за *QA* се следниве [34]:

- **Лимитираност**: Треба точно да се разграничи какви прашања смее да постават корисниците. На пример, доменот за вести не е лимитирачки, бидејќи во него се опфатени различни тематики, за разлика од доменот за биологија. Лимитирачките домени треба да содржат и речник за терминологијата, кој овозможува локализација на доменот.
- **Комплексност**: Доколку доменот не е комплексен, би се користеле обични табели за пребарување како замена за *QA*-системот.
- **Практичност***:* Да постои заедница која би го користела *QA*-системот (во спротивно, воопшто нема потреба од негово развивање).

Најпопуларни домени за *QA*-системите се оние кои ги задоволуваат овие карактеристики. Меѓу нив посебно се истакнува медицинскиот домен. Тој е комплексен, бидејќи побарува анализа на мошне специфични прашања. Исто така е практичен, бидејќи големиот број нови публикации од медицински апект ги принудува медицинските лица да бидат постојано во тек со новите откритија. И секако е лимитирачки, со огромен број публикации кој постојано се зголемува (*MEDLINE*[14]) и содржи детална онтологија за болести, лекови и начини на лекување (*MeSH*[15]).

Во литературата денес може да се сретнат различни *QA*-системи со ограничен домен. Тие користат специфична онтологија и терминологија, што овозможува да се постигне висок квалитет на одговорите. Нивниот главен недостаток е ограниченият број прашања кои можат да ги одговорат заради специфичноста на доменот. Заради тоа, различните *QA*-системи со ограничен домен може да се интегрираат, со цел да се изгради *QA*-систем на општ домен [35]. Таквиот систем, пред сè, треба да изврши анализа на клучните зборови од прашањето и да препознае кој систем со ограничен домен треба да го одговари тоа прашање, што е исклучително предизвикувачка задача.

2.2.2. Класификација на *QA*-системите врз основа на типовите прашања

Како што е наведено во секцијата 2.1.1, успешноста на генерирањето одговори на корисничките прашања е директно поврзана со класификацијата на поставеното прашање. Најголем дел од скорешните пристапи базирани на правила и хибридните пристапи за класификација на прашањата ја користат таксономијата предложена од *Xin et al.* [18], која стана неформален стандард. Причина за тоа е скапоценото множество од 6000 означени прашања[16], познато како *UIUC* (*University of Illinois Urbana-Champaign*) множесто од податоци. Ова множество уште е познато и како *TREC*-множество од податоци, заради неговата широка примена на *TREC*-конференциите. Неговата важност уште повеќе се зголемува откако *Metzler et al.* [36] ја збогатуваат *UIUC* таксономијата уште со две класи, **набројување** и **да-не-објаснување**, креирајќи дополнително множество од 250 прашања прибрани од

[14] https://www.nlm.nih.gov/pubs/factsheets/medline.html
[15] https://www.nlm.nih.gov/pubs/factsheets/mesh.html
[16] http://cogcomp.cs.illinois.edu/Data/QA/QC/

MadSci[17] архивата со прашања.

Она што се забележува кај *UIUC* таксономијата, како и другите актуелни парадигми за класификација на прашањата е нивната фокусираност кон специфични типови прашања. Така на пример, во *UIUC* таксономијата сите груби категории (освен категоријата „опис“) се ориентирани кон факти. Од друга страна, истражувањата направени од *Jansen et al*. [37] за класификација на прашањата, потврдуваат дека денес прашањата на корисниците сè повеќе се нерамномерно распределени низ различни области (ова особено се однесува на прашањата поставени на Интернет). За таа цел *Fan et al*. [38] предлагаат таксономија на прашања погодна за општ QA-систем, која ги содржи следниве категории: факт, набројување, причина, решение, дефиниција и навигација. Во оваа секција подетално се опишани трите најчесто користени категории на прашања на кои се заснова и новата таксономија, дефинирана во ова истражување (секција 4.2), а тоа се: фактовидните прашања, прашањата со набројување и описните прашања.

Фактовидни прашања. Овие прашања се едностави и како одговор побаруваат конкретен факт кој треба да се содржи во кратка фраза или реченица [4]. Денешните *QA*-системи достигнуваат задоволителна точност при одговарање на фактовидните прашања [39]. Главната причина е тоа што овие прашања долги години беа во фокусот на проценката на *QA*-системите, како дел од конференцијата за пребарување на текст (*TREC*) [40]. Исто така, очекуваниот тип одговор (*expected answer types*) за најголем дел од фактовидните прашања најчесто е именуван ентитет (*named entity*), кој лесно може да биде откриен во документите користејќи означувачи на именуваните ентитети [7], [35]. Со тоа му се овозможува на *QA*-системот успешно извлекување на потенцијалните одговари без користење на комплексна обработка на природните јазици. Она што дополнително придонесува за високата точност при одговарањето на овие прашања е можноста како извор на информации да се користи *Wikipedia*, како и медиумите за пренесување на вести [41]. Но, автоматската идентификација на фактовидните прашања се наметнува како значаен недостаток на системите за одговарање прашања, која сè уште е предмет на актуелни истражувања.

***Прашања со набројување*.** Во серијата *TREC*-работилници, прашањата со набројување за првпат се воведени во 2001 година и се актуелни сè до 2007 година, кога привремено се затвора *QA*-работилницата. Овие прашања побаруваат листа од ентитети или факти во одговорот. Од таа причина најголем дел од *QA*-системите ги третираат на сличен начин како и фактовидните прашања. Главната разлика е во тоа што за овие прашања, *QA*-системите не се принудени да го дадат најдобриот одговор, како што е случај со фактовидните прашања. Како и кај фактовидните прашања, очекуваниот тип на одговор кај овие прашања е именуван ентитет. Затоа, техниките кои се користат за успешно одговарање на фактовидните прашања, може да се искористат и за одговарање на прашањата со набројување, без потреба од подлабока обработка на природниот јазик, а притоа да се постигне висока точност.

[17] http://www.madsci.org/

Главните проблеми кои се јавуваат кај фактовидните прашања, а кај прашањата со набројување стануваат особено значајни, се:

- потребата од определување на праг над кој генерираните одговори се смета дека не соодветствуваат на прашањето, и
- неможноста да се определи кога два одговора се еквивалентни [4].

Описни прашања. Описните прашања побаруваат елаборација на одредени специфични настани и опишување на ентитети. Одговорите на овие прашања не се именувани ентитети и побаруваат напредни техники за обработка на природниот јазик, со цел да се изврши подлабока анализа на текстовите заради нивно генерирање [42]. Дизајнот на системите кои ги одговараат описните прашања е исклучително предизвикувачка задача. Проблемите поврзани со нивната ефикасност произлегуваат од системите за прибирање информации кои се имплементирани во нив, а се базирани на моделот множество од зборови. Главниот недостаток на овој модел е неможноста да ги идентификува семантичките информации на зборовите [43]. Уште повеќе, описните прашања имаат одговори кои може да бидат изразени во една реченица, до одговори кои опфаќаат и цел пасус.

2.2.2.1. Системи за рударење на мислењата

Освен најзабележителните категории на прашања наведени во претходно дадената класификацијата, неминовно е да се забележат и прашањата кои се однесуваат на субјективните мислења, односно субјективните информации за одредени ентитети или настани [44]. За одговор на овие прашања, *QA*-системите го користат социјалниот веб (*social web*) и во нив се имплементирани **техники за рударење на мислења** (*opinion mining techniques*). Еден од најзначајните системи за рударење на мислења е *SenticNet*[18]. Овој систем извршува задачи како: утврдување на поларитетот на ставовите[19] (*polarity detection*) и препознавање на емоција (*emotion recognition*), користејќи ја моќта на семантиката и лингвистиката, наместо употребата исклучиво на фреквенциите на зборовите кои се појавуваат [45], [46]. Накратко кажано, *SenticNet* ги обединува најновите откритија во анализата на мислењето на концептуално ниво. Нуди едноставни за користење, модерни алатки за анализа на огромното количество податоци од социјалните мрежи, со што се овозможува автоматизација на активностите како: позиционирање на брендот, откривање на тековните трендови, како и автоматизација на различните методи за маркетинг преку социјалните медиуми.

Системите за рударење на мислењата денес се исклучително полезни за корисниците. Тие овозможуваат анализа и толкување на масивните податоци генерирани од корисниците на социјалните мрежи, блоговите, како и сајтовите за давање оценки. Овие податочни извори претставуваат јавни мислења кои можат да им помогнат на корисниците во донесувањето одлуки при купување на производи, да направат анализа на местата кои сакаат да ги посетат, итн. [47].

Од друга страна, системите за рударење на мислењата се соочуваат со низа

[18] http://sentic.net
[19] Проценка дали искажаното мислење е позитивно, негативно или неутрално.

предизвици, меѓу кои најзабележителни се следниве:

- **Обработка на неформалните прашања**: Системите се соочуваат со проблем при обработка на прашањата од обичните корисници кои се формулирани на неформален јазик. Имено, овие прашања се семантички сиромашни и системите тешко успеваат морфолошки и синтаксички да ги разложат.
- **Класификација на текстот во кој се содржат мислења**: Истражувачките тимови сè уште дефинираат начини како да се утврди дали напишаното мислење е објективно или субјективно [48].
- **Утврдување на границите на реченицата**: Коментарите на корисниците обично се напишани на неформален јазик и не ги следат правилата за интерпункција, затоа истите се тешки за обработка од *QA*-системите.
- **Детекција на лажна содржина во текстот**: Лажната содржина претставува пречка за веродостојно рударење на текстот со мислења.

2.2.3. Класификација на *QA*-системите врз основа на типот на анализа која се врши врз прашањата

Класификација на *QA*-системите се врши и врз основа на анализата која ја прават врз корисничките прашања, како и врз изворните документи. Анализата може да биде: морфолошка, синтаксичка, семантичка, прагматична и дискурс анализа (*discourse analysis*).

Морфолошка анализа. Системите за прибирање информации (како и *QA*-системите) често користат алатки за морфолошка анализа на прашањата и документите од каде се извлекуваат одговорите. Тоа значи определување на зборовната група на зборот, неговите зборформи и коренот. Она што е важно за истражувачите е одлуката како да ги искористат овие информации, со цел да се постигне поголема точност при автоматското одговарање на прашањата. Досегашните истражувања покажуваат различни постигнувања при имплементација на овие алатки за различни јазици (секција 3.2 и 3.3). Поради тоа, нивната примена кај *QA*-системите побарува претходна анализа за нивното влијание врз добиените резултати.

Синтаксичка анализа. Синтаксичката анализа ја идентификува функцијата (службата) на зборовите во реченицата. Обично, секоја реченица се состои од **менливи зборови** (познати и како полнозначни зборови - именки, глаголи, придавки, заменки и броеви) кои се поврзани со **неменливите зборови** (службени зборови - сврзници, предлози, прилози, честички, извици и модални зборови). При синтаксичка анализа на прашањата и изворните документи, *QA*-системот генерира **синтаксичко дрво** (*parse tree*) кое го дефинира начинот на поврзување на зборовите во рамките на една реченица. Користењето на ова дрво овозможува да се намали просторот за пребарување, што од друга страна доведува до поголема ефикасност. Исто така, синтаксичката анализа е пожелна и поради ефективно пребарување, бидејќи ги зема предвид различните зборовни групи на еден ист збор. Сепак, може да се случи примената на оваа анализа да генерира синтаксичка грешка. На пример, при синтаксичка анализа на реченицата „Наведете книга од авторката А.Н. која е најчитана

овој месец“, системот не е во можност со сигурност да утврди дали зборот „најчитана“ се однесува на „книга“ или на „авторката“.

***Семантичка анализа*.** Семантичката анализа го заклучува можното значење на прашањето кое системот треба да го одговори, врз основа на зборовите кои се појавуваат во него. Најчесто, за семантичка анализа се користи синтаксичкото дрво, генерирано со синтаксичката анализа, и со негова помош се интерпретира прашањето. Една од најважните техники за семантичка анализа е **означувањето на семантичките улоги** во текстот (*semantic role labeling*). **Означувањето на семантичките улоги** е задача од обработката на природните јазици, која се состои од утврдување на неопходните реченични членови на глаголите[20] во реченицата и нивна класификација [49]. Оваа задача е на повисоко ниво на апстракција од синтаксичкото дрво, бидејќи две реченици може да имаат различна синтаксичка структура, но исто семантичко значење.

Се покажува дека имплементацијата на семантичката анализа кај *QA*-системите обезбедува поефективно пребарување на одговорите на прашањата, во споредба со пребарувањето базирано на клучни зборови. Недостатоците од примена на оваа анализа се следни:

- Денешните *QA*-системи функционираат на лексичко ниво и ниво на реченица за определување на значењето на текстот [7]. До овој момент, сè уште не постои истражување каде семантичката анализа се врши на ниво на документ.
- Проблемите кои се појавуваат при разрешување на кореференците, идентификација на именуваните ентитети, извлекување на релациите, означување на зборовните групи, итн., предизвикуваат потешкотии при извршување на семантичката анализа на текстот.

Прагматична и дискурс анализа. При овие анализи, прашањата и документите се интерпретираат на ниво на надреченична целина. Така, **прагматичната анализа** го разгледува и контекстот во кој се појавува одредена реченица. Од друга страна, **дискурс анализата** ја зема предвид и структурата на самиот документ, со цел да „разбере“ која е специфичната улога на одредена информација во документот, на пример, дали е заклучок, мислење, предвидување или факт [50]. Ваквите типови на анализа генерално се потребни при пребарување на подолги одговари на комплексни прашања, како прашања кои започнуваат со прашалните зборови „зошто“ и „како“, и помагаат да се извлече значењето од текстот.

[20] Во македонскиот јазик, станува збор за предметите (директен, индиректен и предмет со прилог) и прилошките определби.

2.2.4. Класификација на *QA*-системите врз основа на карактеристиките на податочните извори

Клучните критериуми за класификација на *QA*-системите врз основа на карактеристиките на податочните извори се: големината на податочниот извор, природниот јазик применет во изворот, хетерогеноста, жанрот и медиумите вклучени во изворот [33].

Големина на податочниот извор. Задачата за наоѓање одговори на прашања поставени на природен јазик е тесно поврзана со големината на изворот податоци низ кој се пребарува одговорот. Огромните колекции со податоци имаат свои предности и недостатоци. Така, доколку една колекција која се анализира содржи голем број документи, тогаш веројатноста да се најде одговорот на поставеното прашање е секако поголема. Исто така, поголемиот број појавувања на одговорот во различни документи ја зголемува оправданоста за неговата точност. Ваквите колекции се поволни за примена на статистичкиот пристап и пристапот базиран на правила, бидејќи овие методи постигнуваат повисока точност со зголемување на изворот податоци. Од друга страна, огромните колекции побаруваат повеќе време за нивна обработка. Одредени *NLP*-алатки дури и се преоптоваруваат, како резултат на огромниот број податоци кои треба да ги обработат.

Природниот јазик применет во изворот. Доколку во податочните извори се применети повеќе јазици, тогаш задачата за генерирање на одговорите е исклучително тешка. Имено, не постои единствен начин за разбирање на сите природни јазици, бидејќи секој од нив следи специфични правила и има своја карактеристична морфологија и синтакса. И покрај фактот што овие извори содржат информации раштркани низ различни јазици, кои може да се комбинираат за да се добие повеќе знаење, за нивна обработка е неопходно да се применат различни техники.

Хетерогеноста. Огромни количества информации се зачувани на различни локации и во различни формати. За обработка на различните типови податочни извори не постои единствен модел, и заради тоа истите претставуваат исклучителен предизвик за *IR* и *QA*-системите. Овде треба да се потенцира и потребата од генерирање на соодветни типови прашалници во зависност од податочниот извор. Добрата страна поврзана со вклучување на хетерогените извори во процесот на автоматско одговарање прашања е огромното количество информации кое го содржат (иако во различен формат, како бази на податоци, текстуални документи, мултимедиски документи, итн.), а кое може да се искористи за откривање на одредено специфично знаење.

Жанрот. Јазикот применет во одреден податочен извор може да биде лингвистички правилен или неправилен (формален или неформален). Неформалниот јазик е исклучително предизвикувачки да се обработи, бидејќи истиот не следи никакви формални правила. Во овој случај, најчесто се генерира погрешно синтаксичко дрво, заради што прибирањето на точните одговори е навистина проблематично.

Медиумите вклучени во изворот. Најголем дел од истражувањата направени до денес во *QA*-областа се однесуваат на наоѓање специфична информација во текстуални документи. Новите перспективи се насочени кон прибирање информации од мултимедиски содржини (како видео, аудио, звук), што е значително попредизвикувачко во однос на прибирањето информации од текст [4]. Сепак, се очекува во блиска иднина напредокот на технологијата да овозможи широка примена на техниките за прибирање информации од мултимедиски содржини.

2.2.5. Класификација на *QA*-системите врз основа на генерираниот тип на одговор

Постојат два клучни типа на одговори кои ги генерираат *QA*-системите за да ја задоволат потребата од информација на корисникот, и тоа: извлечен одговор и одговор како резултат на сумирање на текст [33].

Извлечен одговор. Ваквите одговори се во форма на кратки текстуални сегменти (најчесто фактовидни информации), или во форма на пасуси (кои претставуваат кратки дефиниции или описи) извлечени од документи од колекцијата.

Одговор како резултат на сумирање на текст. Сумирањето (*summarization*) претставува процес на извлекување на најважните информации од текстот со цел да се создаде скратена верзија за одредена тема [51]. Притоа, сумирањето може да се примени на единствен документ или на повеќе документи. Во првиот сличај, целта е да се генерира резиме на самиот документ, со кое се карактеризира неговата содржина. Од друга страна, сумирањето на повеќе документи подразбира продуцирање на кондензирана содржина на целата група.

2.3. Значајни пристапи за анализа на природните јазици

Ефективноста на системите за одговарање прашања во голема мера зависи и од тоа колку добро корисниците ги формулираат своите прашања. Фактот дека природните јазици се богати со двосмисленост, дополнително ја зголемува комплексноста на *QA*-системот, односно потребата од точна логичка репрезентација на прашањата поставени на природен јазик. Во литературата се забележуваат четири значајни пристапи кои ги применуваат *QA*-системите за анализа на прашањата и изворните документи, и тоа: учење од податоци (*learning from data*), лингвистички пристап (*linguistic approach*), пристап базиран на правила (*pattern matching approach*) и хибриден пристап (*hybrid approach*) [39], [52].

2.3.1. Учење од податоци

Учењето од податоци претставува заедничка цел на пристапите за статистичко моделирање и машинско учење. За разлика од статистичкото моделирање, кое во практични задачи се применува со векови, важноста на машинското учење доаѓа до израз со забрзаниот пораст на достапните он-лајн текстуални архиви и веб податоци. Алгоритмите за машинско учење можат успешно да се справат со огромното количество „големи податоци“ (*big data*) и нивната хетерогеност. За разлика од нив,

статистичките методи генерално се применуваат врз проблеми со помала димензионалност. Заради поголема прецизност, методите за учење од податоци побаруваат задоволително количество податоци, но штом се добро обучени, даваат подобри резултати од останатите пристапи. Она што е уште позначајно е дека наученият метод лесно може да се прилагоди на нов домен и не зависи од природниот јазик. Сепак, главен недостаток на овие пристапи е тоа што секој термин (кој најчесто е збор) го сметаат за независен. Тоа им оневозможува да идентификуваат одредени зборовни состави, како на пример синтагмите, а не го земаат во предвид и контекстот во кој се појавува терминот (за што е неопходно дополнително испитување). И покрај ова, општо земено, учењето од податоци денес успешно се применува во трите клучни модули на *QA*-системите.

2.3.1.1. Учење од податоци во модулот за обработка на прашањето

Како што е потенцирано во секцијата 2.1.1, акцентот во првиот модул е ставен врз класификација на прашањата кои системот треба да ги одговори (*Question classification* - *QC*). Во повеќедеценискиte истражувања во *QA*-областа, се дадени повеќе формални дефиниции кои се однесуваат на поимот **класификација на прашањата** [53], [54], [55]. Така, *Sundblad* [56] ја дефинира класификацијата на прашањата како задача за доделување на булова вредност на секој пар $\langle q_j, c_i \rangle \in Q \times C$, каде $Q = \{q_1, q_2, \dots, q_{|Q|}\}$ е множеството прашања, додека $C = \{c_1, c_2, \dots, c_{|C|}\}$ е множеството претходно дефинирани категории, каде $j = \overline{1, |Q|}$, $i = \overline{1, |C|}$. Доделувањето вредност T на парот $\langle q_j, c_i \rangle$ потврдува дека q_j припаѓа во категоријата c_i . Во спротивно, доделувањето на вредност F потврдува дека q_j не припаѓа во категоријата c_i . Најчесто се дозволува припадност на прашањата само во една категорија [11], [57], [58].

Анализите направени врз различни истражувања потврдуваат дека надгледуваните пристапи за учење се најмногу применувани во овој модул. Тие обучуваат класификатор со помош на дадено множество за обука, кое се состои од означени прашања. Надгледуваните пристапи, пред сè, се разликуваат во изборот на класификатор, како и од карактеристиките кои се извлечени од прашањата. Различни истражувања избираат различни класификатори. За класификација на прашањата најмногу се користат: машините со носечки вектори (*Support Vector Machines* – SVM), Бајесовите класификатори (*Bayesian Classifiers*) и моделите со максимална ентропија (*Maximum Entropy Models*). Исто така, исклучително важен проблем е како да се извлечат карактеристиките и кое е оптималното множество карактеристики. Во задачата за класификација на прашањата се забележуваат следниве типови карактеристики:

- **Лексички карактеристики** на прашањето, кои се извлекуваат од зборовите во прашањето. Наједноставен пристап за нивно дефинирање е користењето на униграм карактеристиките (пристап повеќе познат како „множество од зборови“), при што секој збор од прашањето претставува посебна карактеристика. Униграм карактеристиките се специјален случај од

таканаречените n – грам карактеристики. Важноста на карактеристиката се дефинира со задавање на одредена тежина (како на пример, честотата на појавување). Од другите лексички карактеристики значајни се: прашалниот збор (како карактеристика сама за себе) [59], формата на зборовите (зборови со мала, голема буква, мешани, цифри и други) [58], како и бројот на зборови во прашањето [60].

- **Синтаксички карактеристики**, кои се извлекуваат од синтаксичката структура на прашањето. Сепак, најчесто користени синтаксички карактеристики се зборовната група и главниот збор во прашањето (*headword*) (најинформативниот збор во прашањето, односно зборот кој специфицира што се бара во прашањето). Истражувањата на *Silva et al.* [11] и *Loni et al.* [61] откриваат дека главниот збор припаѓа во најуспешните карактеристики за класификација на прашањата.
- **Семантички карактеристики**, кои се извлекуваат врз основа на значењето на зборовите во прашањето. Нивното дефинирање побарува дополнителни извори, како *WordNet*, или располагање со речник од зборови. Најчесто користени семантички карактеристики се: хипернимите (*hypernymys*), поврзаните зборови (*related words*) и именуваните ентитети.

Релевантни истражувања врз TREC-множеството податоци. Во овој дел е даден преглед на истражувањата чиј фокус е ставен врз класификација на прашањата, со примена на пристап за учење од податоци, а за проценка е искористено *TREC*-множеството податоци (табела 1).

– *Zhang et al.* [57] експериментираат со пет алгоритми за машинско учење, и тоа: методот на најблиски соседи (*Nearest Neighbors*), наивниот Бајесов метод (*Naïve Bayes*), дрвата на одлучување (*Decision Trees*), ретката мрежа за одбирање (*Sparse Network of Winnows*) и машините со носечки вектори (*SVM*), за кои користат два типа карактеристики: униграм и n – грам. Експерименталните резултати покажуваат дека со овие едноставни карактеристики, *SVM* со линеарен кернел ги надминува останатите класификатори. Уште повеќе, тие предлагаат специјална кернел функција, наречена **кернел за дрва** (*tree kernel*), со цел да се овозможи *SVM* да ја искористи и синтактичката структура на прашањата.
– Наместо богат простор од карактеристики, *Huang et al.* [58] предлагаат користење на компактно и ефективно множество карактеристики, со кое успеваат да добијат повисока точност во однос на претходните истражувања.
– *Loni et al.* [61] развиваат нов класификатор на прашања базиран на учење. Тие анализираат неколку лексички, синтаксички и семантички карактеристики и ја испитуваат нивната полезност. Анализите потврдуваат дека највисока точност се добива со тежинско комбинирање на шест множества карактеристики (табела 1). Оптималните тежини се добиени со поединечно комбинирање на секое множество карактеристики само со униграм карактеристиките.
– *Silva et al.* [11] користат комбинација од пристап базиран на правила (секција 2.3.3) и пристап базиран на учење. Информациите добиени од класификаторот базиран на

правила (главниот збор и категоријата) се ползуваат за обучување на класификатор преку генерирање на множество карактеристики за обука, заедно со униграм карактеристиките на прашањето. Добиените резултати потврдуваат дека оваа комбинација овозможува успешно класифицирање на прашањата кои рачно дефинираните правила не успеваат да ги поврзат со ниту една категорија.

- *Chen et al.* [62] ја испитуваат улогата на семантичките карактеристики (како, именуваните ентитети, паровите зборови кои се во одредена релација во согласност со *WordNet* и семантичките класи) и предлагаат два различни кернели за дрва, кои ги вклучуваат овие карактеристики во *SVM*-моделот. Секој кернел се покажува соодветен за класификација на одредени категории прашања. Со цел да се искористат поволностите од двете функции, авторите го користат пристапот за учење со повеќе кернели (*Multiple Kernel Learning*) за нивно комбинирање.
- За класификација на прашањата, *Hardy et al.* [63] користат машина за екстремно учење (*Extreme Learning Mashine - ELM*). Овој класификатор имплементира семантички карактеристики, со што драстично се намалува димензионалноста на просторот од карактеристики. Од друга страна, намалувањето на димензионалноста придонесува *ELM*-класификаторот да биде значително побрз во споредба со *SVM*, и тоа без значајни промени на точноста.
- *Van-Tu et al.* [64] се фокусираат на проблемот како да се извлечат и селектираат ефикасни карактеристики соодветни за различните типови прашања (типовите прашања се поврзани со прашалните зборови: „кој“, „каде“, „што“, „кога“, „чиј“, „зошто“ и „како“). За таа цел, авторите предлагаат алгоритам за селекција на најсоодветното множество карактеристики (лексички, синтаксички и семантичките) за секој тип прашања и дизајнираат нов тип карактеристики кои се базираат на шаблони за прашањата. Од табелата 1 се забележува дека ваквиот избор на карактеристики од прашањата засега дава најдобар резултат на *TREC*-множеството податоци.

Истражување	Класификатор	Карактеристики	Точност	
			Груба	Фина
Zhang et al. [57]	*SVM* со кернел за дрва	Поддрва	90.0%	/
	SVM со линеарен кернел	*N* - грами од зборови	87.4%	79.2%
Huang et al. [58]	Максимална ентропија	У+ПЗ+ФЗ+ГЗ+Х+ИХ	93.6%	89.0%
	SVM со линеарен кернел	У+ПЗ+ФЗ+ГЗ+Х+ИХ	93.4%	89.2%
Loni et al. [61]	*SVM* со линеарен кернел	У+Б+ФЗ+ГЗ+Х+П	93.6%	89.0%
Silva et al. [11]	*SVM* со линеарен кернел	У+ГЗ+Х+ИХ	95.0%	90.8%
Chen et al. [62]	Учење со повеќе кернели	ИЕ+С+Х+СК	95.8%	/
Hardy et al. [63]	Машина за екстремно учење	ПЗ+ГЗ+Х	92.8%	84.6%
Van-Tu et al. [64]	*SVM* со линеарен кернел	У+Б+ФЗ+ГЗ+П+ПП+КП+ШП	95.2%	91.6%

Табела 1. Споредба на различни надгледувани пристапи за учење од податоци, применети за класификација на прашањата од *TREC*-множеството податоци

Скратеници: У: Униграми, **Б**: Биграми, **НГ**: *N* - грами, **ПЗ**: Прашален збор, **ФЗ**: Форма на зборот, **ГЗ**: Главен збор, **Х**: Хиперними, **ИХ**: Индиректни хиперними, **С**: Синоними, **ИЕ**: Именувани ентитети, **П**: Поврзани зборови, **СК:** Семантички класи, **ПП:** Проширување на прашање, **КП:** Категорија на прашање, **ШП:** Шаблони за прашања

Релевантни истражувања врз други множества податоци. Наредните неколку истражувања се однесуваат на класификација на прашањата, чија проценка не е направена врз *TREC*-множеството податоци. Ова особено е значајно за развој на системи за одговарање прашања на други јазици, како европските јазици, но и развој на системи за одговарање прашања на турски, персиски, малезиски, и други. Најголем дел од неодамнешните истражувања се однесуваат на развој на системи за одговарање прашања на кинески јазик.

- Така, системот за одговарање прашања поставени на кинески јазик, развиен од *Zhang et al.* [65], имплементира машини со носечки вектори (*SVM*) кои користат четири карактеристики, и тоа: зборови, зборовна група, именувани ентитети и семантика. Експериментите се направени врз вкупно 5514 прашања, најголем дел извлечени од базата на податоци на Институтот за технологија *Harbin* (*Harbin Institute of Technology*), како и прашања извлечени од Интернет форум, заради зголемување на разновидноста. Анализирајќи ги потребите на својот *QA*-систем, авторите дефинираат сопствена таксономија која се состои од 7 груби и 50 фини категории. Направените анализи покажуваат дека нивниот систем врши квалитетна класификација на прашањата, односно системот успева точно да класифицира 92% од прашањата во грубата категорија и 85% во фината категорија.
- *Fan et al*. [38] предлагаат нов класификатор на прашања соодветен за општ *QA*-систем, односно систем способен да одговара прашања од различни извори. Новиот класификатор го комбинира пристапот базиран на правила и статистичкиот пристап претставен од Марковата логичка мрежа (*Markov Logic Network* - *MLN*). Со цел да потврдат дека нивната методологија е соодветна за одговарање прашања од различни извори, класификаторот го обучуваат и тестираат врз прашања преземени од два кинески сервиса на заедницата за одговарање прашања (5800 прашања за обука и 1400 прашања за тестирање). Резултатите потврдуваат дека предложениот класификатор е супериорен во однос на традиционалните пристапи за класификација на прашања (како *SVM* и *MLN*), за кои како карактеристики се избрани биграмите.
- *Mollaei et al*. [66] предлагаат класифицирање на прашања поставени на персиски јазик користејќи ги условените случајни полиња (*Conditional Random Fields* - CRF). Карактеристиките кои ги користат се: главниот збор, зборовите од прашањето (земајќи го во предвид и нивниот редослед), прашалниот збор, *N* – грамите од зборови, зборовната група и други. Авторите дефинираат таксономија од 6 груби и 58 фини категории, а анализите ги прават врз множество од 5000 прашања извлечени од материјалите за основно и средно образование, како и често поставувани прашања од неколку веб-локации. Точноста која ја добиваат е 80.35% за грубите и 78.71% за фините категории.

2.3.1.2. Учење од податоци во модулот за прибирање пасуси

Многу од системите за одговарање прашања користат техника за прибирање пасуси со цел да се олесни процесот на наоѓање на точниот одговор.

Релевантни истражувања. Во следниов дел се наведени неколку истражувања кои за прибирање на пасуси вклучуваат пристап за учење од податоци. Табелата 2 дава краток преглед на постигнувањата од овие истражувања.

- Во своето истражување *Othman et al*. [67] предлагаат нов пристап за прибирање пасуси и подобрување на рангирањето кој не зависи од природниот јазик. Овој пристап најпрво користи мерка за сличност базирана на N – грами, при што начинот на нивно извлекување е комплетно различен од претходните приоди. Мерката ги вклучува само N – грамите кои ги споделуваат прашањето и пасусите кои се испитуваат (како и поднизите од низата N – грам). Уште повеќе, тежината на N – грамите зависи од нивната должина и тежините на термините кои се појавуваат во нив. Тогаш, сличноста меѓу прашалникот и пасусот се дефинира како однос на тежината на пасусот и тежината на прашалникот. Со помош на оваа мерка се прибираат пасуси кои веројатно го содржат точниот одговор. Со цел да се гарантира прибирање на високо релевантни пасуси, авторите го интегрираат и *SVM*-моделот, кој вклучува лексички, синтаксички и семантички карактеристики.
- Системот за одговарање прашања развиен на *Carnegie Mellon University* (*CMU Open Advancement*) [68] ги користи фрагментите прибрани од различни веб-пребарувачи и ги рангира врз основа на оценувач на релевантноста. Како најефективен оценувач се покажува хеуристичката комбинација за задавање тежини која ги вклучува *Okapi BM25* формулата и пристапот базиран на рекурентна невронска мрежа [69], [70].
- За рангирање на пасусите во *RMIT* системот, *Mackenzie et al*. [71] имплементираат дрва на одлука со градиентен раст (*Gradient Boosted Decision Trees*), за кои користат неколку карактеристики: бројот на термини од прашалникот кои се појавуваат во пасусот, бројот на нивни синоними кои се појавуваат во пасусот, бројот на термини во пасусот, дали прашалникот се јавува како потстринг во пасусот, итн. Овој систем се покажува значително подобар од просекот на сите системи кои учествуваат на *LiveQA*-работилницата 2016.
- *Murdock et al*. [72] имплементираат едноставен транслациски модел (*translation model*) за прибирање на пасуси на ниво на реченица.
- *Gomez-Soriano et al*. [73] опишуваат нов едноставен модел за прибирање пасуси, исклучително погоден за *QA*-системите. Моделот се заснова на структурата на прашањето и ги фаворизира пасусите кои содржат што подолга низа од N - грами (зборови) од прашањето. Уште позначајно за овој модел е што не користи лингвистички информации, што значи дека е независен од природниот јазик врз кој се применува. Експерименталните резултати на шпанскиот, италијанскиот и францускиот јазик ја потврдуваат оваа негова карактеристика и покажуваат дека предложениот модел е стабилен за различни јазици.
- Статистичкиот систем за одговарање прашања на *IBM* [74] имплементира *IR*-модул кој се состои од две фази. Во првата фаза модулот користи модифицирана *Okapi*

формула, со цел да се рангираат пасуси од енциклопедиски документи. Користејќи ги најдобро рангираните пасуси, системот врши проширување на прашалниците за пребарување со помош на модификација на техниката анализа на локален контекст (*Local Context Analysis – LCA*). Во втората фаза проширените прашалници се користат за рангирање на пасуси од *TREC-9 QA*-корпусот.

Учење од податоци	Истражување	Карактеристики
Комбинација од сличност базирана на *N* – грами и *SVM*	*Othman et al.* [67]	Покажува значително подобрување во однос на неколку предложени пристапи, проценети на *CLEF*[21] *RezPubliQA 2010* колекцијата [75]
Комбинација од *BM25* и рекурентна невронска мрежа	*Wang et al.* [68]	*CMU OAQA* се покажува подобар од сите учесници на *TREC LiveQA*-работилницата 2015 [76]
Дрва на одлука со градиентен раст (*Gradient Boosted Decision Trees*)	*Mackenzie et al.* [71]	*RMIT* системот се покажува значително подобар од просекот на сите системи учесници на *TREC LiveQA*-работилницата 2016
Транслациски модел	*Murdock et al.* [72]	Покажува значително подобрување во однос на моделот на јазик без користење на Веб, надворешни лексикони или онтологии
Сличност базирана на *N* – грами	*Gomez-Soriano et al.* [73]	Значително подобар во однос на моделот на векторски простор (при прибирање на пасуси со една и три реченици). Проценката е направена на *CLEF*-множеството податоци 2004
Модифицирана *Okapi* формула со *LCA* техника за проширување на прашалник	Статистичкиот *QA*-систем на *IBM* [74]	Проширувањето на прашалникот дава подобрување на мерката среден реципрочен ранг (*Mean Reciprocal Rank – MRR*)

Табела 2. Пристапи за учење од податоци применети во фазата за прибирање на пасуси

2.3.1.3. Учење од податоци во модулот за обработка на одговорите

Постојат различни начини за извлекување и валидација на одговорите кои се користат во последната фаза од *QA*-системот. Сепак, последниве години клучна улога во модулот за обработка на одговорите имаат методите базирани на карактеристики, како: невронските мрежи [77], максималната ентропија [78], *SVM* [79], логистичката регресија (*Logistic Regression*) [80], и други. Целта на овие методи е да ја определат „блискоста“ меѓу кандидатите за одговор и прашањето, врз основа на различни карактеристики за сличност.

Релевантни истражувања. Во табелата 3 е даден преглед на неколку познати истражувања од *QA*-областа, кои за извлекување и валидација на одговорот, користат пристап за учење од податоци.

– *Berger et al.* [81] ги испитуваат можностите за примена на статистички методи со цел да се премости лексичкиот јаз меѓу прашањата и одговорите, во фазата на извлекување одговор на поставено прашање кај *QA*-системите. Во своето емпириско истражување анализираат четири статистички техники, и тоа: прилагодената $tf - idf$ (*adaptive* $tf - idf$), проширување на прашалникот (преку

[21] The Cross-Language Evaluation Forum; http://clef.iei.pi.cnr.it/

имплементација на заедничката информација (*mutual information*), искористена за да се научи пресликување меѓу термините од прашалникот и одговорот), транслациски модел (изведен од фамилијата транслациски модели на *IBM* [82]) и модел на латентна променлива (*latent variable model*). Квантитативната анализа е направена врз две реални множества податоци: *Usenet FAQs* [22] и множество прашања поставени од корисници до одредена компанија за продажба. Притоа, анализите откриваат дека сите овие техники даваат значително подобри резултати отколку основниот $tf - idf$ модел. Тие потврдуваат и дека во зависност од карактеристиките на множеството податоци (како големината на речникот, до која мера се совпаѓаат прашањето и одговорите, до која мера се совпаѓаат повеќекратните одговори, итн.) и посакуваното ниво на перформанси, секоја од овие техники може да се покаже како најдобра во одредена ситуација.

– Во своето истражување, *Soricut et al*. [83] опишуваат систем за одговарање прашања без ограничување на типот на прашањата. Во фазата за извлекување на одговорот авторите разгледуваат два различни алгоритми. Како основа користат нов алгоритам за извлекување одговори кој не го премостува лексичкиот јаз меѓу стринговите на прашањето и одговорот, а се заснова на N - грами од термини кои заедно се појавуваат во нив. Вториот алгоритам ја користи лексичката корелација меѓу прашањата и одговорите добиена со помош на техники за статистичко машинско преведување (*Statistical Machine Translation*), со цел да се најде најдобриот одговор за дадено прашање [84]. Проценката на системот е направена врз корпус од еден милион парови прашање/одговор, извлечени од страници со често поставувани прашања (*FAQ*) достапни на веб. Истата покажува дека системот постигнува пристојна точност користејќи го корпусот од разновидни, нефактовидни прашања.
– *Cai et al*. [85] презентираат систем за одговарање прашања поставени на кинески јазик, при што како извор на информации ги користат првите сто фрагменти прибрани од веб-пребарувачот Гугл. Системот користи метод за извлекување на одговор кој се заснова на определување на сличноста меѓу реченицте (*sentence similarity*) кои ги претставуваат прашањето и одговорот. Во овој метод сличноста се определува како линеарна комбинација од фактори кои ги земаат предвид: бројот на заеднички клучни зборови, должината на реченицте, редоследот на зборовите во двете реченици и растојанието меѓу клучните зборови во нив. Она што е нова карактеристика за овој систем е неговиот метод за валидација на одговорите (*answer validation*). Со цел да се овозможи селекција на одговор од мноштвото извлечени одговори во претходната фаза, овој метод ја определува и корелацијата меѓу прашањето и кандидатот за одговор, користејќи ја заедничката информација (*mutual information*). Проценката на системот е направена врз дел од прашањата за тестирање од *HIT-IRLab*[23]. При тоа, авторите забележуваат дека вклучувањето на корелацијата при валидација на одговорите доведува до 3.3% релативно подобрување на точноста на системот.

[22] http://www.faqs.org
[23] Information Retrieval laboratory, Harbin Institute of Technology

– Во своето истражување, *Verberne et al.* [86] испитуваат неколку техники за машинско учење за задачата рангирање на одговорите на прашања кои започнуваат со прашалниот збор „зошто“, заедно со $tf - idf$ моделот како основа за анализа. За техниките за машинско учење, авторите користат множество од 36 лингвистички карактеристики кои ги опишуваат прашањата и идентификуваните одговори од претходниот модул. Целта е да се направи споредба на следниве техники: наивниот Бајес, класификацијата со носечки вектори (*Support Vector Classification*), регресијата со носечки вектори (*Support Vector Regression*), логистичката регресија, рангирачкиот *SVM* (*Ranking SVM*) и генетскиот алгоритам (*Genetic Algorithm*). Во истражувањето се користи *Wikipedia* корпусот од *INEX 2006* [87] и множество од 186 прашања од *Webclopedia* [88] за кои постои барем еден одговор во корпусот. Анализата на резултатите потврдува дека за сите испитани техники се добива *MRR* вредност значително подобра од $tf - idf$ моделот.
– Стремежите на *Agarwal et al.* [89] се насочени кон подобрување на рангирањето на кандидатите за одговор за фактовидните прашања. Испитувањето е направено на две податочни множества, и тоа: прашањата од квиз шоуто *Jeopardy* и множество со медицински податоци. Во ова истражување авторите ги имплементираат следниве статистичките методи: логистичка регресија, *RankBoost* [90], *AdaRank* [91], *Coordinate-Ascent* [92], *LambdaRank* [93], при што заклучуваат дека нивната успешност зависи од големината на множеството за обука, како и податочното множество врз кое се применуваат. Затоа, наместо примена на единствен метод за рангирање, предлагаат каскаден пристап, каде резулатите добиени од еден метод се користат како влез во друг. Анализите покажуваат дека ваквиот пристап не само што обезбедува подобрување на резултатите, туку и го намалува времето за предвидување.

Учење од податоци	Истражување	Карактеристики
Статистичко машинско преведување	*Soricut et al.* [83]	Задоволителни перформанси дури и за нефактовидни прашања
Модел на сличност на реченици	*Cai et al.* [85]	Методот се покажува поефективен од вообичаените методи за селекција на одговорите
Повеќе техники за машинско учење	*Verberne et al.* [86]	Испитаните техники се значително подобри од класичниот $tf - idf$ модел
Комбинација од техники за машинско учење	*Agarwal et al.* [89]	Системот е ефикасен и покажува значително подобрување во однос на резултатите добиени со *Watson*
Конволуциски невронски мрежи (*CNN*)	*Feng et al.* [94]	Супериорни резултати во споредба со косинус сличноста. Постигнува точност од 65.3% на ново-креирано множество податоци[24]

Табела 3. Пристапи за учење од податоци применети за извлекување и/или валидација на одговорите

[24] https://github.com/shuzi/insuranceQA.git

– Технологијата за **длабоко учење** (*Deep Learning*) широко се применува во задачи за машинско учење и често демонстрира супериорни перформанси во споредба со традиционалните методи. Последниве години сè повеќе е актуелна и кај *IR* и *QA*-системите. Пример претставува истражувањето на *Feng et al.* [94], кое применува длабоко учење во задачата за селекција на одговорот на нефактовидните прашања. Нивниот пристап не подлежи на примена на лингвистички алатки и е независен од јазикот и доменот. Се базира на конволуциските невронски мрежи (*Convolutional Neural Networks - CNN*), а добиените резултати претставуваат цврст доказ дека одговарањето прашања со помош на длабокото учење е насока за истражување која ветува.

2.3.2. Лингвистички пристап

Основата на првите системи за одговарање прашања лежи во методите од вештачката интелигенција (*AI*). Овие методи ги интегрираат техниките за обработка на природните јазици (*NLP*) и потребата од постоење на база (корпус) со знаење. Знаењето е организирано во формална репрезентација, како: логика, семантички мрежи, дијаграми за концептуална зависност или репрезентации засновани на шаблони. Врз барањето на корисникот најпрво се применуваат одредени лингвистички техники, како: токенизација (*tokenization*) [8], означување на зборовната група и синтаксичко разложување. Со тоа се овозможува негово трансформирање во прецизен прашалник, со кој може да се извлече соодветен одговор од структурираната база на податоци. Сепак, употребата на база со знаење за специфичен домен е лимитирачка, бидејќи различни домени побаруваат различни граматики и правила за пресликување. Уште повеќе, креирањето на соодветна база со знаење е долготраен процес, и затоа овие системи вообичаено се применуваат кај долгорочни барања за информации за одреден домен.

Првите *QA*-системи, како *Baseball* [1] и *Lunar* [2], претставуваат само интерфејс кон структурираните бази на податоци. Прашањата од корисниците се анализираат користејќи *NLP*-техники со цел да се создаде канонична форма, која потоа се користи за конструкција на стандарден прашалник за базата на податоци. Дијалог-системите (*dialogue systems*), како *ELIZA* [95] и *GUS* [96], исто така користат структурирана база на податоци како извор на знаење. Бидејќи кај овие системи знаењето зачувано во структурираната база на податоци е во можност да даде одговор на прашања во рамките на ограниченият домен, нивната употреба е ограничена. Значајно е да се забележи дека најголем дел од традиционалните *NL*-интерфејси кои вршат детална јазична обработка остануваат истражувачки прототипи, бидејќи нивната комерцијално додадена вредност не ги покрива трошоците.

Со цел да се прошири доменот, некои од постоечките *QA*-системи (како *START* [97], *QA*-системите развиени од *Chung et al.* [98] и *Mishra et al.* [99]), го користат вебот како извор на знаење. Овие системи применуваат свои хеуристики за да ги складираат информациите од веб документите во локална база на податоци, кон која се пристапува подоцна и се применуваат лингвистички техники за генерирање на одговорите. Денес,

лингвистичките техники длабоко се инкорпорирани во најсовремените системи за одговарање прашања.

2.3.3. Пристап базиран на правила

Развојот на *QA*-системите базирани на правила е прилично предизвикувачка задача, бидејќи истражувачот треба да ги земе предвид скоро сите теми за кои системот може да биде тестиран. Овие системи не користат длабинско разбирање на јазикот, ниту спефицични софистицирани техники. Од *NPL* техниките најчесто се применуваат: морфолошката анализа, означувањето на зборовна група и означувањето на семантичките улоги, со цел да се постигне повисока точност во процесот на прибирање одговори.

Релевантни истражувања. Рачно креираните правила се главниот пристап за класифицирање на прашањата во првиот модул од *QA*-системите. За нивно креирање, истражувачите најчесто ги земаат предвид прашалниот збор и одредени комбинации од зборовите во прашањето (како и нивните карактеристики), со цел да се определи неговата категорија и очекуваниот тип на одговор. Процесот на креирање на овие правила е долготраен и бара исклучителен напор, но анализите потврдуваат дека вака креираните правила се многу прецизни.

- *Silva et al*. [11] градат независен класификатор базиран на правила кој се состои од 60 рачно дефинирани правила. Притоа, секое прашање се споредува со овие правила и доколку се забележи совпаѓање, прашањето се класифицира во согласност со неговата категорија. Во спротивно, класификаторот се обидува да го дефинира главниот збор (*headword*) во прашањето. За разрешување на овој проблем, авторите го користат синтаксичното дрво на прашањето, имплементирајќи го Беркли разложувачот (*Berkeley Parser*) [100]. Врз дрвото се применува и множество од правила со цел да се утврди кој јазол го содржи главниот збор. На крај, категоријата на прашањето се добива користејќи ги хипернимите на главниот збор дефинирани со лексичката база *WordNet*, доколку некој од нив може да се поврзе со некоја од утврдените категории на прашањето. Користејќи ги само рачно дефинираните правила, класификаторот има точност од 87.0% за грубата и 83.2% за фината категорија врз *UIUC* множеството податоци. Притоа, треба да се потенцира дека за сите 38 некласифицирани прашања, множеството правила точно го извлекува главниот збор, но и со примената на *WordNet* не се утврдува нивната категорија (бидејќи главниот збор не се содржи во оваа лексичка база).
- За класификација на прашањата *Biswas et al*. [101] ја користат природата на прашањата. За таа цел, рачно анализираат колекција од 2000 прашања од податочното множество на *Li et al*. [13]. Седумте различни типови прашања утврдени во согласност со прашалните зборови: „кој“, „каде“, „што“, „кога“, „чиј“, „зошто“ и „како“, ги класифицираат грубо во три категории: дефиниција, опис и факт. Користејќи го означувачот на зборовни групи на *Stanford*[25] доаѓаат до

[25] https://nlp.stanford.edu/software/tagger.shtml

неверојатен заклучок дека прашањата кои припаѓаат во одредена категорија имаат иста синтаксичка структура. Заради тоа, предлагаат примена на синтаксички правила при класификација на прашања со кои постигнуваат одлични резултати.

Креирањето на правила се применува и при извлекување на точниот одговор на дадено прашање.

– *Soubbotin et al.* [102] дефинираат систем кој успешно користи правила за извлекување на одговорот. За секоја категорија на прашања генерираат низа од правила. Секое правило претставува низа од карактери, интерпункциски знаци, празни места, броеви или зборови. Правилата се добиваат рачно анализирајќи ги изразите, кои обично се одговори на соодветната категорија прашања. Уште повеќе, на секое правило рачно му се задава одредена вредност. На овој начин, системот може да одбере помеѓу неколку кандидати за одговор, во зависност од нивото на доверба дадено на секое правило соодветно на прашањето.

Некои од истражувачите градат правила користејќи подлабоки лингвистички информации. Сепак, исцрпувачкиот процес за рачно дефинирање на правила е причина последниве неколку години сè поголем акцент да се става на пристапот базиран на учење, или на комбинацијата од овие два пристапи.

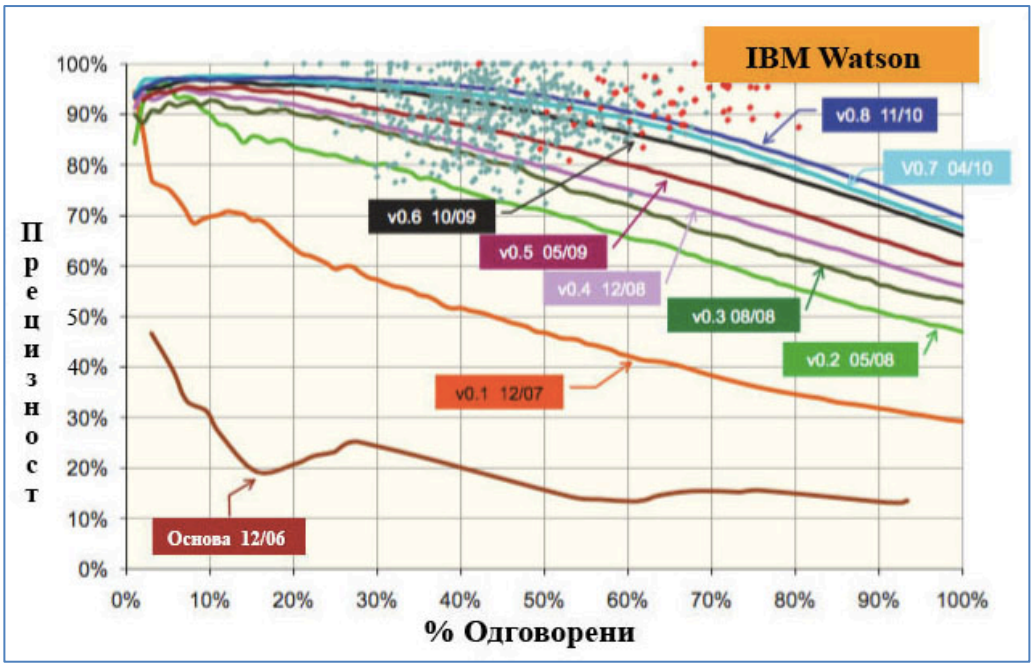
2.3.4. Хибриден пристап

Како што е веќе забележано, трите клучни пристапи се покажуваат прилично добро во рамките на она за што се развиени, но имаат одредени ограничувања доколку се применат надвор од нив. Токму овој факт последниве години доведе до развој на хибридни системи за одговарање прашања, кои ги надминуваат тие ограничувања и го користат потенцијалот на секој пристап. Анализите потврдуваат дека хибридниот пристап покажува различен успех при одговарање на различни типови прашања. Тоа значи дека е потребно вложување на уште поголеми напори за успешно интегрирање на лингвистичкиот пристап и пристапите базирани на учење и на правила, со цел системот ефективно да се справи со барањата на корисниците.

2.3.4.1. IBM Watson

Системот *IBM Watson* претставува најзначајното достигување во областа на хибридните системи за одговарање прашања, и воопшто во *QA*-областа. Тој е резултат на неколкугодишно интензивно истражување на *IBM*-истражувачкиот центар и се докажува како систем чија прецизност, доверба и брзина ги надминува најдобрите натпреварувачи на квизот *Jeopardy* [103]. Развојот на овој систем вклучува иновации од различни истражувачки области, како: обработката на природните јазици, прибирањето информации, машинското учење, компјутерската лингвистика и претставување на знаењето и заклучувањето. *Watson* користи повеќе од 100 клучни класификатори (наречени **експерти**), при што секој од нив се покажува ефективен на одреден тип прашања. Подетално, системот врши: анализа на прашањето поставено на природен јазик, идентификување на извори со соодветни информации, наоѓање и генерирање на хипотези (кандидати за одговор), наоѓање и вреднување на докази (*evidence*), спојување и рангирање на хипотези. Она што е поважно од самите

класификатори кои ги користи *Watson* е начинот на кој тие се комбинираат со цел таквиот пристап да даде највисока можна точност, доверба (*confidence*) и брзина.



Слика 3. Прогрес на прецизноста за одговарање на *Watson*, преземено од [104]

Како резултат на многуте напори направени во текот на неколку години, прецизноста на *Watson* успева да достигне до 85% за 70% прашања кои *Watson* се обидува да ги одговари, што е сосема доволно овој систем да се натпреварува на квизот *Jeopardy*. На сликата 3 е даден преглед на подобрувањето на прецизноста на *Watson*, во периодот од 2007 до 2010 година, во однос на предизвикот *Jeopardy*. Заради подобра споредливост, на сликата се претставени и перформансите на победниците во одредена *Jeopardy* игра (означени со поединечните точки на графикот).

Во делот кој следи накратко е опишан *Watson*-системот, при што посебен акцент е ставен врз она по што тој се разликува од останатите *QA*-системи.

Анализа на прашањето [105]. Во првата фаза *Watson* врши анализа на прашањето со цел да утврди на што се однесува истото и како најдобро да се пристапи кон негово одговарање. За реализирање на оваа задача, системот користи комбинација од техники, како: површно и длабинско разложување на прашањето [106], означување на семантичките улоги, разрешување на кореференците, дефинирање на релациите [107], утврдување на именуваните ентитети, итн. Применувајќи бројни класификатори и правила за детекција врз карактеристиките утврдени со оваа анализа, се овозможува откривање на клучни елементи од прашањето. Меѓу нив најважни се:

- **фокусот**, дефиниран како дел од прашањето кој доколку се замени со одговорот, го прави прашањето самостоен исказ,
- **лексичкиот тип на одговорот** (*lexical answer types - LAT*), дефиниран како збор (или фраза) од прашањето кој го означува типот на бараниот одговор, без идентификување на неговото значење,

- **класификација на прашањето** во една или повеќе категории (факт, дефиниција, скратеница, итн.), и
- утврдување на **елементи од прашањето кои имаат важна улога** и побаруваат посебно справување (како на пример, вгнездени потпрашања кои мора пооделно да бидат одговорени).

Откривањето на овие четири елементи побарува мошне опсежна и детална анализа на прашањето, во споредба со претходните истражувања кои вклучуваат само некои од нив. Дел од пристапите кои ги користи тимот на *Watson* за оваа цел се следниве:

– **Утврдување на фокусот и лексичкиот тип на одговорот (*LAT*)**. Анализирајќи примерок од 20000 прашања, тимот утврдува 2500 различни лексички типови на одговори (*LATs*). Притоа, се забележува дека најчестите 200 типови покриваат скоро 50% од податоците. Основната имплементација за нивно утврдување, како и утврдување на фокусот, се состои од примена на одреден број рачно дефинирани правила, кои се лесно воочливи од самите *Jeopardy* прашањата. Овие едноставни правила се покажуваат релативно прецизни. Сепак, во случаи на покомплицирани реченични конструкции, тие не се во можност да детектираат одредени, мошне корисни лексички типови на одговори. Заради нивно утврдување, *Watson* применува и обучен класификатор (логичка регресија). Резултатите потврдуваат дека со вклучување на статистичкиот класификатор, системот е во можност точно да одговори дополнителни 3.5% од прашањата од множеството за тестирање (кое содржи вкупно 3500 прашања).
– **Класификација на прашањето и утврдување на клучни сегменти од прашањето**. *Jeopardy* вклучува широк опсег прашања, заради што практично е невозможно на единствен начин да се пристапи кон нивно одговарање. Затоа е неопходно да се утврдат категориите во кои припаѓа прашањето (наречени *QClasses*), кои го водат процесот на неговото одговарање (примена на различни техники и модели за машинско учење, во зависност од тоа во која категорија (категории) припаѓа прашањето). Различните категории се утврдуваат со примена на различни техники кои се независни една од друга (на пример, клучни фрази, регуларни изрази, рачно дефинирани правила во комбинација со статистички класификатор кој ја носи конечната одлука, и други). Клучните сегменти од прашањето (наречени *QSections*) се идентификуваат со правила применети врз неговата синтаксичка структура или преку регуларни изрази. Анализите потврдуваат дека со вклучување на *QClasses* и *QSections* во анализата на прашањето, се подобрува севкупната точност на системот. Со нив *Watson* успева точно да одговори дополнителни 2.9% од прашањата од множеството за тестирање.

Имплементацијата на техниките од оваа фаза се покажува доволно ефикасна за да се анализира прашањето во дел од секунда, што е клучно *Watson*-системот да биде конкурентен на квизот *Jeopardy*.

Пребарување и генерирање на кандидатите за одговори [108]. Првиот чекор за разрешување на *QA*-проблемот е идентификување и прибирање на содржина, која *Watson* ја користи како извор за извлекување на одговорите. Корпусот се состои од неструктурирана и структурирана содржина, добиена во комбинација од мануелни и автоматски чекори. Неструктурираната содржина вклучува документи чии наслови се одредени концепти или ентитети (односно документи кои даваат информации за нив, како енциклопедии, речници, литературни дела, сл.), како и документи кои изразуваат одредено мислење или опишуваат некој тековен настан (како порталите за вести). Пред да се премине на дефинирање стратегии за наоѓање на одговорите на прашањата, направена е детална анализа на произволно множество од 3500 *Jeopardy* прашања. Притоа, забележано е дека за скоро сите испитани прашања, одговорот може да се најде во документ чиј наслов претставува концепт или ентитет (поточно, 95.47% од овие прашања имаат одговор кој е наслов на документ од Википедија). Останатите прашања имаат одговор во документ чиј наслов се содржи во самото прашање, а постојат и многу мал број прашања за кои овие две набљудувања не важат. Пребарувањето во неструктурираната и структурираната содржина се извршува на следниов начин:

- Во согласност со направените анализи на прашањата, *Watson* имплементира два постоечки системи за прибирање документи и пасуси од неструктурирани извори (кои го дефинираат **примарното пребарување**). Поточно, *Watson* ги користи системите *Indri*[26] и *Lucene*[27], кои вклучуваат различни модели за рангирање. Имено, *Indri* го комбинира моделот на јазици со мрежа за заклучување [109], додека *Lucene* го користи $tf - idf$ моделот за задавање тежини [110]. Емпириските резултати потврдуваат дека групирањето на резултатите добиени од овие два система, дава повисоко отповикување во споредба со примената на само еден од нив. Од друга страна, рангот на документот (пасусот) подоцна се користи како карактеристика со цел да се зададе вредност на кандидатот за одговор извлечен од тој документ (пасус).
- *Watson*, исто така, идентификува релевантни содржини и од структурирани извори, за што користи два клучни пристапи. Првиот пристап (*Answer Lookup*) е насочен кон извори со знаење во кои се дефинирани семантички релации, како на пример *DBpedia*[28] и *Internet Movie Database (IMDB)*[29]. Ефективноста на овој пристап зависи од можноста да се утврди совпаѓање меѓу именуваните ентитети во прашањата и оние во структурираните извори, како и од квалитетот и квантитетот на семантичките релациите кои може да се идентификуваат во прашањата. За реализирање на оваа задача, *Watson* користи рачно дефинирани правила. Развојот на правилата побарува опсежна анализа на доменот, со цел да се идентификуваат најчестите релации кои се појавуваат во прашањата, и да се најдат соодветни структурирани извори кои ги опфаќаат тие релации. Вториот пристап (*PRISMATIC*

[26] http://www.lemurproject.org/indri.php
[27] http://lucene.apache.org
[28] http://www.dbpedia.org
[29] http://www.imdb.com

search) ја користи наменски изработената база на знаење *PRISMATIC* [111], во која се вклучени синтаксички и семантички релации извлечени од огромна колекција текстуални документи.

За содржината прибрана од структурираните извори, самите прибрани резултати (кои се неинстанцирани аргументи од прашалникот) претставуваат кандидати за одговорот. За генерирање на кандидати за одговорот од резултатите добиени со пребарување на неструктурираните извори, *Watson* ги користи метаподатоците на документите, како: насловите, текстовите што се покажуваат на хиперлинковите и пренасочувањата во Википедија, за да идентификува важни концепти кои се однесуваат на секој од документите. Овие метаподатоци се применуваат во три стратегии за генерирање на кандидатите за одговор. Односно, кандидат за одговор претставува наслов на секој документ утврден со примарното пребарување, именските фрази од прибраните пасуси кои претставуваат наслов на *Wikipedia* документ и текстовите што се покажуваат на хиперлинковите од прибраните *Wikipedia* документи.

Проценката направена врз множество од 3344 прашања покажува дека различните стратегии за пребарување и генерирање на одговори имаат различно влијание врз ефективноста на системот. Сепак, комбинирани заедно имаат позитивен придонес и постигнуваат бинарно отповикување од 87.17% (процент на прашања за кои е генериран точниот одговор како кандидат за одговор).

Прибирање докази [112]. Со цел да се оцени кандидатот за одговор, *Watson* користи нова техника која се нарекува **прибирање докази за поддршка** (*Supporting Evidence Retreival*) (нејзината имплементација е слична на *Indri* алгоритмот за прибирање пасуси [108]). За секој кандидат за одговор, оваа техника генерира прашалници за пребарување кои го вклучуваат и самиот одговор, формирајќи исказ. Потоа, се прибираат пасуси кои се тесно поврзани со тој исказ и истите се рангираат имплементирајќи комплет од четири алгоритми. Алгоритмите користат различни аспекти и релации меѓу термините во прашањето и пасусот. Комплетот ги вклучува следниве алгоритми:

- **Совпаѓање на термините** (*Passage Term Match*) – алгоритам кој задава вредност на пасусот во согласност со тоа колку термини од прашањето се совпаѓаат со оние од пасусот, независно од нивната граматичка поврзаност или редослед (за тежини на термините алгоритмот ги користи $tf - idf$ вредностите),
- **Оценувач кој ги вклучува биграмите** (*Skip-Bigram Scorer*) – алгоритам кој задава вредност врз основа на тоа колку парови термини од прашањето и пасусот се совпаѓаат. Притоа, се работи за термини кои се директно поврзани во синтаксичко-семантичките графови на прашањето и пасусот, или се индиректно поврзани преку единствен јазол. За генерирање на графовите *Watson* користи две клучни компоненти за длабоко разложување, *ESG* разложувачот (*English Slot Grammar parser*) и конструкторот на структурата прирок – неопходни реченични членови (*Predicate-argument structure builder – PAS builder*) [106].

- **Текстуално усогласување** (*Textual Alignment*) – алгоритам кој го утврдува степенот на согласност на редоследот на зборовите во пасусот со оние во прашањето, за што е потребно да се направи замена на фокусот со кандидатот за одговор [113], и
- **Оценувач на кандидатот за одговор во согласност со логичката форма** (*Logical Form Answer Candidate Scorer*) – дава вредност на пасусот (кој го содржи кандидатот за одговор) во согласност со степенот на усогласеност на синтаксичко-семантичкиот граф на содржината на прашањето, со оној на пасусот.

Секој од овие алгоритми задава различна вредност за секој пасус во кој се појавува кандидатот за одговор. Со цел да се генерира единствена вредност од поединечните алгоритми, различните вредности се соединуваат. Добиената вредност претставува карактеристика која се користи во статистички модел за рангирање на кандидатите за одговор во последната фаза од *Watson*-системот. Вообичаен начин за соединување претставува изборот на максималната вредност, сумирањето и сумата на исчезнување (*decaying summing*). Направените анализи од тимот на *Watson* потврдуваат дека најдобри резултати се добиваат доколку се користи сумирањето за **оценувачот кој ги вклучува биграмите** (*Skip-Bigram Scorer*), сумата на исчезнување за **текстуалното усогласување** и **совпаѓањето на термини** и максимална вредност за **оценувачот во согласност со логичката форма**. Проценката направена на множество од 3508 прашања покажува дека вклучувањето на овие четири алгоритми во процесот на рангирање на доказите во *Watson*-системот, ја подобруваат неговата точност од 67.1% на 70.4%.

Рангирање на кандидатите за одговор [114]. Последната фаза во архитектурата на *Watson*-системот вклучува рангирање на сите кандидати за одговорот, искористувајќи ги добиените вредности за доказите. Исто така, во оваа фаза се оценува и довербата дека кандидатот за одговор е навистина точниот одговор. Рачното справување со илјадниците добиени вредности за кандидатите за одговор, со цел да се изврши финалното рангирање, е практично невозможно. Затоа *Watson* користи техники за машинско учење, кои пак од друга страна наметнуваат одредени предизвици. Меѓу нив најзначајни се:

- Справување со кандидатите за одговор кои може да бидат еквивалентни или тесно поврзани (во овој случај, доказот генериран за одреден одговор, може да биде релевантен и за друг одговор).
- Справување со ситуации кога е достапно мало множество за обука за одредени категории прашања (тука се наметнува и потреба од утврдување на значењето на различните карактеристики во техниките за машинско учење, бидејќи искуството потврдува дека нивното значење зависи и од категоријата на прашањето).
- Утврдување на степенот на корисност на карактеристиките, во различните фази при рангирањето.

- Справување со хетерогеноста на карактеристиките генерирани од различни алгоритми, како и справување со карактеристики чии вредности недостасуваат или ретко се појавуваат во множеството за обука.
- Разрешување на небалансираноста на класите (ова значи дека системот мора да се справи со случаи кога се генерираат огромен број кандидати за одговор, но само неколку од нив навистина се точни).

Концептот на машинското учење во оваа фаза го користи множеството генерирани кандидати за одговор и вредностите со кои се оценети нивните докази. Тоа овозможува обука на модел за нивно рангирање и проценка на довербата дека одговорот е точен. Во овој случај, инстанца (*instance*) претставува парот прашање – одговор, додека вредности на карактеристиките се оценките на кандидатот за одговор. За градење на ефективен модел, *Watson* е обучен со примена на множество од приближно 25000 прашања од квизот *Jeopardy*, кои сочинуваат 5.7 милиони парови прашање – одговор (**инстанци**), при што секој од нив има 550 карактеристики. Клучните идеи кои *Watson* ги користи за справување со горенаведените предизвици се следниве:

- **Соединување на еквивалентните одговори** (како и оние кои се тесно поврзани): Оваа имплементација вклучува повеќе независни компоненти кои користат различни техники. Најзначајни се морфолошката анализа и анализата базирана на правила. Притоа, се испитуваат сите парови кандидати за одговор и се донесува бинарна одлука за нивно соединување. Доколку се соединат, кандидатот кој има повисок ранг се избира за претставник.
- **Рангирање со примена на класификатори:** Тимот на *Watson* експериментира со повеќе класификатори, како: логистичка регресија, машини со носечки вектори со линеарен и нелинеарен кернел, невронски мрежи, дрва на одлука и други. Она што треба да се потенцира е фактот дека логистичката регресија се покажува како техника со константно подобри перформанси, и затоа се применува во оваа фаза во *Watson*-системот.
- **Нормализација на карактеристиките**: Постоечките карактеристики се разликуваат во нивните вредности, заради што е неопходно да се направи стандардизација. Стандардизацијата се извршува по прашалник, односно секоја карактеристика се нормализира со одземање на средната вредност од сите карактеристики од ист тип, и делење со стандардната девијација.
- **Индикатор на вредноста која недостасува**: Недостатокот на вредност на одредена карактеристика е сосема различен сигнал од вредност еднаква на нула за таа карактеристика. Имено, вредностите кои се непознати може да обезбедат важни информации во процесот на оценување на точноста на одговорот. За таа цел, *Watson* користи дополнителен индикатор за секоја карактеристика, кој укажува дека вредноста недостасува. Експерименталните резултати покажуваат огромна добивка со вклучување на оваа стратегија во однос на конвенционалното вметнување на средната вредност од множеството за обука.

- **Специфицирање на посебен модел за секоја категорија прашања**: За одговарање на различни категории прашања, *Watson* применува различни карактеристики, што повлекува обучување на индивидуален модел за секоја категорија.
- **Селекција на карактеристиките**: Моделите кои располагаат со многу податоци за обука (следствено на тоа, во многу од карактеристиките ретко се појавуваат вредности (*sparse features*)), може да доведат до преобучување (*overfitting*), бидејќи моделот нема доволно искуство од тие инстанци. За разрешување на овој проблем *Watson* применува техника за елиминирање на одредени карактеристики, користејќи претходно дефиниран праг. Од друга страна, за моделите кои имаат малку податоци за обука а многу карактеристики, селекцијата на карактеристики *Watson* ја прави со примена на оценувач на конзистентно подмножество карактеристики (*consistency subset attribute evaluator*) [115] од *Weka* (*Waikato Environment for Knowledge Analysis*) алатките за машинско учење[30].
- **Задавање тежини на инстанците:** Односот на позитивни и негативни инстанци од кандидатите за одговор генерирани од *Watson* е 1 спрема 94. За разрешување на оваа небалансираност, тимот анализира повеќе методи. Направените експерименти потврдуваат дека најдобри резултати се добиваат со задавање на тежина 0.5 на негативните инстанци во логистичката регресија.

Примената на овие компоненти ја подобрува точноста на *Watson*-системот за 4.5%, во однос на основната верзија чија точност е 67%.

Од краткото резиме за *IBM Watson*-системот може да се заклучи дека доминантни принципи во него се: масивниот паралелизам, мноштвото експерти, сеприсутната проценка на довербата и интеграцијата на површното и длабокото знаење [104]. Со неговата методологија за брз развој и тестирање на поединечните компоненти, *Watson* созреа во суперконкурентен систем за одговарање прашања во рок од 5 години. Искуството стекнато со овој систем претставува поттик за сегашните и идните истражувачи во *QA*-областа.

2.4. Проценка на системите за одговарање прашања

Истражувањата во *QA*-областа го поттикнаа и интересот за развој на техники за проценка на овие системи. Автоматската проценка е истражувана од различни перспективи, како: користење на тест-колекции [116], тестови за разбирање на прочитаното (*reading comprehension tests*) [117], [118] и примена на автоматски системи за проценка на точноста на генерираните одговори, преку нивна споредба со одговори дадени од луѓе за истото множество прашања [119].

Најважните тест-колекции кои се достапни во моментов се генерирани од податоци и резултати од *QA*-проценките развиени на *TREC*[31]. *TREC*-конференциите се организирани од Националниот институт за стандарди и технологија (*National Institute*

[30] http://www.cs.waikato.ac.nz/ml/weka/
[31] http://trec.nist.gov

of Standards and Technology - NIST) и на нив се реализираат различни **работилници** (*tracks*) поврзани со прибирањето информации. За секоја од тие работилници се дефинира рамка за проценка, која овозможува учесниците да генерираат резултати од нивните системи на заеднички корпус. Во 1999 година, *TREC* ја претставува првата работилница за проценка на *QA*-системите. За оваа *QA*-работилница, координаторите дефинираат рамка за проценка која се состои од тест-колекција и метрики за проценка, со цел да се овозможи мерење на перформансите на системите. Следните *TREC QA*-проценки еволуираат во согласност со промените и дополнителните барања кои се воведени во наредните *QA*-работилници. Овие промени главно се однесуваат на зголемување на колекциите со документи, воведување на построги барања, со цел одговорите да се сметаат за точни, зголемување на множеството прашања, како и зголемување на нивната комплексност. Особено влијание има воведувањето на различни категории прашања во овие проценки (фактовидни, прашања со набројување, дефиниции, итн.), што придонесе за развој на специфични техники за проценка во согласност со специфичните карактеристики на секоја категорија прашања. *TREC*-работилниците континуирано се одржуваа сè до 2007 година [120], по што следи повеќегодишна пауза. Сепак, интересот за оваа област поттикна други форуми за проценка (како *Text Analysis Conference - TAC*[32]). Уште повеќе, самото враќање на *QA*-работилницата на *TREC*-конференцијата во 2015 година (наречена *LiveQA*), ја потврди значајноста на QA-системите во насока на задоволување на потребите на моменталните и идни корисници.

2.4.1. Метрики за проценка

Метриките за проценка кои се користат за мерење на перформансите на *QA*-системите континуирано се развиваат низ годините. Причината за тоа е задоволување на потребите кои произлегуваат од проширување на проценката со вклучување на нови категории прашања и ограничувања кои се однесуваат на одговорите. Во овој дел се наведени главните метрики за проценка кои се користат на *TREC QA*-работилниците. Во суштина, методологијата и метриките за проценка произлезени од *TREC*, станаа стандард во ова поле и се прифатени како референтни од страна на другите форуми за проценка. Меѓу нив, најзначајни се *NII Test Collections for IR Systems* (*NTCIR*[33]) и *Cross-Language Evaluation Forum* (*CLEF*[34]), кои се насочени кон проценка на системите за одговарање прашања поставени на популарните источно-азиски и европски јазици, соодветно. Исто така, овие форуми се фокусирани и на прибирање информации каде прашалниците се поставени на еден природен јазик, додека колекцијата содржи документи напишани на еден или повеќе други јазици. Овој пристап се нарекува меѓујазично прибирање информации (*cross-language information retrieval*).

[32] http://www.nist.gov/tac/
[33] http://research.nii.ac.jp/ntcir
[34] www.clef-campaign.org

Првата мерка за проценка која се користи на *TREC* е **среден реципрочен ранг** (*mean reciprocal rank – MRR*). Имено, кога на системот му е дозволено да прикаже неколку рангирани одговори за одредено прашање (на првите *TREC*-работилници дозволено е прибирање на пасуси со одредена големина), тогаш тоа прашање добива резултат еднаков на реципрочната вредност од позицијата на првиот пасус кој го содржи точниот одговор. Доколку ниту еден од пасусите не го содржи одговорот, резултатот кој му се придружува е нула. *MRR* се пресметува како средна вредност од резултатите на сите прашања од колекцијата. Оваа мерка стана стандардна мерка за проценка на *QA*-системите. Подоцна, кога на следните работилници е дозволено прикажување на само еден одговор по прашање, се користи мерката **резултат зависен од довербата** (*confidence-weighted score*). За дадена листа од Q прашања рангирани во согласност со довербата зададена од системот дека е најден точниот одговор, **резултатот зависен од довербата** се дефинира на следниов начин:

$$\frac{1}{Q}\sum_{i=1}^{Q}\frac{n_i}{i}, \tag{1}$$

каде n_i е бројот на точни одговори во првите i рангирања.

Вклучувањето на нови категории прашања на *TREC QA*-работилницата, како прашањата со набројување и прашањата означени како „други“ (дефиници, објаснувања, опис и слично), побара вклучување на посебни проценки со различни метрики за секоја категорија прашања. Системот добива конечен резултат со комбинирање на резултатите добиени за секоја категорија. Фактовидните прашања се проценуваат користејќи ја **точноста** (*accuracy*) како метрика, односно користејќи го процентот на прашања кои системот точно ги одговорил. Прашањата со набројување се проценуваат со добро познатата мерка во *IR*-областа, **F-резултатот** (*F-score*), кој ги комбинира **прецизноста** P и **отповикувањето** R:

$$F = \frac{2 \times P \times R}{P + R}. \tag{2}$$

Најкомплексни за проценка се прашањата означени како „други“, за што е неопходно определување на множество сегменти од информации (S), кои треба да се појават во одговорот. Оценувачите ги класифицираат тие сегменти од информации како „витални“, доколку мора да се појават во одговорот, и „невитални“, доколку нивното појавување во одговорот е прифатливо (но, незадолжително). Конечниот резултат е комбинација од прецизноста и отповикувањето на множеството сегменти од информации, при што трипати поголема важност се задава на отповикувањето:

$$F(\beta = 3) = \frac{10 \times P_S \times R_S}{9 \times P_S + R_S}, \tag{3}$$

каде P_S и R_S се прецизноста и отповиукањето на множеството сегменти од информации S, соодветно. Овој метод во голема мера зависи од способноста на оценувачите да ги определат сите сегменти од информации кои го дефинираат одговорот. Сепак, истиот

дава квантитативна мерка која може да се користи за споредба на различни *QA*-системи.

2.4.2. Проценка на *TREC LiveQA*-работилниците

Во 2015-тата година за првпат се организира *LiveQA*-работилницата на *TREC*, која се фокусира на одговарањето прашања во реално време, и тоа прашања кои се поставени од реални корисници [76]. Во суштина, станува збор за неодговорени прашања поставени на веб-локацијата за одговарање прашања на *Yahoo* (*Yahoo Answers site – YA*). За разлика од претходните *QA*-работилници, овие прашања се значително покомплексни и вклучуваат различни категории (како: прашања кои побаруваат мислење, совет, прашања за анкетирање, и други), со што самата задача за одговарање прашања е мошне пореална и исклучително предизвикувачка. Прашањата припаѓаат во неколку области, познати за учесниците (уметност и хуманост, убавина и стил, компјутери и Интернет, здравје, дом и градинарство, домашни миленичиња, спорт и патување). Следува пример на прашање од множеството за тестирање на *TREC 2015*:

Пример: **Број на прашање: 572**
Прашање:
Колкаво е времето на патување од Лос Анџелес до Њујорк и од Њујорк до Лос Анџелес?
Доколку патувате од Лос Анџелес до Њујорк и од Њујорк до Лос Анџелес и двата авиона патуваат со иста брзина и по истата рута, дали времето на патување ќе биде исто? Или, дали побрзо би било патувањето во насока на ротација на Земјата?

Одговорот кој за нив треба да се генерира е ограничен на 1000 карактери, додека времето за одговарање е ограничено на една минута, со цел да се оневозможи рачно одговарање на истите или користење на одговорите од корисниците кои симултано се акумулираат на *YA* веб-локацијата. Уште повеќе, системот може да одлучи да одговара само на дел од прашањата, враќајќи „нема одговор“ (*null response*) за прашањата за кои ќе одлучи да не ги одговара. Со цел да се земат предвид наведените промени кои се однесуваат на прашањата и форматот на одговорите, за проценка на системите учесници дефинирани се нови метрики. Тие ја земаат предвид и вкупната успешност на системот (кога прашањата кои системот не ги одговорил се третираат исто како и погрешно одговорените прашања), како и прецизноста на системот (кога се наградува одлуката системот да не одговара одредено прашање).

Множеството за тестирање доставено до учесниците содржи 1087, односно 1015 прашања за *TREC Live QA 2015* и *TREC Live QA 2016*, соодветно. Одговорите генерирани од системите учесници ги оценуваат *NIST*-оценувачи, кои ги задаваат следниве оценки: **0** значи дека прашањето не е одговарано, **1** укажува на слаб одговор, **2** укажува на задоволителен, **3** укажува на добар и **4** означува одличен одговор. Притоа, за проценка на системите се користат 7 мерки, и тоа:

- ***avgScore(0-3)***: Претставува главна мерка за оценка на системите и означува просечна оценка добиена од сите прашања. Оваа мерка ги трансформира

оценките од 1 до 4 во 0 до 3, соодветно, што значи дека оценката 1 ја третира исто како и оценката 0 за неодговараните прашања.

- ***succ@ i +***: Претставува број на прашања со оценка i или повеќе ($i \in \{\overline{2,4}\}$), поделен со вкупниот број на прашања. На пример, ***succ@ 2 +*** го мери процентот на прашања одговорени од системот кои имаат барем задоволителна оценка.
- ***prec@ i +***: Претставува број на прашања со оценка i или повеќе ($i \in \{\overline{2,4}\}$), поделен со бројот на прашања одговорени од системот. Со оваа мерка се оценува прецизноста на системот, а притоа не се казнува системот заради неодговорените прашања.

Проценката на системите во 2015 година потврдува дека *OAQA*-системот предложен од *Carnegie Mellon University* [68] е водечки систем во однос на сите мерки [76]. Вредностите од 1.081 и 0.542 за мерките ***avgScore*** и ***prec@2+***, соодветно, потврдуваат дека автоматското одговарање прашања од овој систем е оценето како задоволително во просек и дека околу половина од одговорите се оценети барем како задоволителни. Ваквиот резултат е далеку од максималната можна вредност за ***avgScore*** од 3.0, но ова не е изненадувачки со оглед на комплексноста на работилницата и заради фактот што за првпат се реализира одговарање прашања во реално време. Исто така, оваа проценка потврдува дека системите најтешко ги одговараат прашањата од области кои процентуално се најмалку застапени во множеството прашања дадено до учесниците.

Задачата за одговарање прашања на *TREC LiveQA 2016* е идентична како и претходната година. И оваа година како најдобар се покажува *OAQA*-системот од вкупно 26 проценети системи [121]. Севкупниот пристап кој го користат истражувачите од *Carnegie Mellon University* е мошне сличен како и пристапот во 2015-тата година (секција 2.3.1.2), со таа разлика што системот е проширен со нов метод за рангирање на одговорите, кој се заснова на рекурентните невронски мрежи со внимателно кодирање и декодирање (*attentional encoder-decoder recurrent neural networks*) [122].

Резиме:

Системите за одговарање прашања им овозможуваат на корисниците да пристапат до одредено знаење на природен начин, поточно преку поставување прашања на природен јазик. Главните предизвици со кои се соочуваат *QA*-системите се: разбирање на прашањата поставени на природен јазик (независно од начинот на нивна репрезентација), разбирање на знаењето дадено во огромното количество достапни документи (структурирани, полуструктурирани или неструктурирани) и наоѓање на точни, концизни одговори кои ќе ги задоволат потребите на корисниците. Со цел да се разрешат овие предизвици, низ годините развиени се различни пристапи вклучени во процесот на креирање „идеален“ *QA*-систем. Направените анализи потврдуваат дека изборот на конкретна техника во овој процес е исклучително зависен од проблемот кој се разработува. Често, хибридниот пристап, кој претставува разумна мешавина на очигледно различни техники, обезбедува подобри резултати во форма на брзина, зголемена релевантност и повисоки вредности на мерките за прецизност и отповикување. Сепак, јасно е дека техниките за одговарање прашања кои ги комбинираат лингвистичкиот пристап и пристапите базирани на правила и учење од податоци, и во иднина ќе бидат во фокусот на голем број *QA*-истражувачи.

3. Комбинирање на техниките за успешно одговарање прашања

„Математиката е јазикот на универзумот.“

Galileo Galilei

Природниот јазик претставува алатка со која човекот може да се изрази себе си и да се разбере со околината. Тој го учи јазикот преку откривање одредени шаблони, како: спецификите на јазикот, начинот на поврзување на зборовите со цел да изгради одреден исказ, итн. Додека зборовите претставуваат градежни елементи на значењето, нивната релација во структурата на реченицата (документот), го дава вистинското значење на текстот [123]. **Обработката на природните јазици** претпоставува дека доколку овие шаблони можат да се дефинираат и да му се опишат на системот, тогаш тој ќе може да научи како луѓето зборуваат и се разбираат меѓу себе. Имено, човекот го извлекува значењето од текстот или говорниот јазик на повеќе нивоа, и тоа:

- фонетско (ниво кое е клучно кај системите за препознавање на говорот),
- морфолошко,
- синтаксичко,
- семантичко,
- дискурс (*discourse*), и
- прагматично ниво (секција 2.2.3).

Сепак, не секој *NLP*-систем ги имплементира сите нивоа. Бидејќи секое од овие нивоа за разбирање на јазикот следи одредени дефинирани шаблони, користејќи ги нивните дефиниции може да се вметне одредено разбирање на јазикот во самиот систем. Колку е повисоко нивото на сфаќање на текстот и говорот, толку оваа задача станува покомплексна. Еден од најзначајните предизвици на системите кои имплементираат *NLP*-техники е потребата од апсолутно прецизна и недвосмислена репрезентација на содржината на реченицте (документите). Оваа формална репрезентација треба да ги содржи зборовите и нивното значење во реченицата, како и структурата на самата реченица, со цел да му се овозможи на системот подобро да ја разбере.

Обработката на природните јазици може да се вметне во секој од модулите на еден систем за одговарање прашања, преку примена на некои или на сите наведени нивоа на разбирање. Комплетната примена на *NLP* значи интерпретирање и зачувување на значењето и на прашањата и на документите, во сите нивоа во *QA*-процесот. Изборот колку и каде да се вметне *NLP* е предмет на практична анализа. Односно, она што треба да се одговори е до кој степен дополнителната обработка го забавува целокупниот *QA*-процес и дали добиените резултати се толку многу подобрени, заради што вреди да се прифати намалувањето на ефикасноста.

Проблеми при обработка на природните јазици. Секој природен јазик има свои специфичности кои силно влијаат врз успешноста на *IR* и *QA*-системите. Тука, пред сè, се вбројуваат **лингвистичката варијација** и **лингвистичката**

двосмисленост [124]. Под **лингвистичка варијација** се подразбира можноста да се користат различни зборови и изрази, со цел да се искаже една иста мисла, додека **лингвистичката двосмисленост** се јавува кога одреден збор или фраза дава можност да се интерпретира на повеќе начини. Двата феномена силно влијаат врз процесот на разбирање на прашањата и документите каде се пребаруваат нивните одговори, и тоа на сосема два различни начина. Лингвистичката варијација доведува до затајување на документи, што значи системот пропушта документи кои се релевантни на прашањето, бидејќи зборовите од прашањето не се сретнуваат во него (како синонимите, т.е. зборовите со слично или идентично значење). Од друга страна, двосмисленоста имплицира прибирање на несуштински документи заради погрешната интерпретација на зборовите кои се јавуваат во прашањето, а имаат различно значење во документот (како хомонимите, т.е. зборовите со иста форма, а различно значење). Токму ваквите карактеристики го прават процесот на автоматска обработка на јазикот исклучително напорен.

Друг клучен проблем со кој се соочува *NLP*-областа е автоматското утврдување на припадноста на зборот во различни зборовни групи. **Зборовните групи** или **класи** групираат зборови врз основа на некој општ, заеднички признак. Во науката на јазикот постојат повеќе класификации на зборовите во групи, а најчесто се применуваат: **значенската** (семантичка), **формалната** (морфолошка) и **функционалната** (синтаксичка класификација) [123].

- **Значенската класификација** ја зема предвид внатрешната, смисловната, значенската страна на зборовите. Зборовите се класифицираат врз основа на нивните општи лексички и граматички значења. Тоа се најопшти поимски или категоријални значења, при кои се занемарува посебното, сопственото значење на зборот. Според овој пристап, во македонскиот јазик се изделуваат единаесет зборовни групи (класи): именки, придавки, броеви, заменки, глаголи, прилози, предлози, сврзници, честици, извици и модални зборови.
- **Морфолошката класификација** ја зема предвид менливоста/неменливоста на зборовите. Врз основа на тоа дали ја менуваат или не ја менуваат својата форма во исказот, сите зборови се делат на две големи групи: **менливи** и **неменливи**. Менливите зборови се појавуваат во различни форми во исказите, додека неменливите секогаш се појавуваат во една иста форма. Во македонскиот јазик, во менливи зборови спаѓаат: именките, придавките, броевите, заменките и глаголите. Неменливи зборови се: прилозите, предлозите, сврзниците, честиците, извиците и модалните зборови.
- Кога се зема предвид функцијата (службата) на зборовите во реченицата, се зборува за **синтаксичка класификација** на зборовите. Во реченицата некои зборови претставуваат самостојни членови (како: подмет, прирок, предмет, др.), додека останатите се во служба на тие самостојни делови (т.е., ги поврзуваат, помагаат во изразување на различни односи меѓу нив, итн.).

Припадноста на еден збор во некоја зборовна група најуспешно се определува ако се имаат во предвид сите три пристапи и ако зборот не се разгледува изолирано,

туку во рамките на реченицата. Денес, сè уште не може да се каже дека постојат *NLP*-техники кои без грешка можат да го извлечат точното значење на реченицата (документот). Во суштина, научната заедница е поделена околу процедурата која треба да се следи за да се постигне оваа цел.

Во следниов дел се наведени клучните *NLP*-техники кои се имплементираат во *IR* и *QA*-системите, со цел делумно да се разрешат наведените проблеми и подобро да се разбере природниот јазик. Исто така, анализирано е и влијанието на близината меѓу зборовите од прашалникот (прашањето) и зборовите од понудените (генерираните) одговори за дадено прашање, во процесот на утврдување на точниот одговор. Секоја секција во продолжението на ова поглавје е проследена со детален приказ на соодветниот пристап имплементиран во системот за одговарање прашања на македонски јазик.

3.1. Стоп зборови

Еден од клучните пристапи кои се применува во претпроцесирачката фаза кај *IR* и *QA*-системите е дефинирање на листата од таканаречените **стоп зборови** (како службените зборови, зборовите кои имаат висока фреквенција, и слично), односно зборови кои немаат никаква или имаат сосема мала важност во процесот на прибирање одредена информација. Отфрлувањето на стоп зборовите од документите (прашалниците) обично ја подобрува успешноста на системот. Но, сепак постојат и примери со кои системот не може да се справи по нивното отстранување, како на пример со исказот: „Да се биде или не“. Коригирањето на листата со стоп зборови за одредена задача може значително да ги подобри резултатите [125]. И покрај тоа што дефинирањето на стоп зборовите генерално не се смета за *NLP*-задача, сепак *NLP*-техниките можат да помогнат во креирањето на специфични листи кои можат да се справат со примери слични на претходно наведениот исказ.

3.2. Означување на зборовите во согласност со значенската класификација

Важноста да се идентификува значенската група на зборот (во понатамошниот текст, **зборовна група** (*part-of-speech*)) при обработка на јазикот лежи во огромната количина информации кои таа ги дава за самиот збор, но и за неговите соседи. Ова е секако јасно за главните групи зборови (како: именките, глаголите, придавките, итн.), но исто така важи и за многу пофини разграничувања. На пример, познавањето дали еден збор е присвојна заменка или лична заменка може да долови кои зборови веројатно ќе се појават во неговата околина (пример, по присвојните заменки најчесто следи именка, додека по личната заменка следи глагол). Располагањето со ваква информација е клучно при дефинирањето на статистичкиот модел на јазик за препознавање на говор. Во одредени природни јазици, како англискиот, зборовната група исто така го одредува и изговорот на одредени зборови [126]. Познавањето на оваа информација овозможува поприроден изговор на системите за синтеза на говор и повисока точност на системите за препознавање на говор.

Познавањето на зборовната група во која припаѓаат зборовите може да се искористи и во процесот на нивно сведување на основен збор, познат како **стемирање** (*stemming*), бидејќи зборовната група ги открива морфолошките афикси (префикси и суфикси) на зборовите (секција 3.3). Достапноста на информацијата за зборовната група на зборовите може да ја подобри успешноста на *IR* (*QA*) системите преку примена само на одредени зборовни групи (секција 3.2.1). Автоматското доделување на зборовната група игра значајна улога и при синтаксичкото разложување (*parsing*), во задачата за селектирање на точното значење на зборот (*word sense disambiguation*) и при површното разложување на текстот, со цел брзо да се најдат имињата, да се определи времето, местото или другите именувани ентитети во задачата за извлекување информации (*IE*). На крај треба да се потенцира дека, корпусите документи чии зборови се означени со зборовна група, се исклучително корисни за лингвистичко истражување (на пример, за наоѓање на инстанци или фреквенции на одредени конструкции).

Во литературата се предложени многу алгоритми за означување на зборовите со зборовна група, кои генерално се класифицираат во три групи: означувачи базирани на правила (*rule-based taggers*), стохастички означувачи (*stochastic taggers*) и хибридни означувачи.

- **Означувачите базирани на правила** најчесто вклучуваат две фази. Во првата фаза користат речник за создавање на листа од можните зборовни групи за секој збор. Втората фаза вклучува огромна база од рачно напишани правила за да се утврди точната зборовна група од генерираната листа. На пример, во англискиот јазик за двосмислен збор се потврдува дека е именка (а не глагол), доколку му претходи членот (*determiner*). Еден од првите означувачи е *Taggit* [127], кој за означување на зборовите применува речник и 3300 рачно дефинирани хеуристички правила (правилата го земаат предвид и контекстот во кој се појавува зборот). Предноста на означувачите базирани на правила е лесната разбирливост на правилата дефинирани рачно и можноста да постигнат добри резултати и со мал број правила. Некои од овие означувачи постигнуваат точност дури и нешто повеќе од 99%. Од друга страна, недостатоците кои се врзуваат со нив се следниве: правилата се специфични за јазикот и корпусот (а нивното дефинирање побарува одлично познавање на природниот јазик) и програмирањето на навистина добар означувач одзема многу време.
- **Стохастичките означувачи** најчесто ја разрешуваат двосмисленоста користејќи корпус за обука со цел да се пресмета веројатноста даден збор да има одредена ознака во даден контекст. Најголем дел од овие означувачи се базираат на скриените Маркови модели (*Hidden Markov Model – HMM*), а за обука потребен им е огромен рачно анотиран корпус, со цел да ја определат веројатноста за секој збор. Постигнуваат точност од 95-97%, користејќи означени множества кои вклучуваат неколку десетици до 130 ознаки. повторна Со повторна обука на моделот, стохастичките означувачи лесно можат да се прилагодат на други јазици и нов корпус. Единствениот недостаток е можноста тешко да се програмираат.

- Еден од најпознатите **хибридни означувачи** е *CLAWS4* [128]. Во првиот чекор, овој означувач применува осум теста за задавање на иницијални ознаки, по кој секој збор има една или повеќе ознаки. Во вториот чекор се применува *HMM* со цел да се утврди единствена ознака. Неговата точност се проценува на 96-97%.

Множества со ознаки за англискиот јазик. Бројот на популарни множества со ознаки за англискиот јазик е мал. Најголем дел од нив се развиени користејќи го означениот корпус со 87 ознаки од Брауновиот Универзитет (*Brown University*) [129], [130]. Овој корпус претставува примерок од еден милион зборови извлечени од текстови со различен жанр (весници, новели, научно-истражувачка литература, итн.). Означувањето на зборовите кои се појавуваат во него со зборовни групи, е направено во два чекора. Првиот чекор ја користи *Taggit* програмата за означување, која успева точно да означи 77% од зборовите од Брауновиот корпус. Комплетното означување е завршено рачно со корекција на останатите грешки. Освен Брауновото множество со ознаки (*Brown tagset*), едно од најпознатите и најчесто користени множества со ознаки е *Penn Treebank* множеството со 45 ознаки [131]. Множествата со ознаки се применуваат за проценка на алгоритмите за означување. Притоа, нивната корисност зависи од тоа колку информации и се потребни на апликацијата во која се применуваат.

Колку е тежок проблемот за задавање ознаки? Постојат ситуации кога носењето на одлука за означување на зборовите во одредена реченица е тешка и за човекот. Од друга страна, дури и едноставни примери може да ја направат оваа задача нетривијална за автоматско означување. На пример, во македонскиот јазик постојат зборови кои не се еднозначни и може да припаѓаат во неколку зборовни групи (како зборот „игра“, кој може да биде именка или глагол, или зборот „право“, кој може да биде именка, придавка или прилог). Целта на означувачот на зборовните групи е разрешување на оваа двосмисленост и избирање на вистинската ознака за дадениот контекст. Но, она што се наметнува како прашање е колку често се појавува двосмисленоста при означувањето, за одреден природен јазик. Анализите потврдуваат дека најголем дел од зборовите во англискиот јазик се недвосмислени, односно имаат единствена ознака. Сепак, многу од најфреквентните англиски зборови се двосмислени (на пример, зборот *„can“* може да биде помошен глагол (може), именка (метална кутија) или глагол (става нешто во металната кутија)). Така, *DeRose* [132] утврдува дека 11.5% од вкупниот број **типови** [35] (*word types*) во Брауновиот корпус се двосмислени, а тоа важи за повеќе од 40% од **токените**[36] (*tokens*).

Што се однесува до македонскиот јазик, добиените сознанија ги потврдија претпоставките дека сè уште не е направена ниту една анализа која би го утврдила процентот на леми [37] во македонскиот јазик што имаат повеќе од една ознака за

[35] **Тип** е класа од сите **токени** кои ја содржат истата низа од карактери.
[36] **Токен** е примерок од низа карактери во одреден документ кои се групирани како корисна семантичка единица за процесирање [133].
[37] **Лексема** претставува множеството од флективни форми на одреден збор (кој може да има повеќе семантички значења), додека **лема** е претставникот на тоа множество во речникот на одреден јазик.

зборовна група. Ваква статистичка анализа е направена само за ограничен корпус документи, за кој се потврдува дека повеќе од 30% од зборовите се двосмислени [134].

3.2.1. Влијание на различните зборовни групи во процесот на прибирање информации

Перцепцијата дека познавањето на зборовната група на зборовите може до одреден степен да укаже на присуство или отсуство на информативна содржина во одреден јазик, не е воопшто нова. Истата се забележува уште во 4-тиот век пред нашата ера, кога во старогрчкиот јазик е направена груба категоризација на зборовите само на три категории: именки, глаголи и придавки (вклучувајќи ги и партиципите (*particilpes*)), и сè останато [135]. Понова формулација за разграничување на зборовните групи претставува „Теоријата за рангирање“ на *Jepspersen* [136], кој забележува дека граматичките категории се семантички одредливи и можат да бидат предмет на рангирање. Тој ги идентификува следниве степени на зборовни групи:

- прв степен (примарни) зборовни групи → именки,
- втор степен (секундарни) зборовни групи → глаголи и придавки (вклучувајќи ги и партиципите),
- трет степен (терцијални) зборовни групи → прилози, и
- сè останато.

Jepspersen го дефинира поимот **степен** на зборовните групи земајќи ги предвид нивните комбинаторни својства. Односно, дека секоја зборовна група е подложна на промени од зборовната група со повисок степен (на пример, именките се модифицирани од глаголите, а глаголите од прилозите). Денес, широко прифатено разграничување на зборовните групи е нивната поделба на две категории: **отворени** (**полнозначи**) и **затворени** (**функциски** или **службени зборови**). Во англискиот јазик категоријата отворени ги вклучува: именките, глаголите, придавките, прилозите, додека сите останати влегуваат во категоријата затворени зборовни групи (заменки, членови, сврзници, предлози, извици, помошни глаголи, броеви, партиципи).

3.2.1.1. Релевантни истражувања

NLP-истражувачите често ги сметаат зборовите од затворената категорија како стоп зборови и ги исклучуваат од севкупниот процес, заради нивниот незначителен придонес кон содржината на текстот кој се обработува. Но, и покрај вклучувањето само на категоријата отворени зборовни групи, се наметнува прашањето дали некоја од нив е позначајна од другите и до кој степен.

– Одредени истражувања потврдуваат дека именките се најдобри индикатори за содржината на документите [137]. Во области како биологијата и рекламирањето, кои ги потенцираат разликите меѓу „нештата“ и нивните својства, како исклучително значајни се покажуваат придавките. Од друга страна, во области како музиката, кои се главно богати со прилози, улогата на оваа зборовна група се потврдува како одлучувачка во толкувањето на текстот [138]. *Klavans et al.* [139] спроведуваат студија со која потврдуваат дека појавувањето и распределбата на глаголите во статии за вести, претставуваат значајни индикатори за типот и

содржината на статиите. Во одредени *IR*-задачи дури и стоп зборовите се покажуваат како корисни [140].

- Иако претходно наведениве истражувања потрдуваат дека важноста на зборовните групи во голема мера зависи од доменот кој се разгледува, *Shah et al.* [138] се обидуваат да ја проценат нивната улога во општ случај. Во задачата за класификација на документите заклучуваат дека највисока точност се добива доколку се користат само именките од документите како индикатори на нивната содржина, потоа само глаголите, само придавките и најмала точност се добива при користење само на прилозите. Точноста уште повеќе се зголемува доколку во класификацијата се користат четирите наведени зборовни групи.
- Забележувајќи ја важноста на значенската ознака на зборовите во пребарувањето, *Xu* [141] предлага инкорпорирање на таа информација во добро познатиот $tf - idf$ алгоритам ($tf - idf$ тежини за термините и примена на мерка за сличност со цел да се споредат векторите на документот и прашалникот). Овој пристап ја зема предвид и зборовната група на секој термин од прашалникот и во согласност со тоа, задава специфична тежина на терминот. Врз основа на фактот дека глаголите и именките се обично поважни од придавките и прилозите, а тие пак се поважни од останатите зборовни групи (заклучоци изведени за англискиот јазик), авторот предлага инкорпорирање на **тежинска функција** во $tf - idf$ тежината на секој термин. Функцијата е следна:

$$w_{POS}(t) = \begin{cases} w_1, & \text{ако } t \text{ е глагол или именка} \\ w_2, & \text{ако } t \text{ е прилог или придавка,} \\ w_3, & \text{инаку} \end{cases}$$

  каде $w_1 > w_2 > w_3 > 0$. Емпириските резултати потврдуваат дека модифицираниот алгоритам вклучен во предложениот систем *Courses* од самиот автор (систем за пребарување на онлајн отворени курсеви) постигнува исклучително позитивни резултати.
- *Lioma et al.* [142] предлагаат нова шема за задавање тежини на термините, кои произлегуваат од статистиките за n - грамите од зборовните групи. Ваквиот начин на задавање тежини укажува колку терминот е генерално информативен, врз основа на n - грамите од зборовни групи во кои тој термин најчесто се појавува, за одреден природен јазик. Експерименталните резултати потврдуваат дека интегрирањето на ваквиот тип тежини во пребарувањето, дава подобрување од +33.7%, во однос на стандардните $tf - idf$ и $BM25$ тежини.
- *Wang et al.* [143] анализираат до кој степен инкорпорирањето на ознаката за зборовната група може да го подобри пребарувањето на клиничките информации. За утврдување на оптималните тежини применуваат алгоритам за машинско учење. Со цел да потврдат дека методот кој ја вклучува и ознаката за зборовната група (со задавање на оптималната тежина) е валиден за секој модел за пребарување, ги проценуваат перформансите на повеќе модели (меѓу кои: $tf - idf$, $Okapi\ BM25$, моделот на јазици со порамнувањето на *Dirichlet* и порамнувањето на *Jelinek-Mercer*) [144]. Методот е тестиран на работилниците со медицински записи на *TREC 2011* и *TREC 2012*. За секој модел за пребарување, експерименталните

резултати потврдуваат дека пребарувањето кое ги вклучува и ознаките за зборовната група со оптимални тежини дава значително подобри резултати, отколку пребарувањето во кое се зададени еднакви тежини на зборовните групи или таа информација не се користи воопшто.

3.2.2. Означување на зборовната група во други јазици

Алгоритмите за означување на зборовната група дефинирани за англискиот јазик, можат да се применат и врз други природни јазици. Во одредени случаи, овие методи функционираат добро без поголеми модификации. Таков експеримент реализира *Brants* [145], кој го применува добро познатиот *TnT* (*Trigrams'n'Tags*) статистички означувач на зборовни групи (во кои се користат Маркови модели од втор ред). Во своето истражување авторот потврдува дека овој означувач постигнува точност за означување на германскиот корпус *NEGRA* од 96.7%, која е идентична со точноста постигната за означување на англискиот корпус *Penn Treebank*. Но, кога се работи за **аглутинативните јазици** [38] (како: унгарскиот, финскиот, естонскиот, турскиот, и други) и **флективните (фузиони) јазици**[39] (како: македонскиот, грчкиот, полскиот, чешкиот, и други), неопходно е да се направат бројни промени и проширувања.

Главниот проблем кај овие јазици е големиот број зборови со кои располагаат (вклучувајќи ги сите нивни можни зборoформи, кои определуваат лице, род, број, начин, време, итн.), во споредба со англискиот јазик. Така, *Oravecz et al*. [146] забележуваат дека во англиски корпус кој содржи половина милион зборови, има околу 19000 различни **типови**, додека во унгарски корпус со иста големина има скоро 50000 различни **типови**. Проблемот станува уште позабележителен со зголемување на корпусот. Ваквиот голем број **типови** е причина за значителна деградација на перформансите на *TnT* означувачот, доколку истиот се примени врз аглутинативните јазици. Така, *Oravecz et al*. [146] забележуваат дека неговата точност за означување на унгарски корпус изнесува 92.88%, во споредба со точноста од 96.7%, добиена за означување на корпусите на англиски и германски јазик. Уште повеќе, авторите посочуваат дека точноста добиена за познатите зборови (98.32%) е споредлива со англиските резултати. Но, исклучителен проблем претставуваат непознатите зборови (*unknown words*) за кои се постигнува точност од само 67.07% за унгарскиот јазик, во споредба со 84-85% за непознатите зборови од англиските јазик. *Hajic* [147] го забележува истиот проблем кај многу други јазици (вклучувајќи ги и чешкиот, словенскиот, естонскиот и романскиот). Точноста на статистичките означувачи може да се подобри со вклучување на речник кој дава можност за подобро моделирање на непознатите зборови. Наведениве анализи потврдуваат дека главен проблем кај аглутинативните и флективните јазици е означувањето токму на непознатите зборови.

---

[38] **Аглутинативни јазици** се оние јазици чија морфологија главно користи **аглутинација** (лингвистички процес што се однесува на изведбената морфологија, во која комплексните зборови се формираат со спојување на различни морфеми, кои остануваат непроменети во секој аспект по спојувањето).

[39] **Флективен јазик** е оној јазик во кој се менува формата на одредени зборови при промена на начинот на кој тие се користат во речениците.

Втората потешкотија поврзана со овие јазици е огромното количество информации кое го носи морфологијата на зборот. Во англискиот јазик, многу од информациите за синтаксичката функција на зборот се содржани во редоследот на зборовите, или соседните функциски зборови. Во јазиците богати со формообразувачки морфеми, информациите како: род (машки, женски, среден), лице (прво, второ, трето), број (еднина, множина), итн., се содржат во самите зборови, а при тоа редоследот на зборовите игра помала улога во означување на синтаксичката функција. Бидејќи означувањето често се користи како претпроцесирачки чекор за други *NLP*-алгоритми, (секција 3.2), исклучително е важно да се утврди оваа морфолошка информација. Тоа значи дека за овие јазици се наметнува потребата од вклучување на морфолошките информации во ознаката за зборовна група, со цел истата да биде корисна како и ознаката за зборовна група на зборовите од англискиот јазик, која заради специфичноста на јазикот не ги содржи.

Од овие причини, множествата со ознаки за аглутинативните и флективните јазици обично се многу поголеми од множествата со ознаки за англискиот јазик, кои најчесто вклучуваат 50 до 100 ознаки. Ознаките во таквите богати множества претставуваат низи од морфолошки ознаки, а не само единствена ознака. Доделувањето ознаки на зборовите од такво множество ознаки значи заедничко разрешување на проблемот за означување на зборовната група и морфолошката интерпретација (*disambiguation*). Токму примената на мофролошката информација во ознаката за зборовна група, значително го зголемува нивниот број. Тоа може јасно да се забележи во морфолошки означениот корпус *MULTEXT-East* [40] на англиски, македонски, чешки, естонски, унгарски, романски и словенски јазик [148]. Во табелата 4 е дадена големината на множествата ознаки за овие корпуси, забележана од *Hajic* [147].

Јазик	Големина на множеството ознаки
англиски	139
македонски[41]	765
чешки	970
естонски	476
унгарски	401
романски	486
словенски	1033

Табела 4. Големина на множествата ознаки за повеќе јазици

За да се применат вакви големи множества ознаки, неопходно е да се изврши морфолошка анализа на зборот, со цел да се изгенерира листа од можните морфолошки ознаки (можни ознаки за зборовната група). Морфолошката анализа најчесто се прави со вклучување на речник за соодветниот јазик, кој ги содржи сите можни форми за секој збор со соодветните ознаки. Потоа, целта на означувачот е да направи разлика помеѓу овие ознаки.

[40] http://nl.ijs.si/ME/
[41] http://nl.ijs.si/ME/V4/msd/html/msd.msds-mk.html

3.2.2.1. Релевантни истражувања за персискиот, арапскиот и кинескиот јазик

Во оваа секција се наведени неколку релевантни истражувања кои ја вклучуваат информацијата за зборовната група на зборовите, при прибирањето информации на персиски, арапски и кинески јазик.

- За означување на зборовите во перскискиот јазик, *Karimpour et al.* [149] го применуваат *TnT* статистичкиот означувач, обучен на рачно означена колекција со повеќе од два милиона збора и 40 различни ознаки. Во своето истражување за прибирање информации на персиски јазик, авторите вклучуваат различни комбинации за преферирање на различните зборовни групи, вклучувајќи и отфрлање на неколку најмалку фреквентни зборовни групи. Анализите потврдуваат дека некои од овие комбинации дури имаат и негативен ефект врз прецизноста на системот и дека најдобри резултати се добиваат доколку на различните зборовни групи им се зададат еднакви тежини. Сепак, највисока прецизност се постигнува кога информацијата за зборовната група се комбинира со едноставно стемирање (продуцирано од персискиот стемер *PERSTEM*, базиран на морфолошки правила) [150].
- *Kanaan et al.* [151] користат комплетно автоматизиран означувач базиран на правила и потврдуваат дека именките се најрелевантната зборовна група за прибирање документи напишани на арапски јазик. Заради тоа, авторите се одлучуваат за индексирање само на именките, кои сочинуваат повеќе од половина од вкупниот број термини во тест-колекцијата која ја анализираат. Ваквиот пристап се покажува дека ја редуцира употребата на дискот и ја подобрува брзината на системот за прибирање, со минимален ефект врз мерките за прецизност и отповикување. Резултатите добиени во ова истражување се комплетно во согласност со резултатите добиени при експериментите направени од *Chowdhury et al.* [152] за англискиот јазик.
- Клучно својство на кинескиот јазик, кој го прави процесот на означување на зборовните групи многу потежок отколку во другите јазици, е тоа што речениците се пишуваат без празно место меѓу карактерите. Бидејќи означувањето зависи од тоа како реченицата е поделена на зборови, неопходност за означувачот претставува реализирање на успешна сегментација на реченициите на зборови. Во таа насока, *Zhang et al.* [153] формулираат пристап кој ги интегрира сегментирањето на реченициите на зборови и означувањето на зборовните групи во единствен проблем. За таа цел, авторите користат унија од карактеристиките специфични за секоја од двете задачи во единствен алгоритам за учење (перцептрон) [154]. Ваквото интегрирање покажува подобрување на F – мерката околу 10% - 15%, во споредба со пристапот кој се реализира во две фази (прво сегментирање на реченициите, кое потоа е следено од означувањето на зборовите со соодветната зборовна група).
- *Ning et al.* [155] предлагаат хибриден означувач за кинескиот јазик, кој го комбинира пристапот базиран на правила со *HMM*. Овој означувач постигнува точност околу 92% - 94%, врз податоци за тестирање кои содржат и непознати зборови.

3.2.3. Означување на зборовната група во системот за одговарање прашања на македонски јазик

Македонскиот, како и останатите словенски јазици, е високо флективен јазик со богат морфолошки систем, кој разликува различни граматички функции и релации. За утврдување на влијанието на спецификите на македонскиот јазик во процесот на прибирање информации, една од целите на ова истражување е да се утврди колкава е значајноста на различните зборовни групи во самиот процес на пребарување. Поради недостапност на статистички означувач за означување на зборовите во согласност со нивната зборовна група (*part-of-speech tagger*), ова истражување го користи претходно развиениот, анотиран речник за македонскиот јазик[42]. Речникот ги содржи најчесто користените македонски зборови, заедно со множеството зборформи за секој поединечен збор. Во согласност со истражувањата направени за другите природни јазици, во испитувањата се вклучени пет зборовни групи од македонскиот јазик, и тоа: именките, глаголите, придавките, броевите и прилозите. Останатите зборовни групи (како: предлозите, сврзниците, честиците, извиците и модалните зборови) се зборови кои се проценети дека имаат исклучително мало влијание во утврдување на содржината на документите. Исто така, се изоставени и заменките, заради нивната честа поврзаност со проблемот за разрешување на кореференците. Односно, нивната поврзаност со процесот на спојување повеќе изрази кои се однесуваат на ист ентитет, што претставува клучна задача во областа извлекување информации (*IE*).

За секоја од петте наведени зборовни групи во македонскиот јазик својствени се различни категории [123], и тоа:

- за **именките** карактеристични се: родот, бројот и определеноста;
- **придавките** имаат исти граматички белези како и именките до кои стојат;
- потипични **глаголски** категории се: времето, начинот, лицето, видот, преодноста, залогот (дијатеза), родот и бројот;
- **броевите** немаат одделни форми за изразување на граматичките значења, освен за определеноста (можат да бидат со член или без член)[43], и
- **прилозите** се неменливи зборови во однос на можноста за суфиксација, а менливи од аспект на нивното степенување, кое се реализира со префикси.

Во системот за одговарање прашања на македонски јазик, се искористени карактеристиките само на именките. Секоја именка е означена во согласност со основната поделба на именките на општи и сопствени, и искористена е категоријата род (машки, женски, среден) (табела 7). За останатите зборови е земена само ознаката за зборовната група, без карактеристиките кои таа ги носи. Прелиминарните резултати потврдија дека најголемо влијание во пребарувањето имаат личните именки (*proper nouns – np*), следени од општите именки (*common nouns – n*). Помалку значајни се глаголите (*verbs – v*) и придавките (*adjectives – a*), додека прилозите (*adverbs – r*) и броевите (*numbers – m*) се зборовни групи со најслабо влијание. За да се оцени до кој

[42] http://nl.ijs.si/ME/V4
[43] Исклучок од ова прават само броевите еден и два.

степен зборовните групи помагаат во утврдувањето на релевантните информации, тестирани се различни вредности како нивни тежини. Распределбата на тежините на зборовните групи е дадена со следниве равенки:

- $np = x$,
- $n = x - a_x$,
- $a = x - 2a_x$, $v = x - 2a_x$,
- $m = x - 3a_x$, $r = x - 3a_x$,

(4)

каде x и a_x се променливи за кои се тестирани различни вредности. Земајќи го предвид фактот дека тежините имаат позитивни вредности, последната равенка имплицира дека $a_x < x/3$. Тестирани се вредности за x од 0.5 до 5.0, со чекор 0.5 и во согласност со вредноста на x, вредностите на a_x се движат од $x/10$ до $x/3$, со чекор $x/10$. Во табелата 5 се дадени сите тестирани вредности за x и a_x.

x	a_x
0.5	0.05, 0.1, 0.15
1.0	0.1, 0.2, 0.3
1.5	0.15, 0.3, 0.45
2.0	0.2, 0.4, 0.6
2.5	0.25, 0.5, 0.75
3.0	0.3, 0.6, 0.9
3.5	0.35, 0.7, 1.05
4.0	0.4, 0.8, 1.2
4.5	0.45, 0.9, 1.35
5.0	0.5, 1.0, 1.5

Табела 5. Тестирани вредности за x и a_x

3.3. Утврдување на групите од морфолошки поврзани зборови

Стемирањето е метод кој често се применува во претпроцесирачката фаза (при креирање на индексот со термини), но и во процесот на анализа на корисничките прашалници. Тој претставува суров хеуристички процес, дизајниран за да овозможи утврдување на морфолошки поврзаните зборови. Се заснова на идејата дека зборот може да се сведе на **основен збор** (*stem*), кој се однесува на одредена централна идеја или „значење“ и дека на него се додаваат афикси кои го модифицираат значењето и/или го прилагодуваат зборот за неговата синтаксичка улога [156]. При сведувањето на даден збор на основниот збор (стем), не се зема предвид контекстот во кој тој се јавува. Генерираниот основен збор не секогаш претставува заглавен збор во речникот зборови на одреден природен јазик, но сите варијанти на тој збор треба да се пресликаат во таа форма по извршување на процесот стемирање. Ова значи дека, во просек, стемирањето ја зголемува сличноста меѓу документите или меѓу документите и прашалникот, што резултира со зголемување на отповикувањето, на сметка на намалување на прецизноста.

Во одредени истражувања стемирањето значи отстранување само на флексијата од зборовите, односно отстранување на **формообразувачката морфема**, која означува

промена на родот, бројот, лицето, видот, итн. Во ваков случај, на пример, зборовите „емпирист“, „емпиристи“, „емпиристине“, кои претставуваат зборформи на зборот „емпирист”, се пресликуваат во него. Или стемирањето може да значи и отстранување на **зборообразувачките морфеми** од зборот (најчесто само суфиксот, некогаш и префиксот). Тоа значи пресликување на зборовите „стар“, „старец“, „старица“, „старост“ и „старина“ во зборот „стар“. Генерално, отфрлањето само на флексијата има позитивни ефекти во процесот на прибирање информации. Од друга страна, отфрлањето на зборообразувачките морфеми може различно да влијае врз успешноста на системот за прибирање информации, бидејќи постојат зборови кај кои отфрлањето на овие морфеми комплетно го менува нивното значење (таков пример е пресликувањето на зборот „организација“ во „орган“). Ова особено се однесува при бришење на префиксите, освен во одредени специфични домени, како медицината или хемијата, каде овој пристап има позитивен ефект.

Интересно поле за истражување е автоматското утврдување кога да се примени стемирањето, со цел да се избегне **престемирањето** (*overstemming*) и **потстемирањето** (*understemming*). **Престемирањето** подразбира пресликување на два збора со различен стем во иста основна форма, односно генерирање на лажно позитивни. Додека **потстемирањето** настанува кога за два збора кои имаат ист стем, се генерира различна основна форма, познато како лажно негативни. Претходните истражувања покажуваат дека ефектот од стемирањето зависи од системот во кој се применува и од податоците врз кои се работи. Исто така, бенефитот од стемирањето е тесно поврзан со природниот јазик врз кој се применува, бидејќи се покажува дека стемирањето ја подобрува ефективноста на системите кои обработуваат јазик со богата морфологија [157], [158], [159]. Заради тоа, најдобар начин е негово обучување и оптимизирање заедно со системот.

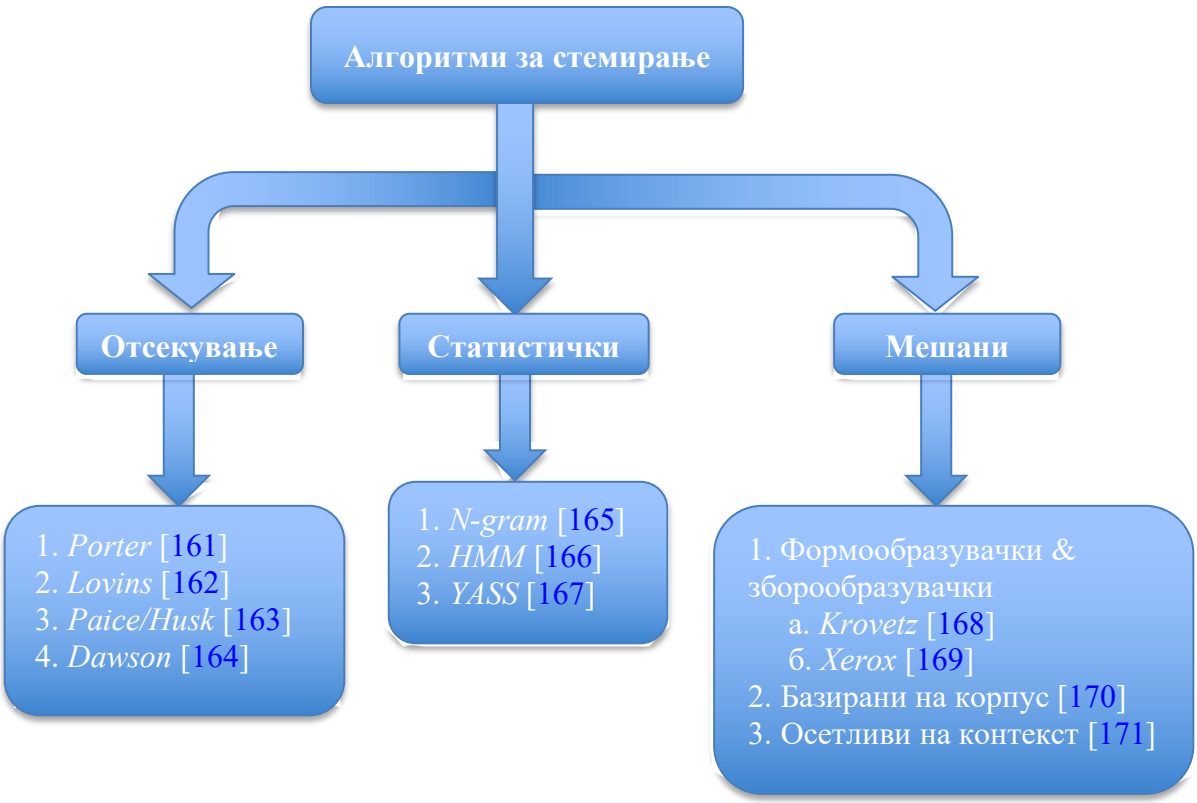
3.3.1. Класификација на алгоритмите за стемирање

Алгоритмите за стемирање генерално се класифицираат во три групи: **методи на отсекување** (*truncating methods*), **статистички методи** (*statistical methods*) и **мешани методи** (*mixed methods*). Секоја од овие групи алгоритми се карактеризира со свој специфичен начин на утврдување на основниот збор (стемот) на даден збор. На сликата 4 се наведени неколку најважни алгоритми од секоја група.

3.3.1.1. Методи на отсекување

Методите на отсекување применуваат дефинирани правила за отстранување на афиксите (префиксите и/или суфиксите) од даден збор. Наједноставниот стемер во оваа група е *Truncate (n)*, кој ги задржува првите n букви од зборот и ги отстранува сите останати. Зборовите чија должина е помала од n остануваат непроменети. Можноста за престемирање при овој метод се зголемува кога должината на зборот е помала. Друг едноставен пристап претставува *S-stemmer*, алгоритам кој ги соединува еднинската и множинската форма на англиските именки. Тоа го извршува користејќи правила за отстранување на суфиксите кои ја дефинираат множинската форма, со цел нивно конвертирање во еднинска форма [160].

Стемерот *Porter* (*Porter stemmer* [44]) е еден од најпопуларните алгоритми за стемирање, кој во повеќе истражувања се покажува мошне ефективен [161]. Предложен е во 1980-та година, а до денес се направени многу модификации и подобрувања на основниот алгоритам. Се заснова на идејата дека суфиксите во англискиот јазик (околу 1200) генерално се комбинација од помали и поедноставни суфикси. Алгоритамот се состои од пет чекори, при што во секој чекор се применуваат одредени правила, сè додека не бидат задоволени условите во едно од нив. Доколку одредено правило се прифати, се отстранува суфиксот соодветно, и се преминува на наредниот чекор. По проверката на сите пет чекори се прикажува генерираниот основен збор (стем). Алгоритмот содржи околу 60 правила и е лесен за разбирање. *Porter* дизајнира рамка за стемирање, позната како *Snowball*[45], чија главна цел е да им овозможи на програмерите да развиваат сопствени стемери за другите јазици. Во моментов постојат имплементации за јазици од романската, германската, руската и турската група, како и за уралските и скандинавските јазици.



Слика 4. Поделба на алгоритмите за стемирање

3.3.1.2. Статистички методи

Дизајнирањето на стемер за одреден природен јазик е исклучително напорна задача, а она што е уште позначајно, е неопходноста од јазична експертиза. Токму затоа, во истражувањата често се применуваат статистички методи во насока на креирање пристапи независни од јазикот, кои овозможуваат групирање на морфолошки поврзаните зборови.

[44] https://tartarus.org/martin/PorterStemmer/
[45] http://snowballstem.org/

YASS (Yet Another Suffix Stripper) Stemmer. Majumder et al. [167] опишуваат пристап базиран на кластерирање со цел да се определат групите зборови со ист корен и нивните морфолошки варијанти. Авторите дефинираат нови метрики за растојание на стрингови (*string distance measures*), D_1, D_2, D_3 и D_4, со цел да се кластерира речникот од зборови. Овие метрики пресликуваат пар стрингови X и Y во реален број, и притоа колку таа вредност е помала, толку сличноста меѓу X и Y е поголема.

Првата метрика D_1 се заснова на задавање тежина на позициите каде двата стринга содржат различен карактер, односно:

$$D_1(X,Y) = \sum_{i=0}^{n} \frac{1}{2^i} p_i \,, \tag{5}$$

каде $X = x_0x_1 \ldots x_n$, $Y = y_0y_1 \ldots y_{n'}$ $(n \geq n')$ и p_i е Буловата функција (за „казна“):

$$p_i = \begin{cases} 0, & \text{ако } x_i = y_i \text{ за } 0 \leq i \leq n' \\ 1, & \text{инаку} \end{cases}. \tag{6}$$

Првите испитувања на авторите потврдуваат дека поефективни би биле метриките кои ги земаат предвид совпаѓањата на карактерите на два стринга, сè до првиот различен карактер, и задавање казна на секоја наредна позиција. На ова согледување се базираат останатите три дефинирани метрики. Во равенките (7), (8) и (9), m ја означува позицијата на првото несовпаѓање во карактерите на X и Y (односно, $x_0 = y_0$, $x_1 = y_1$, … $x_{m-1} = y_{m-1}$, но $x_m \neq y_m$).

$$D_2(X,Y) = \begin{cases} \frac{1}{m} * \sum_{i=m}^{n} \frac{1}{2^{i-m}}, & \text{ако } m > 0 \\ \infty, & \text{инаку} \end{cases}, \tag{7}$$

$$D_3(X,Y) = \begin{cases} \frac{n-m+1}{m} * \sum_{i=m}^{n} \frac{1}{2^{i-m}}, & \text{ако } m > 0 \\ \infty, & \text{инаку} \end{cases}, \tag{8}$$

$$D_4(X,Y) = \frac{n-m+1}{n+1} * \sum_{i=m}^{n} \frac{1}{2^{i-m}}. \tag{9}$$

Точното растојание („сличноста“ во смисла на морфолошката поврзаност) меѓу два стринга (збора), X и Y, се добива преку множење на вкупната „казна“ со факторот кој е наменет да награди, доколку стринговите имаат повеќе заеднички карактери на почетокот, или да ги казни сите несовпаѓања од првиот различен карактер во стринговите (или двете). Колку е поголемо пресметаното растојание, толку е поголема морфолошката различност меѓу двата збора. Експериментите направени врз стандардните множества податоци за тестирање за англискиот и францускиот јазик

(*TREC* и *CLEF* множествата податоци) и врз креираната тест-колекција за *Bengali* јазикот (индиски јазик), потврдуваат дека предложените метрики во комбинација со **кластерирањето со комплетна врска** (*complete-link clustering*), се ефективни за јазици богати со суфикси. Уште повеќе, авторите заклучуваат дека:

- Разликите во ефективноста на четирите метрики за растојание на стрингови не се значајни, но метриката D_3 се покажува како најмалку чувствителна во однос на промена на вредноста на прагот.
- Успешноста на предложениот стемер за кластерирање на речникот зборови, без примена на никакво лингвистичко знаење, е споредлива со успешноста добиена со примена на стандардните стемери базирани на правила, како *Porter* и *Lovins* [162]. Споредливоста се однесува на просечната прецизност, како и вкупниот број на релевантни документи во процесот на прибирање информации.
- Предложениот алгоритам за стемирање го подобрува отповикувањето при прибирање информации на францускиот јазик, како и на индиските јазици.

N-gram stemmer. Утврдувањето на групите морфолошки поврзани зборови може да се направи применувајќи и **метрики за сличност на стрингови** (*string-similarity measures*) кои ги инкорпорираат n – грамите (n – *grams*) генерирани од зборовите. Под **n – грам** се подразбира низа од n последователни карактери, извлечени од даден збор (стринг). Групите може да содржат зборови кои споделуваат исти почетни n – грами, или споделуваат одреден дел од своите n – грами. Нивниот квалитет се подобрува доколку се имплементира соодветен алгоритам за кластерирање. Најчесто користени вредности за n се 2 и 3, односно се користат биграмите и триграмите од зборовите.

Меѓу метриките за сличност базирани на n – грами, најпопуларни се оние кои ги имплементираат коефициентите *Dice*, *Positional Dice* и *Jaccard* и *Overlap*, дадени со равенките (10) до (13), соодветно [165]:

$$Dice_{coef}(X,Y) = \frac{2 * |n(X) \cap n(Y)|}{|n(X)| + |n(Y)|}, \tag{10}$$

$$PosDice_{coef}(X,Y) = \frac{2 * |n_{pos}(X) \cap n_{pos}(Y)|}{|n_{pos}(X)| + |n_{pos}(Y)|}, \tag{11}$$

$$Jaccard_{coef}(X,Y) = \frac{|n(X) \cap n(Y)|}{|n(X)| + |n(Y)| - |n(X) \cap n(Y)|}, \tag{12}$$

$$Overlap_{coef}(X,Y) = \frac{|n(X) \cap n(Y)|}{min(|n(X)|, |n(Y)|)}, \tag{13}$$

каде X и Y се два збора за кои се утврдува „сличноста“ (во случајов, дали двата збора се морфолошки поврзани), $n(X)$ и $n(Y)$ се множествата од n – грами за зборовите X и

Y, соодветно, додека $n_{pos}(X)$ и $n_{pos}(Y)$ се множествата од n – грами, кои ја носат и својата позиција во зборовите X и Y, соодветно.

Релевантни истражувања. Статистичките методи за групирање на зборовите се широко имплементирани во различни истражувања. Кластерирањето базирано на метриките за сличност на стрингови за првпат е предложено од *Adamson et al*. [172] за англискиот јазик. Користејќи го *Dice*-коефициентот, авторите успешно кластерираат примерок од зборови во групи од семантички поврзани зборови. Во своето истражување *Robertson et al*. [173] заклучуваат дека сличноста на стрингови базирана на n – грами е погодна метрика за имплементација во ситуации каде треба да се пребарува голем речник со зборови. Освен за англискиот, статистичкото стемирање се покажува мошне ефективно и за други јазици. Така, *Ekmenkcioglu et al*. [174] го користат *Overlap*-коефициентот базиран на биграми и триграми за групирање на зборови извлечени од множество од политички вести, напишани на турски јазик. Анализите потврдуваат дека триграмите даваат подобри резултати во утврдување на точните варијанти на даден збор. *Snajder et al*. [175] применуваат стемирање базирано на сличност на стрингови врз морфолошки комплексниот хрватски јазик. Авторите испитуваат четири метрики, и тоа: *Dice* (со биграми и триграми) [165], D_3, D_4 од *YASS*-стемерот [167] и *Levenshtein distance* [176], за кластерирање на примерок зборови од тековни вести. Користејќи го алгоритмот за **кластерирање со просечна врска** (*average-link clustering*), за различни вредности на прагот, заклучуваат дека D_4 метриката генерира кластери со најдобар квалитет. Односно, оваа метрика најдобро се справува со зборообразувачките и формообразувачките морфеми во хрватскиот јазик, кој пред сè е богат со суфикси. *Majumder et al*. [177] го применуваат *YASS*-стемерот и врз унгарскиот и чешкиот јазик, и заклучуваат дека неговите перформанси се споредливи со оние добиени од достапните стемери базирани на правила, за овие два јазика.

3.3.1.3. Мешани методи

Еден од најчесто користените мешани методи за стемирање е методот предложен од *Xu et al*. [170]. Причината за дефинирање на овој метод е потребата да се надминат недостатоците на стемерот *Porter*, преку искористување на специфичноста на зборовите од корпусот врз кој истиот се применува. Стемерот *Porter* генерира низа проблеми, меѓу кои најзабележителни се:

- стемерот не е едноставен за разбирање и модифицирање,
- понекогаш здружува зборови кои имаат различно значење (на пример зборовите, *„policy“* и *„police“*), или не успева да здружи зборови кои имаат ист основен збор (како *„index“* и *„indices“*), и
- генерира основни зборови (стем) кои не секогаш се зборови и се тешки за интерпретација од страна на крајниот корисник (на пример, од зборот *„iteration“* генерира стем *„iter“*, од зборот *„general“* генерира стем *„gener“*).

Истражувањата посочуваат уште еден клучен проблем кој се појавува при користење на алгоритамите за стемирање базирани на правила, а тоа е дека нивната успешност

зависи од корпусот врз кој се применуваат. Затоа, *Xu et al.* [170] користат одредени статистики извлечени од корпусот (како: фреквенцијата на зборовите, фреквенцијата на заедничкото појавување на два збора во одреден текстуален сегмент (прозорец), и други.) за автоматска модификација на групите од зборови добиени со стемерот *Porter*, со цел тие да соодветствуваат на карактеристиките на корпусот. Основната идеја е дека варијантите на зборот кои треба да се изгрупираат за даден корпус, се појавуваат заедно (*co-occur*) во документите од корпусот. Поточно, авторите дефинираат нова метрика за мерење на значајноста на заемното појавување на зборoформите, дадена со:

$$em(a,b) = max\left(\frac{n_{ab} - En(a,b)}{n_a + n_b}, 0\right), \tag{14}$$

каде n_a и n_b се бројот на појавувања на зборовите a и b во корпусот, соодветно, додека n_{ab} е бројот на појавувања на a и b во рамките на дефиниран прозорец од зборови во корпусот. Попрецизно, n_{ab} е бројот на елементи во множеството:

$$\{< a_i, b_j > \left| dist(a_i, b_j) < window\},$$

каде a_i и b_j се различни појавувања на зборовите a и b во корпусот, додека $dist(a_i, b_j)$ е растојанието меѓу a_i и b_j, измерено со бројот на зборови во дефинираниот прозорец. $En(a,b)$ е очекуваниот број на заемно појавување, под претпоставка дека a и b се статистички независни. Улогата на $En(a,b)$ во равенката (14) е мошне важна, бидејќи два збора може случајно да се појават заедно. На пример, предлогот *„of“* и определениот член *„the“* од англискиот јазик, скоро секогаш се појавуваат заедно во текстуален прозорец со одредена големина, но сосема е погрешно да се заклучи дека овие два збора се поврзани. За пресметување на $En(a,b)$ авторите ја применуваат следнава формула:

$$En(a,b) = kn_a n_b, \tag{15}$$

каде k е константен фактор, чија вредност зависи од корпусот и големината на прозорецот, а $em(a,b)$ вредноста се користи за редефинирање на групите од зборови генерирани со стемерот *Porter*.

Користејќи го овој концепт, авторите успеваат да подобрат одредени недостатоци настанати како резултат на престемирањето и потстемирањето генерирани од стемерот *Porter*, врз корпус на англиски јазик. Истиот покажува подобри резултати во однос на конвенционалните пристапи и за корпус текстови на шпански јазик. Главната моќ на овој метод е можноста да се одбегнат здружувања на зборови, кои не се соодветни за конкретниот корпус. Од друга страна, главните недостатоци се потребата од утврдување вредност на константниот фактор k, за секој корпус пооделно, и зголеменото време за обработка.

3.3.2. Автоматско групирање на збороформи во македонскиот јазик

Во своето истражување, *Jovanovska et al*. [178] креираат три различни речници од зборовите кои се појавуваат во документите од тест-колекцијата на македонски јазик од областа на информатичките технологии, именувани како: Речник_1, Речник_2 и Речник_3. Тест-колекцијата од документи и прашања со повеќекратен избор е детално опишана во секцијата 5.1. При тоа, Речникот_1 ги содржи уникатните зборови (типовите) од документите, Речникот_2 содржи групи од зборови од Речникот_1 кои припаѓаат во иста лексема, додека Речникот_3 содржи групи од зборови од Речник_1 кои се сведуваат на ист основен збор (стем). Овие речници се вклучени во процесот на одговарање прашања од оваа колекција и е забележано дека највисока точност се добива доколку во прибирањето информации се искористат збороформите на секој клучен збор од прашалникот (односно, доколку се примени Речникот_2). Потоа следи примената на Речник_3, а најниска точност се добива доколку во прибирањето информации се користи Речник_1, односно доколку се искористат само клучните зборови од прашалникот, без инкорпорирање на нивните збороформи и зборовите со кои споделуваат ист основен збор.

Токму поради заклучоците донесени во спомнатото истражување, следната поставена цел е да се утврди ефективен начин за генерирање на Речникот_2, за даден корпус на македонски јазик, кој би се користел во процесот на прибирање информации. Клучен проблем кој овде се наметна е непостоењето на лематизатор (*lemmatizer*) за македонскиот јазик, како и стемер (*stemmer*). Тоа наложи потреба од рачно креирање на Речникот_2 (воедно и Речникот_3) (секција 4.2.1), што претставува исклучително макотрпна работа, за која е неопходна и стручна експертиза. Уште едно прашање кое произлезе од претходните истражувања во однос на прибирањето информации на македонски јазик [178] е дали примената на точно дефинираните групи од зборови кои припаѓаат во иста лексема е најдобрата практика за овој процес. Со цел да се одговори на спомнативе пројавени согледувања, ова истражување имплементира статистички пристап за автоматско генерирање на Речникот_2 (односно утврдување на групите од збороформи) и дава преглед на експерименталните резултати добиени со негово инкорпорирање во процесот на прибирање информации на македонски јазик (секција 4.3.1.3).

3.3.2.1. Метрики за сличност на стрингови базирани на n – грами

За утврдување на групите зборови кои припаѓаат на иста лексема, ова истражување ги користи метриките за сличност на стрингови базирани на коефициентите *Dice*, *Positional Dice* и *Jaccard*. Тие се дадени со равенките (16), (17) и (18), соодветно.

$$dis_{Dice}(X,Y) = 1 - Dice_{coef}(X,Y)\,, \tag{16}$$

$$dis_{PosDice}(X,Y) = 1 - PosDice_{coef}(X,Y)\,, \tag{17}$$

$$dis_{Jaccard}(X,Y) = 1 - Jaccard_{coef}(X,Y)\,. \tag{18}$$

При тоа, „сличноста“ на два збора (која овде означува припадност на зборовите во иста лексема) е утврдена со вклучување на нивните биграми (два последователни карактери од еден збор). Во согласност со дадениве равенки, може да се заклучи дека два збора се идентични доколку растојанието меѓу нив е еднакво на нула. Максимално различни се зборовите чие растојание е еднакво на еден. Треба да се потенцира дека од дадените метрики само метриката која го имплементира *Positional Dice*-коефициентот ја зема предвид позицијата на биграмите во самиот збор, што е исклучително важно за утврдување дали два збора се зборформи од иста лема во македонскиот јазик.

3.3.2.2. Метрика за сличност базирана на триаголниот прозорец

Истражувањето применува и нова метрика за сличност на стрингови базирана на **триаголниот прозорец** (секција 3.4.2). Мотивацијата за примена на прозорска функција (со цел групирање на зборформите), дојде од идејата да се искористат овие функции во последната фаза од системот за одговарање прашања на македонски јазик, односно во фазата на селектирање на точниот одговор. Имено, прозорските функции продуцираат поголеми тежини за зборовите кои се појавуваат поблиску еден до друг во текстуален сегмент (детално објаснување е дадено во секцијата 3.4.2). Од тоа произлезе интересот да се испита дали овие тежини може да се искористат за автоматско групирање на зборформите, односно дали истите може да се применат за поголемо наградување на подолгите идентични низи од карактери со кои започнуваат два збора, и за задавање казна на првиот карактер во кои двата збора се разликуваат, како и на секој наредец. Во продолжение следува објаснување како е дефинирана триаголната функција (*triangular function*).

Слика 5. Триаголен прозорец генериран од равенките $p_1(x)$ и $p_2(x)$

Нека X и Y се два збора претставени со следниве низи карактери:

$$X = x_0 x_1 \dots x_m \dots x_n$$

$$Y = y_0 y_1 \dots y_m \dots y_{n'}.$$

Нека m е позицијата на првиот карактер во кој се разликуваат двата збора (т.е., $x_0 = y_0$, $x_1 = y_1$, $\dots$ $x_{m-1} = y_{m-1}$, но $x_m \neq y_m$) и нека n и n' се должините на X и Y, соодветно. Без губење на општоста, може да се претпостави дека $n \geq n'$. Равенките $p_1(x)$ и $p_2(x)$, кои минуваат низ точките $(0,0)$, $(m,1)$ и $(m,1)$, $(n,0)$, соодветно, го дефинираат триаголниот прозорец (*triangular window*), прикажан на сликата 5. Равенките се дадени со следниве изрази:

$$p_1(x) = \frac{1}{m}x\,, \tag{19}$$

$$p_2(x) = \frac{1}{m-n}x - \frac{n}{m-n}. \tag{20}$$

Важно е да се потенцира дека збороформите во македонскиот јазик се разликуваат во неколку букви (карактери) на крајот од стринговите, кои ги претставуваат тие збороформи. Токму затоа, новата метрика ја користи функцијата $p_2(x)$ за дефинирање на тежини со кои се задава казна на буквите, во кои се разликуваат двата збора во своите завршетоци. Од друга страна, функцијата $p_1(x)$ се користи за да се дефинираат тежини со кои се наградуваат буквите на почетокот од двата збора, во кои тие се совпаѓаат. Максималната казна се дава за првото несовпаѓање во буква меѓу двата збора X и Y, т.е. $p_1(m) = p_2(m) = 1$. Новата метрика е дефинирана со равенката (21):

$$dis_{Triangle}(X,Y) = \frac{p_2(m) + p_2(m+1) + \dots + p_2(n)}{p_1(0) + p_1(1) + \dots + p_1(m-1)}. \tag{21}$$

Користејќи ги равенките (19) и (20), финалната форма на метриката за сличност базирана на триаголниот прозорец е дадена со следниов израз:

$$dis_{Triangle}(X,Y) = \frac{n-m+1}{m-1}. \tag{22}$$

За генерирање на групите зборови со имплементација на новодефинираната метрика (како и метриките дадени со равенките (16), (17) и (18)), искористено е **кластерирањето со комплетна врска** (*complete-link clustering*). Добиените групи се вклучени во самиот процес на прибирање пасуси за дадено прашање поставено на македонски јазик, кои системот потоа подетално ги разработува, со цел да го пронајде точниот одговор. Успешноста во прибирањето пасуси добиена со четирите наведени метрики е презентирана и анализирана во секцијата 4.3.1.3, за различни вредности на прагот, со кој се утврдува дали два збора припаѓаат во иста група или не.

3.4. Близината на термините како значаен фактор во процесот на одговарање прашања

Една од причините заради која широко се применува прибирањето пасуси кај сегашните системи за одговарање прашања, е интуицијата дека одговорите на најголем дел од прашањата може да се пронајдат во прилично кратки текстуални сегменти, како една или две реченици. Ова секако зависи од категоријата на самото прашање, бидејќи некои од нив, како на пример описните и процедуралните прашања, побаруваат поопширни одговори. Фактот дека најголем дел од одговорите се дадени локално во документот, доведува до два заклучока кои се однесуваат на прибирањето документи, како составен дел од процесот на автоматско одговарање прашања. Првиот се однесува на фактот дека прибирањето на документите како метод треба да ја земе предвид близината на термините од прашалникот и да ги рангира повисоко документите каде тие се појавуваат поблиску еден до друг. Вториот заклучок е дека методот на прибирање би требало да прикаже фрагменти од документите, кои покажуваат висока блискост меѓу термините од прашалникот, наместо целите документи.

Токму овие две побарувања ги задоволува **прибирањето на пасуси**. Сепак, експериментите кои може да се најдат во литературата не покажуваат значителни подобрувања на прибирањето пасуси во однос на прибирањето документи, кога се применува во рамки на системите за одговарање прашања [26]. Исто така, се покажува дека во многу ситуации прибирањето пасуси води кон значително намалување на перформансите на *QA*-системите. Но, и покрај овие забелешки не може со сигурност да се каже дека прибирањето пасуси е несоодветно кога се работи за одговарањето прашања. Сепак, треба да се потенцира дека параметрите кои се вклучени во овој процес, како: големината на пасусите, степенот на поклопување меѓу пасусите, фиксната наспроти променливата должина, итн., треба внимателно да се селектираат. Уште повеќе, овие параметри може да бидат силно зависни од самата колекција и прашалниците.

Друга можност која ги задоволува двете побарувања дадени претходно, е **прибирањето базирано на близина** (*proximity-based retrieval*), кое не побарува дефинирање на одредени параметри. Во овој случај, близината е изразена со растојанието меѓу термините, за чие дефинирање во литературата може да се сретнат два клучни пристапи, и тоа:

- **Пристап базиран на парови** (*pairwise based approach*). Овој пристап ја мери близината меѓу два термина, во согласност со бројот на термини кои се појавуваат меѓу нив. Дефинирањето на близината меѓу два термини е тривијално, но сепак постојат неколку начини за разрешување на овој проблем, доколку се вклучени повеќе термини [179].
- **Пристап базиран на покривки** (*span-based approach*). Овој пристап утврдува текстуален сегмент од документот, наречен **покривка** (*spans*), кој ги покрива сите термини заеднички за прашалникот и документот.

Моделирањето на близината (*distance modeling*) може да се смета како индиректен начин за доловување на корелацијата меѓу термините. Токму во таа насока, *Beeferman et al*. [180] во своето истражување потврдуваат дека близината на термините има силно влијание врз нивната корелација. Затоа, нејзиното инкорпорирање во системите за прибирање информации (и *QA*-системите) станува сè поактуелно во последно време.

3.4.1. Релевантни истражувања

Во оваа секција се наведени релевантни истражувања кои ја инкорпорираат близината на термините во различни модели за прибирање информации.

3.4.1.1. Инкорпорирање на близината на термините во *Okapi BM25*

Во литературата може да се сретнат истражувања кои ја комбинираат близината на термините со добро познатиот *Okapi BM25* веројатностен модел.

- Така, *Song et al*. [181] предлагаат флексибилен пристап за разделување на документот на покривки, при што секоја покривка започнува и завршува со термин од прашалникот и не мора да ги содржи сите термини од него. Покривките се вклучени во *BM25* моделот за пребарување, на тој начин што фреквенцијата на термин од прашалникот е заменета со вредност која укажува колкав е придонесот на тој термин кон релевантноста (*relevance contribution*). Придонесот е определен од должината на покривките во кои се појавува терминот, бројот на други термини од прашалникот кои се појавуваат во тие покривки и неколку параметри за подесување. Оваа техника авторите ја означуваат како *BM25-P3* и истата ја тестираат врз *TREC-9, 10, 11* тест-колекциите. Анализите потврдуваат подобрување на прецизноста во пребарувањето, во споредба со неколку други модели, меѓу кои и стандардниот *Okapi BM25*.
- *He et al*. [182] предлагаат подобрување на класичниот *Okapi BM25* модел, каде наместо унигрмите (зборовите), ги вклучуваат n – грамите од документите. Тие дефинираат нови методи, и тоа: методот за броење на n – грамите во рамки на прозорец (*window-based N-gram Counting method*) и анализата на преживување врз различни статистики (*Survival Analysis over different statistics*). Екстензивните експериментирања врз стандардните *TREC*-колекции потврдуваат дека овој метод, наречен *BM25P*, доведува до значителни подобрувања во пребарувањето во однос на стандардниот *BM25* и позициските модел на јазик.
- *Tao et al*. [179] систематски ја испитуваат близината на термините од прашалникот и ја проценуваат нејзината корелација со релевантноста на документот во процесот на прибирање информации. Авторите дефинираат три метрики базирани на парови, и тоа: **минимално** (*MinDist*), **просечно** (*AveDist*) *и* **максимално растојание по парови** (*MaxDist*). Тие се соодветно определени со најмалата, просечната и најголемата вредност на растојанието меѓу сите парови термини од прашалникот, кои се појавуваат во документот. Во своите анализи користат и пристапи базирани на покривки. При тоа, **покривката** ја дефинираат како должина на фрагмент од документот кој ги покрива сите појавувања на терминте од прашалникот,

вклучувајќи ги и нивните повторувања, додека **минималната покривка** ја дефинираат како должина на најкраткиот фрагмент од документот кој го покрива секој термин од прашалникот барем еднаш. Експериментите направени врз пет стандардни *TREC*-тест-колекции покажуваат дека *MinDist* инкорпорирана во *Okapi BM25* дава најдобри резултати во однос на останатите метрики за близина и значително подобрување на пребарувањето во однос на основниот *Okapi BM25* модел.

3.4.1.2. Инкорпорирање на близината на термините во моделите на јазик

Примената на близина на термините покажува одредено ниво на успешност и во контекст на пристапот кој ги користи моделите на јазик (*language models*).

– *Lv et al*. [183] предлагаат нов модел, наречен **позициски модел на јазик** (*Positional Language Model - PLM*). Клучната идеја на авторите е дефинирање на модел на јазик за секоја позиција во документот. Моделите на јазик за сите позиции во документот се комбинираат со цел да се оцени неговата релевантност во однос на даден прашалник. Авторите претпоставуваат дека термин на одредена позиција може да „пропагира“ доказ (*evidence*) за своето појавување на други позиции во склоп на истиот документ, при што позициите кои се поблиску до терминот добиваат поголем дел од доказот од оние кои се подалеку. На овој начин, секоја позиција би добила пропагиран доказ (тежина) од сите термини во документот, со тоа што поголема тежина доаѓа од термините кои се поблиску до таа позиција. За да се процени функцијата за пропагирање, која се инкорпорира во моделот на јазик и го дефинира моделот на јазик за одредена позиција, авторите анализираат неколку кернел функции базирани на близина. Од нив најдобри резултати дава Гаусовата кернел функција. Што се однесува до порамнувањето во моделот на јазик, *Dirichlet* методот се покажува подобар отколку порамнувањето на *Jelinek-Mercer*. Експерименталните резултати на *TREC*-тест-колекциите покажуваат дека *PLM* е ефективен за прибирање пасуси и е подобар од актуелните модели за прибирање базирани на близина [179].
– *Zhao et al*. [184] развиваат нов начин за интегрирање на факторот за близина во униграм моделот на јазик (добиениот модел е наречен *Proximity Language Model-PLM*). За добивање на овој модел ја користат метриката за минимално растојание по парови на *Tao et al*. [179]. Збирот од минималните растојанија го конвертираат во вредност која директно ја вметнуваат во фреквенцијата на термините во *KLD* (*Kullback-Leibler Divergence*) моделот на јазик. Иако нивните резултати покажуваат повисока **средно просечна прецизност** (*Mean Average Precision - MAP*) во споредба со оние на *Tao et al*. [179], во повеќе од 50% од направените експерименти, резултатите не покажуваат значително подобрување во однос на основниот *KLD* моделот на јазик.

3.4.1.3. Поврзаност на близината на термините со машинското учење

Најголем дел од пристапите *Learning to Rank* користат карактеристики изведени од заемното појавување на термините од прашалникот. На пример, *Cummins et al*. [185] го применуваат машинското учење за да развијат функција за близина на термини, која ја оптимизира средно просечната прецизност (*MAP*). Тие земаат предвид различни статистики, како на пример, просечното и минималното растојание меѓу термините од прашалникот во документот. Иако функциите за прибирање добиени како резултат на процесот на машинското учење ја зголемуваат *MAP* на различни тест-колекции, се покажува дека овие подобрувања не секогаш се статистички значајни. Она што е особено значајно е дека вака генерираните функции не му овозможуваат на човекот да разбере на кој начин близината и релевантноста навистина се поврзани. Едноставно, само ја потврдуваат интуицијата дека таква врска навистина постои.

Конечно, треба да се потенцира дека близината на термините може да се искористи во секоја фаза од *IR* и *QA*-системите. Постојат истражувања кои покажуваат дека оваа информација може да ја подобри дури и селекцијата на термините за проширување на прашалникот за прибирање [186], [187].

3.4.2. Прозорски функции

Одредени истражувања ја користат и густината на појавување на клучните зборови (во одреден текстуален сегмент), која може да изрази тесна поврзаност меѓу нив [188]. Кај системите за одговарање прашања, растојанијата меѓу низата зборови која го претставува точниот одговор и клучните зборови извлечени од прашањето се исклучително важни [189]. За таа цел, пожелно да се утврди функција која задава поголема тежина на клучните зборови, доколку тие се појавуваат погусто во одреден сегмент. Ова својство го задоволуваат **прозорските функции** (*window functions*). Постојат повеќе типови на прозорски функции, но во пребарувањето најчесто се користени функциите кои се засноваат на правоаголениот и триаголениот прозорец (*rectangular and triangular window*) и прозорецот познат како *Hanning* (*Hanning window*).

Суштинската разлика меѓу различните типови прозорски функции е различната важност која ја задаваат на термините. Така, при користење на правоаголен прозорец се претпоставува дека термините кои се наоѓаат во тој прозорец имаат исто влијание во задачата за која се применува прозорската функција. Триаголниот прозорец рефлектира линеарна корелација меѓу влијанието и растојанието на термините, додека *Hanning*-прозорецот овозможува поблиските термини да имаат поголема важност отколку термините кои се повеќе раздалечени. Споредбите на различните прозорци направени во неколку истражувања, потврдуваат дека *Hanning*-прозорецот дава најдобри резултати во процесот на прибирање документи [190], [191]. Така, *Kise et al*. [192] ја имплементираат *Hanning*-прозорската функцијата со цел прибирање пасуси во процесот на одговарање прашања. Авторите забележуваат дека прво рангирани се пасусите во кои густината на термините од прашалникот е највисока. Резултатите од извршените експерименти потврдуваат дека овој метод ги надминува конвенционалните методи за прибирање (како методот на векторски простор, латентното семантичко

индексирање (*Latent Semantic Indexing – LSI*) и псевдорелевантната повратна врска (*Pseudo Relevance Feedback*)), особено при пребарување на долги документи со кратки прашалници. Исто така, се забележува дека за нешто повеќе од половина од прашалниците, точниот одговор е лоциран во прворангираниот пасус, во прозорец со 50 термини.

3.4.2.1. Инкорпорирање на *Hanning*-прозорецот во системот за одговарање прашања на македонски јазик

Нека W е големина на прозорецот. *Hanning*-прозорската функција е дефинирана со равенката:

$$f(x) = \begin{cases} \frac{1}{2}(1 + cos\, 2\pi \frac{x}{W}), \text{ако } |x| \leq W/2 \\ 0, \text{инаку} \end{cases}, \quad (23)$$

додека нејзината форма е прикажана на сликата 6.

Слика 6. *Hanning*-прозорска функција

Hanning-прозорската функција е имплементирана во системот за одговарање прашања на македонски јазик во последната фаза, односно фазата за утврдување на точниот одговор од четирите понудени одговори. Нека p_1, p_2, p_3, p_4 и p_5 се петте најдобро рангирани пасуси во фазата за прибирање информации (за одредени прашања дополнително се вклучени и пасусите чија рангирачка вредност е еднаква на вредноста на петтиот рангиран пасус). За секое прашање q и четирите понудени одговори a_1, a_2, a_3 и a_4, се креирани четири прашалници: $q_1 = q + a_1$, $q_2 = q + a_2$, $q_3 = q + a_3$ и $q_4 = q + a_4$. Односно, секој прашалник ги содржи зборовите од прашањето, заедно со зборовите од соодветниот понуден одговор. *Hanning*-прозорската функција е искористена за да се утврди во кој дел од петте прибрани пасуси, најгусто се распоредени зборовите од секој прашалник. Доколку зборовите од прашалникот q_i (за некое $i = \overline{1,4}$) најгусто се распределени во некој од прибраните пасуси, тогаш се претпоставува дека тој прашалник го содржи точниот одговор на прашањето.

Утврдувањето на точниот одговор во системот за одговарање прашања на македонски јазик се врши со примена на следниов алгоритам:

– За прашалникот q_1 постапката е следна:
  – За пасусот p_1:
    – Ги лоцираме сите зборови од понудениот одговор a_1 кои се наоѓаат во пасусот p_1. Нека нивните позиции се означени со: l_{11}, l_{12}, ... и l_{1a_1}.
      – За позицијата l_{11} се разгледува прозорецот $\left(l_{11} - \frac{w}{2}, l_{11} + \frac{w}{2}\right)$, каде w е големината на прозорецот, и во него се лоцираат останатите зборови од прашалникот q_1. **Густината на распределба** (density *distribution* – *DD*) за овој пасус за позицијата l_{11} и прозорецот w е дефинирана со вредноста [192]:

$$DDp_1(l_{11}) = \sum_{x=-w/2}^{x=w/2} f(x)b_1(l_{11} - x),$$

      каде $f(x)$ е *Hanning*-прозорската функција, додека $b_1(l_{11} - x)$ е тежината на терминот кој се наоѓа на позицијата $l_{11} - x$.
      – Слично се пресметуваат и вредностите на DDp_1 за останатите позиции, односно се пресметуваат: $DDp_1(l_{12})$, ..., $DDp_1\left(l_{1a_1}\right)$.
    – Нека $D_1 = max\{DDp_1(l) | l \text{ е позиција на збор од } a_1 \text{ во } p_1\}$, односно D_1 е вредноста со која се проценува максималната густина на зборовите од q_1 во пасусот p_1 (тоа се постигнува со *Hanning*-прозорската функција, која задава повисоки тежини доколку зборовите се појавуваат поблиску еден до друг).
  – Постапката се повторува и за останатите пасуси p_2, p_3, p_4 и p_5. Нека соодветните максимални вредности се D_2, D_3, D_4 и D_5.

    Нека $S_1 = max\{D_1, D_2, D_3, D_4, D_5\}$, односно вредноста на q_1 е S_1 (добиена во согласност со петте искористени пасуси).
– Истата постапка се повторува и за прашалниците q_2, q_3 и q_4, чии соодветни вредности се S_2, S_3 и S_4. На крај, се претпоставува дека прашалникот со максимална вредност го содржи точниот одговор.

Резиме:

Врската меѓу *NLP*, *IR* и *QA*-областите е заемна и исклучително силна. Така, задачите кои се клучни за **обработката на природните јазици** (како: морфологијата, синтаксата и семантиката), претставуваат важни аспекти и за системите за **прибирање информации** (*IR*), за општ и специфилен домен. Од друга страна, *IR*-системите се само еден сегмент од ланецот на комплексната обработка на текстуалните корпуси. Во таа насока е и дизајнот и имплементацијата на системите за **одговарање прашања** (*QA*), како и системите за **рударење на мислења** (кај кои прибраниот текстуален сегмент не е поврзан за конкретна тема, туку за лични размислувања). Во ваквите случаи, прибраните сегменти мора да се подложат на постобработка, применувајќи комплексен *NLP*-систем, со цел да се извлечат кратки и концизни одговори на прашањата, или да се утврди дали истите содржат лични размислувања за одредена тема. Последниве години, исклучително актуелна е примената на техниките за машинско учење во развојот на *IR* (*QA*) системите, со значајна *NLP*-компонента.

Поконкретно, ова поглавје дава преглед на влијанието на зборовните групи и збороформите во процесот на прибирање информации на англиски јазик (како и други јазици). Во него е оценета и улогата која ја има близината на клучните зборови од прашалникот во овој процес. Забележаните позитивни резултати за другите јазици, се причина за комбинирање на различни техники од *NLP* и *IR*-областите, во насока на градење систем за успешно одговарање прашања поставени на македонски јазик. Во отсуство на лематизатор за македонскиот јазик, ова истражување дефинира нова метрика за автоматско групирање на збороформите во македонскиот јазик, која се базира на триаголниот прозорец (како пример на прозорска функција). Истражувањето дефинира и постапка за имплементација на *Hanning*-прозорската функција (за определување сегменти во кои клучните зборови се густо распределени), со цел да се открие единствениот точен одговор меѓу четирите понудени за даденото прашање од тест-колекцијата на македонски јазик.

4. Тест-колекција и експериментални резултати

„Само преку неуспехот и експериментирањето можеме да учиме и да се развиваме.“

Isaac Stern

Ова поглавје започнува со приказ на предизвиците карактеритични за процесот на креирање и анализирање на **прашањата со повеќекратен избор**. Во него е вкучен и преглед на истражувањата кои укажуваат колкаво влијание може да има начинот на формулирање на самите прашања и изборот на можни одговори врз исходот добиен со нивната примена во процесот на утврдување на знаењето. Потоа, опишана е тест-колекцијата документи и прашања со повеќекратен избор на македонски јазик од областа филозофија, врз која се извршени експериментите. Дадена е распределбата на прашањата по категории, опишан е системот за одговарање прашања на македонски јазик, методите и техниките кои се имплементирани во секој модул, и на крај детално се прикажани експерименталните резултати. Поглавјето завршува со сумирање на пристапите кои даваат најдобар резултат за опишаната тест-колекција.

4.1. Прашања со повеќекратен избор

Прашањата со повеќекратен избор (*Multiple Choice Questions – MCQ*), во продолжение означени со **ППИ**, може да бидат ефективен и ефикасен начин за оценување на постигувањата во процесот на учење. Тие се состојат од зададен проблем, познат како **корен**, и листа од предложени решенија, наречени **алтернативи**. Алтернативите најчесто вклучуваат еден **точен одговор**, додека останатите се инфериорни алтернативи, наречени **дистрактори** [193]. Притоа, задача на испитаникот е да селектира една алтернатива, за која смета дека претставува најдобар одговор на поставениот проблем. Тестовите кои се базираат на овие прашања имаат неколку потенцијални предности, како што се:

- **Многустраност (*versatility*):** Тестовите може да се искористат за утврдување на различни нивоа на постигнувања во учењето, од најосновни знаења, па сè до правење анализи и проценки. Бидејќи се прави избор од множество потенцијални одговори, сепак овие прашања не можат да се применат за секое тестирање. На пример, тие не претставуваат ефективен начин за тестирање на способноста за организација на размислувањата или изразувањето на креативните идеи.
- **Веродостојност (*reliability*):** Веродостојноста се дефинира како степен до кој тестот доследно го мери исходот од учењето. Имено, оценувањето со примена на прашањата со повеќекратен избор е пообјективно во однос на оценувањето со помош на есејските прашања, кои се подложни на неконзистентноста на оценувачот.
- **Валидност (*validity*):** Валидноста се дефинира како степен до кој тестот успева да го измери исходот од учењето за кој е наменет. Бидејќи прашањата со повеќекратен избор обично се одговараат побрзо од есејските прашања,

тестовите кои се базираат на нив можат да се фокусираат на релативно поголем дел од материјалот, со што се зголемува валидноста на оценувањето [194].

И покрај овие предности, оценувањето на постигнувањата со помош на тестови базирани на прашања со повеќекратен избор често е критикувано. Одредени автори потенцираат дека овие тестови ја оценуваат способноста на испитаниците за меморирање, а не за разбирање, примена и анализа на одредени информации [195]. Сепак, сосема е јасно дека доколку внимателно се напишани, истите може да се искористат и за тестирање на размислувањето на повисоко ниво, преку креирање на прашања кои се фокусираат на повисоки нивоа на разбирање, дефинирани со Блумовата таксономија [196]. Таквите прашања имаат корен кој претставува проблем, за чие разрешување е неопходна анализа и примена на одредени принципи од испитаната област. Исто така, алтернативите можат да придонесат во овој процес, преку потребата истите да се проценат. Креирањето на ваквите прашања побарува повеќе способности и вештини, отколку креирањето прашања за кои е неопходно меморирање [197].

Клучот за искористување на наведените предности на прашањата со повеќекратен избор е конструкција на „добри“ прашања. Додека многустраноста и веродостојноста се својствени карактеристики на самите прашања со повеќекратен избор (ППИ), нивната валидност не може да се претпостави заради можноста испитаникот без соодветно знаење случајно да го погоди точниот одговор. Затоа, за да се зголеми точноста на проценката применувајќи ги прашањата со повеќекратен избор, се препорачува имплементација на стандардни протоколи за **превалидација** (*prevalidation*) и **поствалидација** (*postvalidation*).

4.1.1. Стандардни протоколи за превалидација на прашањата со повеќекратен избор

Она што особено е важно при креирањето на тестови со прашања со повеќекратен избор, е вклучувањето само на прашања кои се меѓусебно независни. Тоа ќе му оневозможи на испитаникот да искористи информации од едно прашање за да одговори друго и со тоа да ја намали валидноста на тестот. Превалидацијата спречува појава на грешки при конструкцијата на ППИ, преку примена на низа инструкции.

За конструкција на ефективен корен, потребно е да се следат следниве клучни насоки:

- Коренот треба да биде значаен сам за себе и да претставува јасен проблем. Таквиот корен овозможува фокусирање на исходите од учењето (пример Б1).
- Коренот не треба да содржи ирелевантни информации кои можат да ја намалат веродостојноста и валидноста на резултатот од тестот (пример Б2) [198].
- Коренот смее да биде искажан со негација **само** доколку тоа го побарува важен исход од учењето. Истражувањата потврдуваат дека испитаниците имаат потешкотии во разбирањето на проблемите кои се искажани со негација [199]. Доколку навистина е неопходно негативно изразување за оценување на компентенциите (на пример, при идентификување на опасна клиничка пракса и

слично), тогаш негацијата мора да биде потенцирана (со курзив) или запишана со големи букви (пример Б3).

- Коренот треба да биде прашална реченица или евентуално реченица со надополнување [200]. Истражувањата потврдуваат дека когнитивното оптоварување се зголемува доколку коренот се конструира со почетно или внатрешно празно место (пример Б4).

Од друга страна, процесот на креирање на алтернативите треба да ги задоволи следниве побарувања:

- Сите алтернативи мора да бидат уверливи. Неточните алтернативи треба да служат како функционални дистрактори. Вообичаените грешки кои ги прават испитаниците, претставуваат најдобар извор на дистрактори (пример Б5).
- Алтернативите треба да бидат јасно искажани и концизни. Опширните алтернативи ја оценуваат способноста на испитаниците за читање, наместо нивните постигнувања во учењето.
- Алтернативите треба да бидат меѓусебно исклучителни. Оние кои содржат преклопувачка содржина се сметаат за трик прашања од страна на испитаниците и може да ја поткопаат довербата во процесот на тестирањето.
- Алтернативите треба да бидат хомогени по содржина. Хетерогените алтернативи може да му помогнат на испитаникот да дојде до точниот одговор.
- Алтернативите треба да бидат ослободени од показатели кои укажуваат на точниот одговор, бидејќи денешните испитаници се подготвени истите да ги забележат. Заради тоа, важно е алтернативите да бидат граматички конзистентни со коренот, да имаат слична форма, слична должина и да се користи ист стил во нивното изразување.
- Треба да се избегнуваат алтернативите: **„сите погоре наведени“** и **„ниту една од наведените“**. Доколку се користи алтернативата „**сите погоре наведени**“ како точен одговор, испитаникот кој може да идентификува повеќе од една алтернатива како точна, може да го селектира точниот одговор, иако не е сигурен за останатите алтернативи. Истата забелешка важи и доколку се искористи алтернативата **„ниту една од наведените“** како точен одговор. Односно, и во двата наведени случаи, испитаникот може да примени делумно знаење за да дојде до точниот одговор.
- Алтернативите треба да бидат дадени по логички редослед (на пример, азбучен или нумерички редослед), со цел да се избегне пристрасноста кон одредени позиции.
- Бројот на алтернативите може да варира меѓу прашањата со повеќекратен избор, сè додека сите алтернативи се уверливи. Не постојат силни докази кои потврдуваат значајни разлики во тешкотијата и веродостојноста на резултатите од тестот меѓу прашањата кои содржат два, три или четири дистрактори.

4.1.2. Стандардни протоколи за поствалидација на прашањата со повеќекратен избор

Поствалидацијата помага да се идентификуваат прашањата со повеќекратен избор со дискутабилна валидност, за да може истите да се подложат на соодветно модифицирање, пред повторно да се искористат или целосно да се отфрлат. Тоа се постигнува со задавање нумерички вредности на секој проблем и неговите алтернативи. Овие нумерички вредности се утврдени со: **индексот на тежина**, **индексот на дискриминација** и **ефективноста на дистракторите**[46]. Врз основа на прифатените стандардни граници за овие индекси, ППИ може да бидат прифатени, модифицирани или отфрлени.

Направените истражувања потврдуваат дека постои значителен простор за подобрување на квалитетот на многу тестови базирани на прашањата со повеќекратен избор. Анализирајќи примерок од 60 прашања со повеќекратен избор од медицинската област, *Ramakrishnan et al*. [201] заклучуваат дека нешто повеќе од третина од дистракторите не се функционални, односно не се прифатиливи. Истите треба да бидат модифицирани или заменети и повторно тестирани, сè додека не го задоволат критериумот за прифатливост (односно, да се постигне ефективност на дистракторите поголема или еднаква на 5%). *Halikar et al*. [202] детално анализираат 20 прашања од медицинската област и забележуваат дека сите ППИ имаат најмалку еден нефункционален дистрактор, додека вкупниот број на нефункционални дистрактори изнесува 23% од множеството дистрактори. Исто така, резултатите покажуваат дека процентот на прифатливи прашања, врз основа на индексот на тежина и индексот на дискриминација изнесува 35% и 50%, соодветно. Авторите препорачуваат генерирање на базен од валидни ППИ со познати вредности за индексите, од каде испитувачите може да изберат соодветни ППИ за одредено тестирање. Во своето истражување, *DiBattista et al*. [203] истакнуваат дека е неопходна обука и поддршка на оценувачите, со цел тие да бидат сигурни дека нивните тестови со прашања со повеќекратен избор се добро креирани и имаат прифатлива дискриминаторна моќ. Самиот процес на креирање на вакви висококвалитетни тестови е вештина која се учи [204].

4.2. Опис на тест-колекцијата за одговарање прашања на македонски јазик од областа филозофија

Прашањата со повеќекратен избор се инкорпорирани во државната матура во Р. Македонија, која претставува завршен дел на средното образование. Државната матура се полага според посебни испитни програми кои се темелат на целите на наставните програми за соодветните наставни предмети, вклучени во листата на државната матура. За да може да се обезбеди непристрасност во ова истражување, комплетното испитување е направено врз множество прашања со повеќекратен избор, преземени од државната матура[47] на Р. Македонија. Фокусот е ставен врз предметот филозофија (за четврта година гимназиско образование), бидејќи учебникот од оваа област е

[46] http://www.proftesting.com/test_topics/steps_9.php
[47] http://www.matura.gov.mk

единствениот достапен извор со дозвола за преземање [205].

Во истражувањето се вкучени прашањата со повеќекратен избор во периодот од 2010 година, заклучно со 2016 година, кои се јавно достапни на веб-страницата на Државниот испитен центар. Вкупниот број на такви прашања изнесува 313, и истите се подложени на анализа, за да се утврди дали ги задоволуваат стандардните протоколи за превалидација. Притоа, забележано е дека вкупно 82 прашања мора да бидат исклучени од тест-колекцијата заради нивната „несоодветност“. Тоа значи дека најголем дел од нив не ги задоволуваат наведените протоколи (додаток Б), мал дел прашања се повторуваат во различни испитни години, додека останатите се прашања чиј одговор не може да се најде во наведениот извор (учебникот по „Филозофија“). Останатите прашања (вкупно 231 прашање) се со четири понудени одговори, од кои само еден одговор е точен. При тоа, точниот одговор може да се пронајде во книгата „Филозофија“. Следува пример на прашање од тест-колекцијата на македонски јазик.

Аристотел бил основоположник на Универзитетот „Ликеј“ кој уште е нарекуван:

А. Киренска школа
Б. Атинска академија
В. Киничка школа
Г. Перипатетичката школа (точен одговор)

Во табелата 6 е даден преглед на распределбата на прашањата во согласност со нивната категорија. Анализата на множеството прашања иницира дефинирање на нова таксономија на прашања, соодветна за тест-колекцијата на македонски јазик. Таксономијата применува двослојна хиерархија: 3 груби категории (фактовидни прашања, описни прашања и прашања со набројување) и 7 фини категории (додатокот В содржи пример на прашање од секоја од седумте фини категории). Може да се забележи дека најзастапени се описните прашања, кои сочинуваат 56.28% од вкупното множество прашања, следени од фактовидните прашања со 38.96%, додека најмалку се застапени прашањата со набројување (4.76%). Прашањето дадено претходно како пример, припаѓа во групата фактовидни прашања (ентитет).

Груба категорија (ГК)	Фина категорија (ФК)	# во ФК	# во ГК	%
фактовидни прашања	личност	38	90	38.96
	ентитет	52		
описни прашања	опис	19	130	56.28
	исказ	83		
	дефиниција	28		
прашања со набројување	личност	7	11	4.76
	ентитет	4		
Вкупно		**231**	**231**	**100**

Табела 6. Распределба на прашањата по категории

4.2.1. Креирање на речникот од термини

Во согласност со надворешните својства, зборовите во македонскиот јазик се поделени во две различни групи: зборови кои се подложни на промени (како именките, придавките, заменките, броевите и глаголите), и зборови кои се неменливи (прилозите, предлозите, сврзниците, извиците, честиците и модалните зборови) [123]. Постигнувањата во другите јазици јасно потврдуваат дека различните зборовни групи имаат различно влијание во процесот на прибирање релевантни документи од одреден корпус [142]. Поради ова, но и поради сознанието дека ваков заклучок не е донесен што се однесува до прибирањето документи напишани на македонски јазик, дефинирањето на вистинското множество зборовни групи со соодветни тежини, кои би имале позитивно влијание во овој процес, е проблем вреден за истражување.

Во согласност со причините наведени во секцијата 3.2.3, ова истражување ги користи следниве пет зборовни групи: именките (*n*), глаголите (*v*), придавките (*a*), броевите (*m*) и прилозите (*r*). Останатите зборовни групи се исклучени уште во претпроцесирачката фаза, при креирање на речникот за колекцијата документи. За да се оцени влијанието на различните зборовни групи, како и зборoформите, при прибирањето информации на македонски јазик, креирани се два речника од термини[48], именувани како Речник_1 и Речник_2.

Ознака за зборовна група	Број	Значење на ознаката
np	86	лична именка без означен род
npm	766	лична именка од машки род
npf	131	лична именка од женски род
npn	5	лична именка од среден род
ncm	2329	општа именка од машки род
ncf	2599	општа именка од женски род
ncn	571	општа именка од среден род
ngn	894	глаголска именка од среден род
v	3597	глагол
r	641	прилог
a	5159	придавка
m$	515	број запишан со цифри
m(буква)	64	број запишан со букви
unknown	40	непознат збор
latin	244	латински збор
english	6	англиски збор
abb	19	скратеница
prefix	2	префикс
Вкупно	17688	

Табела 7. Распределба на термините од колекцијата документи по зборовни групи

[48] **Термин** е **тип** кој е вклучен во речникот на *IR*-системот [133].

Речник 1. При креирањето на Речник_1, прво се отстранети сите интерпукциски знаци од учебникот „Филозофија“. Утврдено е дека учебникот содржи вкупно 157193 **токени** (неуникатни зборови). Потоа, како термини за Речник_1 се искористени **типовите** (во суштина уникатните зборови) од петте избрани зборовни групи *n*, *v*, *a*, *m*, и *r*, чиј вкупен број изнесува 17668. Означувањето на зборовите во согласност со зборовната група е направено со примена на претходно спомнатиот анотиран речник за македонскиот јазик (секција 3.2.3). Овој речник содржи најголемиот дел од често користените зборови во македонскиот јазик. Зборовите кои не се вклучени во него, а се појавуваат во колекцијата (заради спецификите на терминологијата), се означени рачно. Ова особено се однесува на сопствените именки. Треба да се потенцира дека означувањето на зборовите со помош на речник, не го зема предвид контекстот во кој тие се појавуваат. Тоа значи дека зборовите кои имаат повеќе од една ознака, не може со сигурност да се означат со вистинската ознака. Ваквиот проблем во ова истражување е разрешен со вклучување на овие зборови само еднаш во Речник_1, и тоа со ознаката за зборовна група која има највисока тежина (равенки (4)). Во табелата 7 е прикажана распределбата на термините од Речникот_1, во согласност со нивната зборовна група. Како што може да се забележи, повеќе од 40% од термините во Речникот_1 припаѓаат во зборовната група именки.

Речник 2. Напорите за креирање на Речникот_2 се насочени кон определување на **лексемите** (множествата флективни форми на одреден збор) од термините од Речникот_1. Од секое множество случајно е избран член, кој ги претставува сите останати членови од тоа множество во колекцијата документи, но и во прашалниците при нивната обработка. Користејќи ја стратегијата опишана подолу, Речникот_2 се состои од вкупно 9927 лексеми.

Во отсуство на лематизатор (*lemmatizer*), како и стемер (*stemmer*), за македонскиот јазик, ова истражување применува статистички пристап со цел да ги групира зборовите од Речникот_1 кои припаѓаат во иста лексема. Првиот чекор во креирањето на Речникот_2 е примена на метриката за сличност на низи карактери која го имплементира *Dice*-коефициентот (равенка (16)), заедно со **кластерирањето со комплетна врска** (*complete-link clustering*) [206]. *Dice*-вредноста е определена со примена на биграмите од термините и притоа се утврдува дека два термина припаѓаат во иста лексема доколку имаат одеден број заеднички биграми. Во овој случај, дефиниран е праг од 0.5, што значи дека кластерирањето со комплетна врска се извршува сè додека постојат два термина од Речникот_1 чија *Dice*-вредност е помала од 0.5. Уште повеќе, креирањето на Речникот_2 ја следи следнава стратегија:

- Секоја сопствена именка, како и секој број запишан со цифри, претставува лексема сама за себе.
- Алгоритамот за кластерирање се извршува над множество термини кои припаѓаат на иста зборовна група и започнуваат со ист биграм. При тоа, општите именки се кластерираат само доколку припаѓаат на ист род (машки, женски или среден). Одлуката за кластерирање на ваквите множества термини произлезе од специфичноста на македонскиот јазик, каде зборoформите

припаѓаат во иста зборовна група (зборoформите на општите именки се од ист род) и се разликуваат само во неколку букви на крајот од низите карактери.

Анализата на дефинираните лексеми откри дека *Dice*-метриката не е соодветна за оваа задача (примери од генерираните кластери се наведени во додатокот Г). Во македонскиот јазик постојат парови зборови со комплетно различно значење, кои се разликуваат само во една буква во нивниот запис, како на пример зборовите: „продор“ и „прозор“. *Dice*-метриката за овие два збора дава вредност 0.2, па поставувајќи праг од 0.5 во прелиминарните експерименти, овие два збора се вклучуваат во иста лексема, што е непожелно. За правилно да се дефинира Речникот_2, неопходно е да се извршат корекции рачно. Со цел да се избегне овој долготраен процес, направено е дополнително истражување за наоѓање на најсоодветната метрика за групирање на зборoформите во македонскиот јазик, која би можела да се користи без рачно интервенирање. Резултатите од ова истражување се презентирани во секцијата 4.3.1.3.

Зборовна група	Зборови
именка	филозоф (машки род), филозофија (женски род), филозофијата (женски род)
глагол	филозофира, филозофираше
придавка	филозофски, филозофските

Табела 8. Распределба на седум зборови по зборовни групи

Во табелата 8 се дадени седум зборови од учебникот „Филозофија“, поставени во соодветната зборовна група. Секој од нив претставува посебен термин во Речникот_1. Од друга страна, овие седум зборови формираат четири термини во Речникот_2, бидејќи некои од нив припаѓаат на иста лексема (табела 9).

Речник	Број на термини	Термини
Речник_1	7	филозоф, филозофија, филозофијата, филозофира, филозофираше, филозофски филозофските
Речник_2	4	филозоф[49], филозофија[50], филозофира[51], филозофски[52]

Табела 9. Примерок од термини од двата речника дефинирани од седум зборови (филозоф, филозофија, филозофијата, филозофира, филозофираше, филозофски и филозофските)

4.2.1.1. Речник од зборови кои имаат ист стем

Во своето истражување, *Jovanovska et al*. [178] ја испитуваат успешноста на системот за прибирање информации (и одговарање прашања) со примена и на Речник_3, кој содржи групи од зборови кои споделуваат ист основен збор (стем). Неговото креирање е направено рачно преку здружување на одредени групи од зборови од Речникот_2. Во тој речник, на пример, зборовите од табелата 8 претставуваат еден термин (табела 10).

[49] Претставник на кластерот {филозоф}
[50] Претставник на кластерот {филозофија, филозофијата}
[51] Претставник на кластерот {филозофира, филозофираше}
[52] Претставник на кластерот {филозофски, филозофските}

Речник	Број на термини	Термини
Речник_3	1	филозоф[53]

Табела 10. Примерок од термини од Речник_3 дефинирани од седум зборови (филозоф, филозофија, филозофијата, филозофира, филозофираше, филозофски и филозофските)

Сепак, во следните истражувања (чии акцент е ставен на дополнително подобрување на точноста на *QA*-системот) е исклучена примената на Речникот_3, бидејќи наведеното истражување потврди дека најдобри резултати во прибирањето информации (како и во процесот на одговарање прашања) се постигнуваат со примена на Речникот_2 (секција 3.3.2).

4.2.2. Стретегии за прибирањето пасуси

Во претпроцесирачката фаза, системот за одговарање прашања на македонски јазик го разделува учебникот „Филозофија“ на пасуси кои содржат 300 последователни **токени**. Притоа, искористено е преклопување на пасусите, со цел да се задржи контекстот на реченицитe кои во овој процес би станале случајно пресечени. Оваа постапка резултира со вкупно 1033 пасуси. Задржувајќи ги зборовите чија зборовна група е *n*, *v*, *a*, *m* или *r*, просечниот број на токени во пасусите е околу 157. За рангирање на пасусите (за секој речник) се применети следниве три стратегии:

- **Стратегија_1 - Моделот на векторски простор со косинус сличноста (како основа за споредба)**. Секој пасус (како и секој прашалник) е претставен како вектор од термини од речникот, додека важноста на терминот е определена во согласност со $tf - idf$ тежинската шема. Сликата М1 (додаток М) претставува дел од кодот за пресметување на сличноста меѓу два документа (односно, меѓу прашалник и документ) со помош на оваа стратегија. Пресметувањето на $tf - idf$ тежините на терминте од прашалникот (користејќи ги и зборовите од кластерот зборформи во кој припаѓа овој термин), е дадено на сликата М2.
- **Стратегија_2 - Совпаѓањето на координати**. Секој пасус (како и секој прашалник) е претставен како вектор од термини од речникот, додека важноста на терминот е 1 или 0 во зависност од тоа дали терминот се појавува во пасусот (прашалникот) или не. Сликата М3 претставува дел од кодот за пресметување на пресметување на сличноста меѓу прашалникот и документот со помош на оваа стратегија.
- **Стратегија_3 - Совпаѓањето на координати засилено со информацијата за зборовната група.** Секој пасус (како и секој прашалник) е претставен како вектор од термини од речникот, додека важноста на терминот е во согласност со неговата зборовна група. Вредностите кои се тестирани како тежини за зборовните групи се дадени во табелата 5.

[53] Претставник на кластерот {филозоф, филозофија, филозофијата, филозофира, филозофираше филозофски, филозофските}

4.3. Експериментални резултати

Во оваа секција се дадени резултатите добиени во процесот на прибирање пасуси со примена на двата креирани речника, како и резултатите добиени во последната фаза од *QA*-системот, односно утврдена е точноста на системот во селекцијата на точниот одговор од понудените четири одговори. Во делот за прибирање на пасусите, се дадени постигнувањата со примена на трите опишани стратегии. Исто така, се презентирани и резултатите од примената на метриките за сличност на зборови, со цел автоматски да се групираат зборoформите од Речникот_1, како и резултатите добиени со новодефинираната метрика базирана на триаголниот прозорец. Истите метрики се применети и за проширување на прашалникот со **сличен** збор, во случај да содржи збор кој го нема во речникот од колекцијата. Пристапите кои даваат најдобар резултат подоцна се вклучени и во последниот модул од *QA*-системот, чии резултати се презентирани на крајот од ова поглавје.

4.3.1. Имплементација на системот за одговарање прашања

Системот е имплементиран како *Python* библиотека изградена од повеќе пакети и модули со различни функционалности, кои заедно го градат целиот процес на пронаоѓање одговор на дадено прашање. *Python* е едноставен, но моќен програмски јазик со одлични функционалности за процесирање на линвистичките податоци. Се користи во многубројни научни истражувања, во индустријата, едукацијата, итн., бидејќи овозможува зголемување на продуктивноста, квалитетот и одржиливоста на кодот, притоа намалувајќи го времето за развој.

Со оглед на тоа што се работи со поголеми податоци над кои се вршат повеќекратни пресметки, се наметна потребата да се обрне поголемо внимание на мемориската и временската оптимизација на алгоритмите. Во контекст на ова, доста често користена податочна структура во библиотеката е хеш мапата (*dictionary* или *OrderedDict* во *Python*), која претставува колекција од клучеви и вредности и најчесто има константно време на пристап до бараниот елемент – клуч. Наспроти тоа, кај податочната структура листа (*list*) пребарувањето на елемент е во линеарно време, но ако се знае неговиот индекс, тогаш времето на пристап е константно. Овој факт исто така е искористен при имплементација на библиотеката.

Така, за алгоритмите во кои не е важен редоследот на појавување на зборовите, пасусите и прашањата се чуваат во меморија како листа од хеш мапи, во кои секој клуч е збор, а неговата соодветна вредност е бројот на појавувања на тој збор во пасусот. Во случај кога се користи кластерирањето, вредноста е вупниот број на појавувања на било кој збор од соодветниот кластер. Кога треба да се зачува и редоследот на појавување на зборовите во корпусот, наместо вредноста да биде вкупниот број на појавувања, таа е листа од позициите на кои тој збор се наоѓа. Ова е всушност имплементацијата на инвертираниот индекс.

Кластерите технички се листа од листи од зборови, но од погоре наведените причини, тие се трансформираат во хеш мапа каде клуч е секој збор, а вредност е *ID*-то на соодветниот кластер во кој тој збор припаѓа. Ова е ефикасно од мемориски аспект

со оглед на тоа што секој збор е зачуван само еднаш, а и го олеснува пресметувањето на $tf - idf$ и останатите алгоритми.

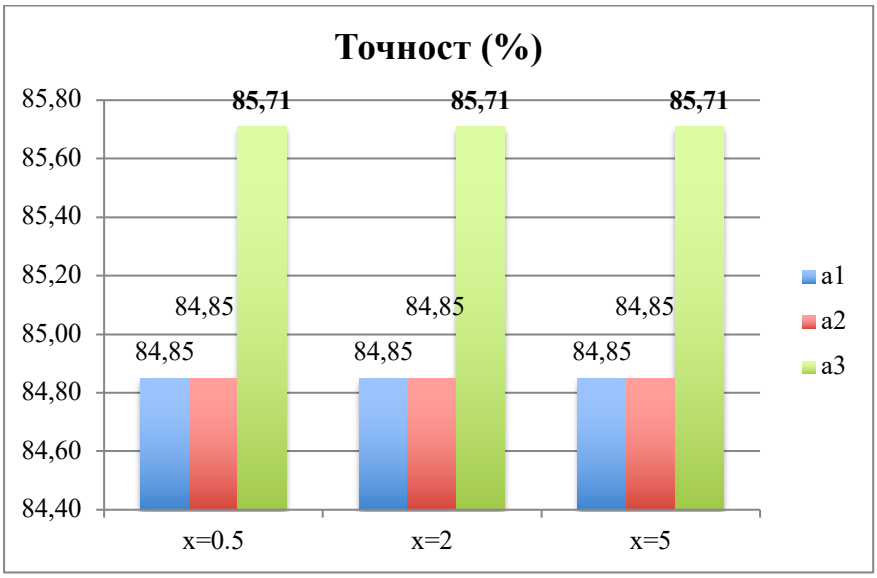
Делови од кодот со кои се реализирани најкарактеристичните функции на системот за одговарање прашања на македонски јазик, се дадени во додатокот М.

4.3.2. Прибирање пасуси

За секое прашање од тест-колекцијата рачно е определено множеството од пасуси кои го содржат неговиот одговор (означено со C). Дизајнираниот систем прибира пет најдобро рангирани пасуси за секој прашалник (прашање) кои најверојатно го содржат точниот одговор, плус сите пасуси кои имаат иста рангирачка вредност како и петтиот рангиран пасус (ова множество е означено со R). Како мерка за проценка во оваа фаза е искористена **точноста** (*accuracy*), со цел да се оцени за колку прашања од тест-колекцијата системот би можел да најде одговор во последната фаза. Се смета дека системот има погодок, доколку пресекот на C и R не е празно множество.

4.3.2.1. Прибирање пасуси со Речникот_1

Оваа секција ги содржи резултатите од процесот на прибирање пасуси применувајќи го Речникот_1. За првите две стратегии (моделот на векторски простор и совпаѓањето на координати) е добиена точност од 87.88% и 87.45%, соодветно (што претставува процент на прашања од тест-колекцијата за кои системот прибира точен пасус, односно пасус кој го содржи одговорот на прашањето).



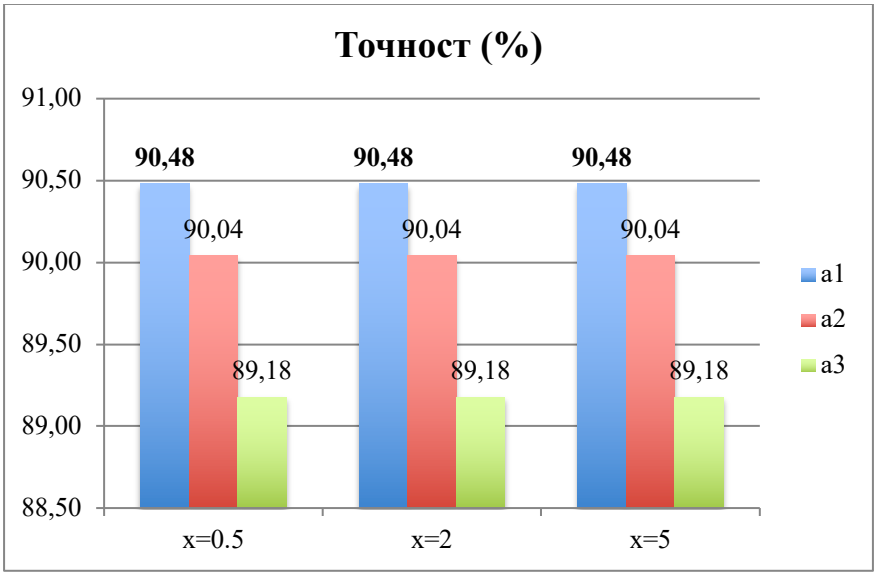
Слика 7. Две различни точности (во %) добиени вклучувајќи ги тежините за различните зборовни групи

За да се оцени влијанието на зборовната група во процесот на прибирање информации, за третата стратегија се тестирани различни тежини за зборовните групи, дадени со равенките (4) и табелата 5. Добиените резултати покажуваат дека за секоја вредност за x, за првите две соодветни вредности на a, се добива идентична точност од 84.85%, додека за третата соодветна вредност се добива точност од 85.71%. На сликата

7 се прикажани постигнатите резултати за три случајно избрани вредности на x. Постигнувањата добиени со Стратегијата_3 потврдуваат дека информацијата за зборовната група не е доминантна карактеристика сама за себе за оваа тест-колекција.

Комбинирање на информацијата за зборовна група со $tf - idf$ шемата за задавање тежини. Земајќи предвид дека резултатите од третата стратегија не се ветувачки, одлучено е да се комбинира информацијата за зборовна група на терминот со неговата фреквенција и инверзната документ фреквенција.

Ваквата комбинација ја потврди претпоставката дека зборовната група не е единствениот фактор за успешно прибирање информации. Имено, ваквата комбинација постигна највисока точност. Притоа, информацијата за зборовната група е вклучена само во векторот на прашалникот, и се тестирани истите вредности дадени со равенките (4) и табелата 5. Од друга страна, $tf - idf$ вредностите се искористени за претставување на векторите на пасусите. За секоја вредност на x и трите соодветни вредности за a, се добија точно три различни точности, 90.48%, 90.04% и 89.18%. На сликата 8 се прикажани постигнатите резултати за три случајно избрани вредности на x. Овие резултати укажуваат дека зголемувањето на разликата на тежините меѓу зборовните групи, поттикнува намалување на точноста на системот во процесот на прибирање пасуси.

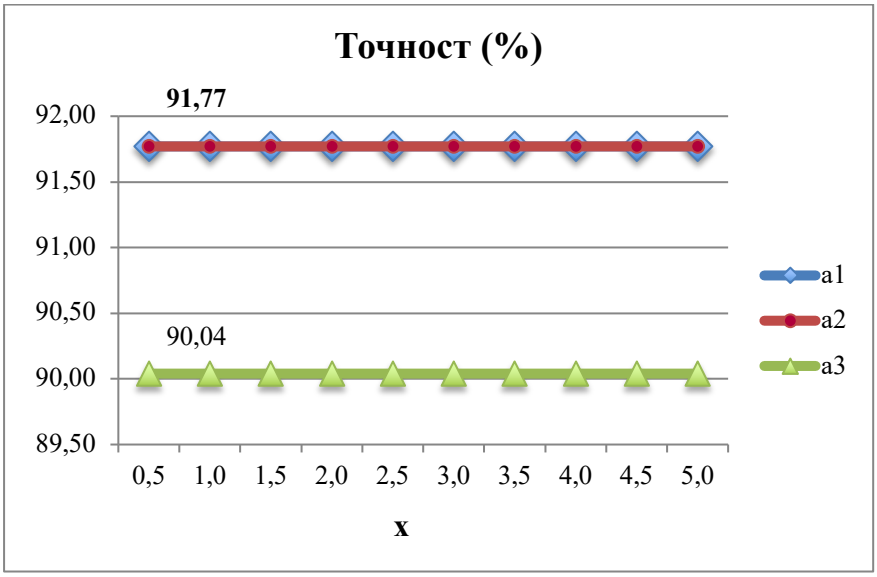


Слика 8. Три различни точности (во %) добиени комбинирајќи ги тежините за различните зборовни групи со $tf - idf$ вредностите

4.3.2.2. Прибирање пасуси со Речникот_2

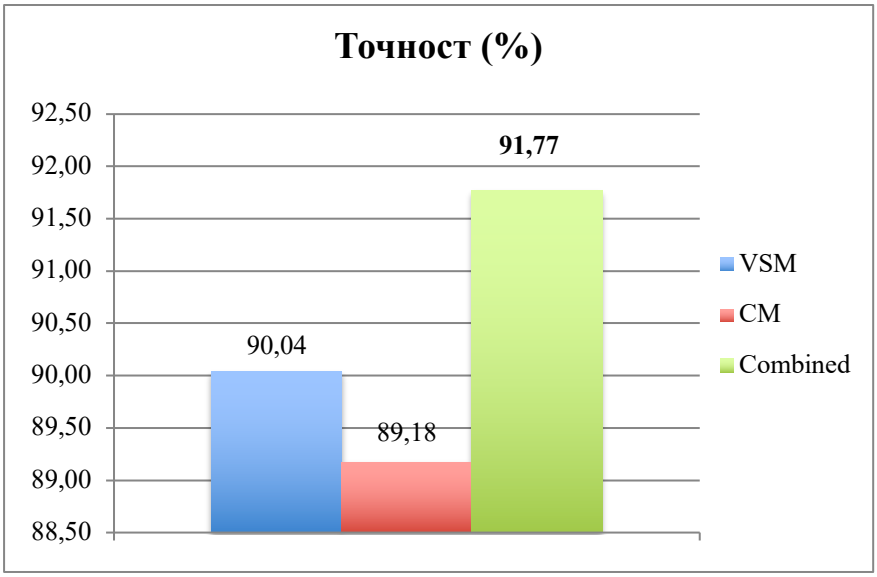
Во овој дел се дадени резултатите од процесот на прибирање пасуси користејќи го Речникот_2, односно со вклучување на зборoформите. За првите две стратегии (моделот на векторски простор и совпаѓањето на координати), постигната е точност од 90.04% и 89.18%, соодветно. Овие постигнувања повторно потврдуваат дека најдобар пристап за прибирање пасуси (документи) напишани на македонски јазик е вклучувањето на различните зборoформи на терминте од прашалникот (независно од моделот за рангирање). За третата стратегија, се тестирани истите вредности за

променливите x и a, дадени со равенките (4) и табелата 5. Највисоката точност постигната на овој начин е 83.98%.



Слика 9. Две различни точности (во %) добиени вклучувајќи ги тежините за различните зборовни групи

Од друга страна, комбинирањето на информацијата за зборовната група со $tf - idf$ шемата за задавање тежини, ја даде највисоката точност. Сликата 9 ги прикажува постигнатите точности при користење на различни вредности на x и трите соодветни вредности на a. Она што може да се забележи, е дека зголемувањето на разликата на тежините меѓу зборовните групи, продуцира намалување на точноста на системот. Подобри резултати се постигниваат за помали разлики (вакво однесување се забележува и при прибирањето пасуси без вклучување на зборoформите (слика 8)).



Слика 10. Постигнати точности (во %) со примена на зборoформите за три различни модели на прибирање

На сликата 10 е прикажан финалниот преглед на постигнатите резултати со моделот на векторски простор (со косинус сличност - *VCM*), совпаѓањето на

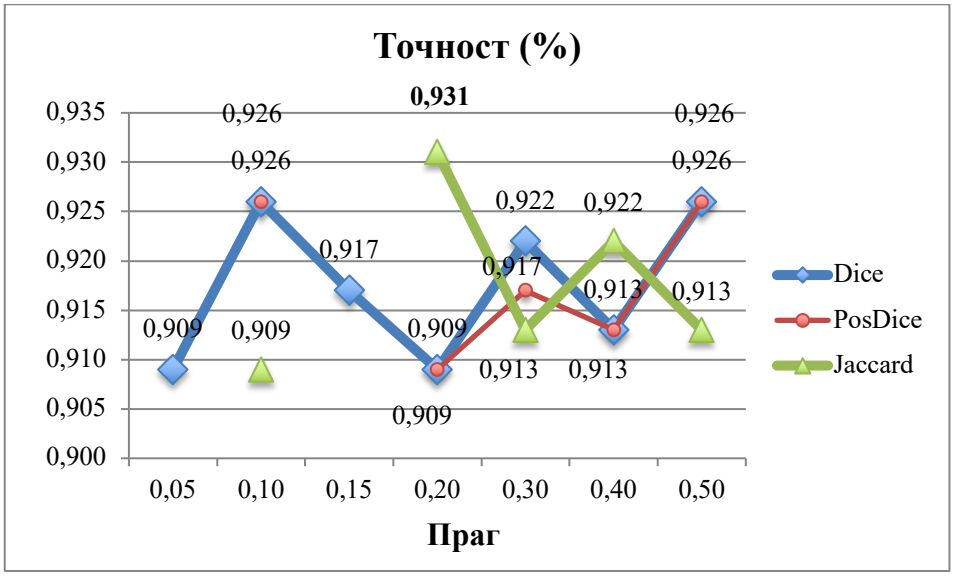
координати (*CM*) и комбинирањето на информацијата за зборовната група со $tf - idf$ шемата за задавање тежини (наречен **комбиниран метод** – *combined method*). Треба да се потенцира дека најдобар резултат во процесот на прибирањето пасуси се добива со примена на Речникот_2 и комбинираниот метод. Релативното подобрување во однос на **основата** (*baseline*) е 3.8949%.

Во сите наредни тестирања кои се направени со цел да се подобри точноста на прибирањето пасуси, е применет **комбинираниот метод**, односно земена е предвид информацијата за зборовната група на термините од прашалникот, и тоа за вредностите $x = 0.5$ и $a = 0.05$, како и $tf - idf$ шемата за задавање тежини на термините од пасусите.

4.3.2.3. Автоматско групирање на зборформите во македонскиот јазик

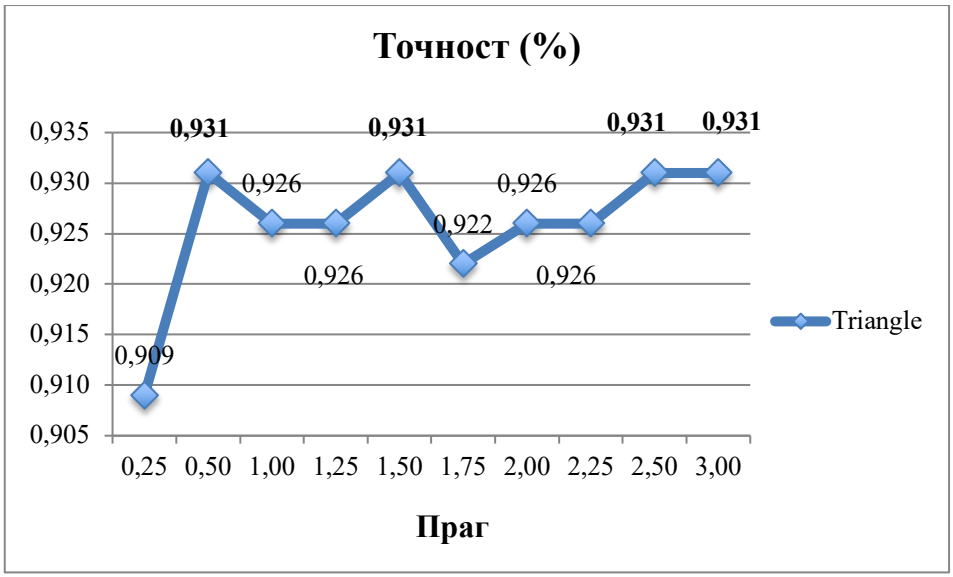
Заради согледувањата наведени во секцијата 3.3.2, ова истражување ги имплементира метриките за сличност на стрингови кои ги вклучуваат коефициентите *Dice*, *Positional Dice* и *Jaccard* (дадени со равенките (14), (15) и (16), соодветно), со цел автоматски да се групираат зборформите во македонскиот јазик. На сликите M4, M5 и M8 (додаток M) се дадени деловите од кодот кои се однесуваат на имплементацијата на овие метрики, соодветно, во дизајнираниот систем за одговарање прашања на македонски јазик.

Резултатите добиени од примената на генерираните групи во процесот на прибирањето пасуси (за различни вредности на прагот искористен во кластерирањето со комплетна врска) се дадени на сликата 11. Може да се забележи дека највисока точност се добива со примена на метриката која го вклучува *Jaccard*-коефициентот за праг 0.2 (93.10%), т.е., системот успешно прибира точни пасуси за 215 прашања од вкупно 231 прашање од тест-колекцијата. Останатите две метрики исто така даваат подобри резултати (за одредени прагови), отколку примената на точните кластери од зборформи.



Слика 11. Постигната точност (во %) во процесот на прибирање пасуси, со примена на три различни метрики, заради автоматско групирање на зборформите

Со цел да се утврди влијанието на групите од збороформи утврдени со новата метрика $dis_{Triangle}(X,Y)$ (равенка (20)) во процесот на прибирање пасуси, повторно се тестирани неколку вредности за прагот при кластерирањето. На сликата М9 (додаток М) е даден дел од кодот кој се однесува на имплементацијата на метрика базирана на триаголниот прозорец во *QA*-системот. Притоа, постигнатите резултати за различни вредности на прагот се прикажани на сликата 12. Може да се забележи дека за сите вредности на прагот (освен за вредноста 0.25) новата метрика за растојание дава подобар резултат, отколку примената на точните кластери од збороформи (дефинирани рачно). Системот успешно прибира точни пасуси за 215 прашања од вкупно 231 од тест-колекцијата, за следниве вредности на прагот: 0.50, 1.50, 2.50 и 3.00.



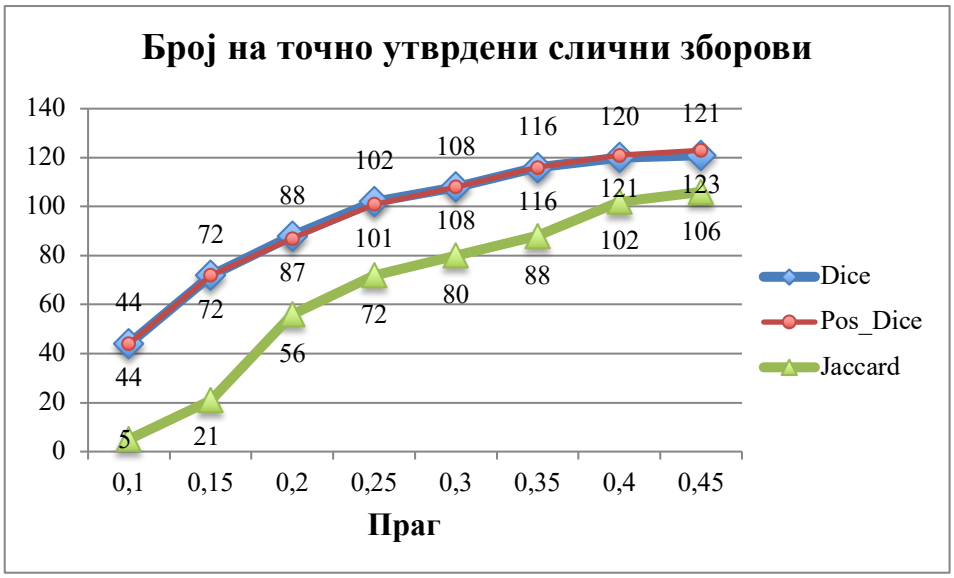
Слика 12. Постигната точност (во %) во процесот на прибирање пасуси, со примена на метриката за растојание базирана на триаголниот прозорец, заради автоматско групирање на збороформите

4.3.2.4. Проширување на прашалникот

Анализата на финалните резултати, добиени во процесот на прибирањето пасуси (со примена на точните кластери од збороформи), откри дека од вкупно 19 прашања од колекцијата за кои системот прибра погрешни пасуси, 38.89% се од категоријата фактовидни прашања, додека 61.11% се описни прашања. Со цел да се утврдат причините за неуспех, направена е подлабока анализа врз овие прашања. Како најзабележителен проблем се покажаа зборовите од прашањата кои не се вклучени во речникот зборови. Имено, 10 од вкупно 19 од овие прашања имаат клучни зборови кои не се застапени во Речникот_1 (примери на такви прашања се дадени во додатокот Д). До овој момент, системот ги исклучува овие зборови при прибирањето, што веројатно придонесува за зголемување на можноста за прибирање несоодветни пасуси за овие прашања. За разрешување на овој проблем, истражувањето е продлабочено со примена на метрики за сличност на стрингови за утврдување на **сличен** збор за збор од прашалникот кој го нема во Речникот_1 (во понатамошниот текст наречени **непознати зборови**). Во таа насока се имплементирани истите метрики (16), (17), (18) и (22) кои

ги вклучуваат коефициентите *Dice*, *Positional Dice* и *Jaccard*, соодветно, како и метриката базирана на триаголниот прозорец (за различни вредности на прагот).

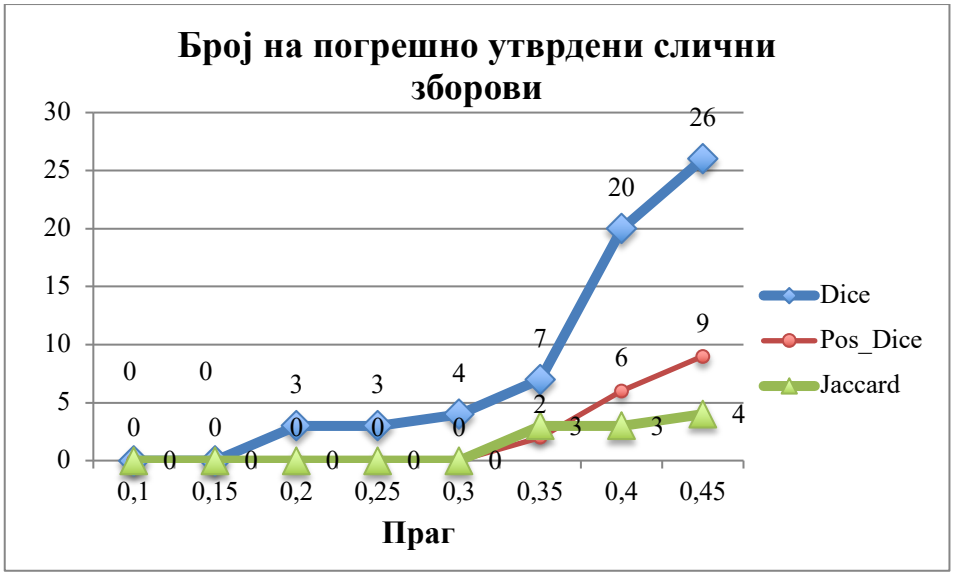
Проценката на успешноста на овие метрики за утврдување на сличен збор од Речникот_1 е направена врз множество од вкупно 165 зборови од прашањата од тест-колекцијата, кои не се појавуваат во учебникот „Филозофија“. Анализата на добиените резултати е направена рачно и емпириски. Притоа, се смета дека метриката правилно го проширува прашалникот со вклучување на **сличен** збор, доколку како **сличен** за непознат збор, утврдува збор кој споделува ист основен збор (стем) како и непознатиот.



Слика 13. Број на непознати зборови за кои точно е утврден сличен збор, со примена на метриките *Dice*, *Positional Dice* и *Jaccard*, за различни вредности на прагот

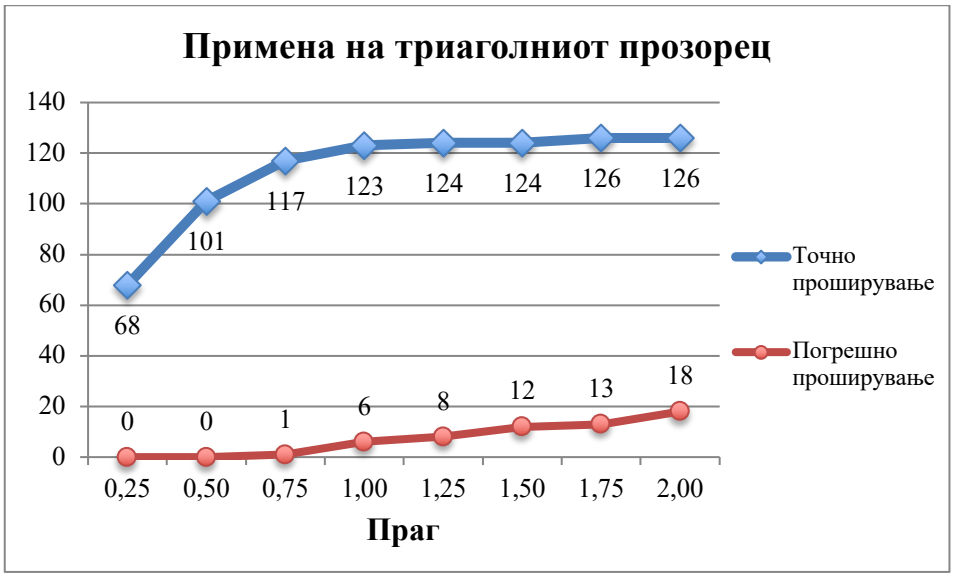
Мануелна анализа. Резултатите добиени со рачната анализа на успешноста на четирите наведени метрики се прикажани на сликите 13, 14 и 15. Сликите 13 и 14 претставуваат преглед на постигнувањата на метриките кои ги вклучуваат коефициентите *Dice*, *Positional Dice* и *Jaccard*, додека сликата 15 ги прикажува постигнувањата на метриката базирана на триаголниот прозорец, за различни вредности на праговите. Во додатокот Ѓ се дадени неколку примери од непознати зборови, заедно со најсличниот збор утврден со секоја од овие четири метрики.

Од графиците дадени на сликите 13 и 14, може да се забележи дека за поголеми вредности на прагот од 0.35 за *Dice*-метриката, зголемувањето на бројот на точно утврдени слични зборови е незначително, во однос на зголемувањето на бројот на погрешно утврдени слични зборови. Мануелната анализа укажува дека веројатно за тој праг оваа метрика би постигнала највисока точност. Од друга страна, *Positional Dice* и *Jaccard* се значително постабилни во бројот на зголемување на погрешно утврдените слични зборови, со зголемување на вредноста на прагот. Во согласност со сликите 13 и 14, се претпоставува дека *Positional Dice*-метриката би требало да даде најдобар резултат, за вредност на прагот поголема или еднаква на 0.35.



Слика 14. Број на непознати зборови за кои погрешно е утврден сличен збор, со примена на метриките *Dice*, *Positional Dice* и *Jaccard*, за различни вредности на прагот

Што се однесува до примената на метриката базирана на триаголниот прозорец за проширување на прашалникот, од сликата 15 може да се заклучи дека веројатно највисока точност во процесот на прибирањето пасуси би се постигнала за вредност на прагот околу 1.50.



Слика 15. Број на непознати зборови за кои точно/погрешно е утврден сличен збор, со примена на метриката базирана на триаголниот прозорец, за различни вредности на прагот

Емпириска анализа. За да се потврдат (или отфрлат) претходно наведените претпоставки, влијанието на проширувањето на прашалникот со сличен збор со примена на четирите метрики е оценето и експериментално, преку негово вклучување во самиот процес на прибирање пасуси. Во оваа фаза се искористени точните кластери од зборoформи, како и автоматски генерираните кластери со помош на метриката базирана на триаголниот прозорец за праговите 0.5, 1.5 и 2.5, за кои е добиена највисока точност во процесот на прибирањето пасуси (слика 12). Резултатите добиени

со примена на коефициентите *Dice*, *Positional Dice* и *Jaccard* се дадени во додатокот Е. Притоа, може да се забележи дека трите метрики ја постигнуваат својата највисока точност, доколку за проширување на прашалникот се применат кластерите генерирани автоматски со метриката базирана на триаголниот прозорец за прагот 1.5. Највисоката точност од 93.9% се постигнува со метриката *Positional Dice*, што значи дека системот успева да прибере точен пасус за 217 од вкупно 231 прашање од тест-колекцијата. Табелите Е1, Е2 и Е3 (додаток Е), ги потврдуваат забелешките дадени при рачната анализа за постигнувањата на овие три метрики.

***Triangle*_Праг**	**Точни кластери**	***Triangle*_0.5**	***Triangle*_1.5**	***Triangle*_2.5**
0.25	0.922	0.931	0.931	0.931
0.50	0.926	0.935	0.935	0.926
0.75	0.926	0.935	0.935	0.926
1.00	0.931	0.935	**0.939**	0.931
1.25	0.931	0.935	**0.939**	0.931
1.50	0.931	0.935	**0.939**	0.931
1.75	0.931	0.935	**0.939**	0.931
2.00	0.931	0.935	**0.939**	0.931

Табела 11. Постигната точност во фазата за прибирање пасуси, со примена на точните кластери и кластерите генерирани со метриката базирана на триаголниот прозорец за праговите 0.5, 1.5 и 2.5, и проширување на прашалникот со истата метрика

Во табелата 11 се прикажани постигнувањата во процесот на прибирањето пасуси со искористување на метриката базирана на триаголниот прозорец за проширување на прашалникот (дел од кодот за нејзината имплементацијата е даден на сликата М10). Како што може да се забележи, највисока точност во процесот на прибирањето пасуси се добива со примена на оваа метрика за проширување на прашалникот (за повеќе вредности на прагот), со вклучување на кластерите од зборформи генерирани автоматски со истата метрика за прагот 1.5. Постигнатата точност изнесува 93.9%, односно системот успева да прибере точен пасус за 217 од вкупно 231 прашање од тест-колекцијата „Филозофија“.

4.3.3. Примена на *Hanning*-прозорецот за утврдување на точниот одговор

Во оваа секција се дадени резултатите добиени од примената на *Hanning*-прозорската функција во последната фаза од *QA*-системот, односно во фазата за селекција на точениот одговор од четирите понудени одговори за секое прашање. Делот од кодот со имплементација на оваа функција е даден на сликата М11.

За големина на прозорецот се тестирани следниве вредности:

$$w \in \{4, 6, 10, 16, 20, 26\}.$$

Во табелата 12 се дадени резултатите добиени со примена на *Hanning*-прозорецот врз точните пасуси (односно пасусите кои го содржат точниот одговор на прашањата), како и врз петте најдобро рангирани пасуси во фазата за прибирање пасуси од *QA*-системот. Притоа, овие пасуси се добиени со вклучување на кластерите од зборформи генерирани со новата метрика базирана на триаголниот прозорец со праг 1.5, како и проширувањето на прашалникот со примена на истата метрика со праг 1.5.

Треба да се потенцира дека резултатите дадени во табелата 12 од примената на *Hanning*-прозорската функција се добиени без користење на зборформите и без проширување на прашаликот, при креирање на четирите прашалници кои се користат за утврдување на точниот одговор (секција 3.4.2.1). Резултатите добиени со примена на пасусите прибрани во *IR*-фазата со вклучување на кластерите со зборформи генерирани со метрика базирана на триаголниот прозорец со прагови 0.5 и 2.5, и проширување на прашаликот со примена на истата метрика со праг 1.5 (за кои е постигната највисока точност при прибирањето пасуси, заедно со прагот 1.5), се дадени во додатокот Ж.

w	**Точни пасуси**		**Врз пасусите добиени во *IR* со примена на кластери од зборформи генерирани со *Triangle*_1.5 и проширување на прашалнник со *Triangle*_1.5**	
	# на точно одговорени	*QA*-точност (%)	# на точно одговорени	*QA*-точност (%)
4	187	80.95	167	72.29
6	197	85.28	177	76.62
10	**203**	**87.88**	**181**	**78.35**
16	202	87.45	180	77.92
20	201	87.01	180	77.92
26	198	85.71	177	76.62

Табела 12. Постигната точност во фазата за селекција на точниот одговор, со примена на *Hanning*-прозорецот, без зборформи и без проширување на прашаликот

Табелата 12 потврдува дека зголемувањето и намалувањето на бројот на точно одговорени прашања при зголемување на вредноста на *w*, го запазува трендот доколку наместо точните пасуси се искористат пасусите добиени со новата метрика искористена за групирање на зборформите и проширување на прашаликот со сличен збор за збор од прашаликот кој не се содржи во речникот.

w	**Врз пасусите добиени во *IR* со примена на кластери од зборформи генерирани со *Triangle*_1.5 и проширување на прашалнник со *Triangle*_1.5**	
	# на точно одговорени	*QA*-точност (%)
4	168	72.73
6	185	80.52
10	190	82.25
16	188	81.39
20	**192**	**83.12**
26	187	80.95

Табела 13. Постигната точност во фазата за селекција на точниот одговор, со примена на *Hanning*-прозорецот, заедно со зборформи и проширување на прашаликот

Во табелата 13 се дадени постигнувањата на *Hanning*-прозорската функција, доколку во последната *QA*-фаза се земат во предвид и зборформите на зборовите од прашаликот и се направи негово проширување во случај на непознат збор. Петте најдобро рангирани пасуси од *IR*-фазата врз кои се применува *Hanning*-прозорецот, се

оние добиени со вклучување на кластерите од збороформи генерирани со метриката базирана на триаголниот прозорец со праг 1.5, како и проширување на прашалникот со примена на истата метрика со праг 1.5.

4.4. Дискусија и препораки

Во оваа секција се сумирани и анализирани резултатите добиени од примената на дизајнираниот систем за одговарање прашања, врз опишаната тест-колекција на македонски јазик од областа филозофија. Максималната точност на системот за оваа колекција изнесува **83.12%**, постигната со имплементација на следниов пристап:

– Во фазата за прибирање пасуси се применети:

- **комбинираниот метод**, односно земена е предвид информацијата за зборовната група на термините од прашалникот, и тоа за вредностите $x = 0.5$ и $a = 0.05$ (равенки (4)), како и $tf - idf$ шемата за задавање тежини за термините од пасусите,
- кластерите од збороформи добиени со метриката за сличност на стрингови базирана на **триаголниот прозорец** за праг 1.5, и
- проширувањето на прашалникот со сличен збор, доколку содржи збор кој не се појавува во речникот од колекцијата, со вклучување на метриката базирана на **триаголниот прозорец** за праг 1.5.

– Во последната фаза за селекција на точниот одговор се применети:

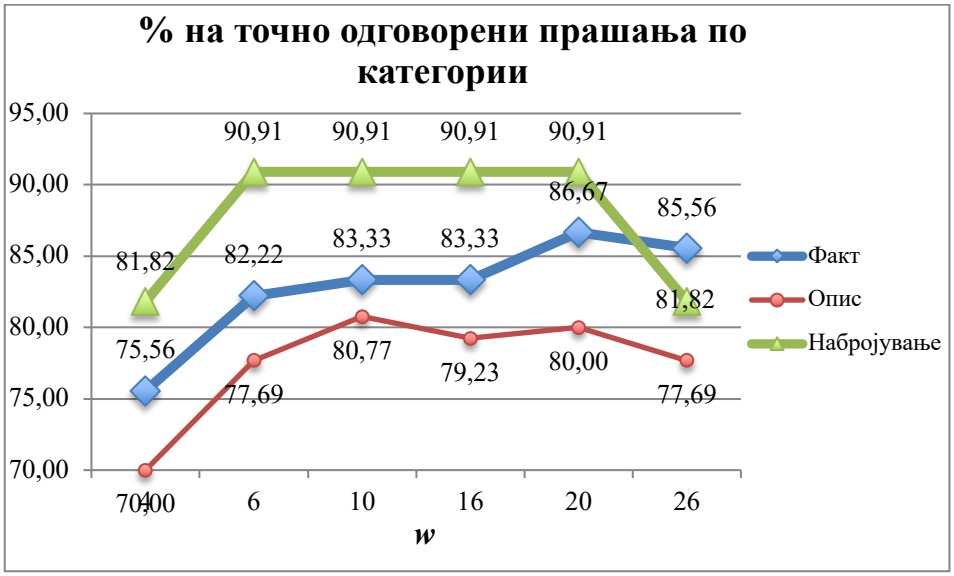
- *Hanning*-прозорската функција за големина на прозорецот $w = 20$,
- кластерите од збороформи добиени со метриката за сличност на стрингови базирана на **триаголниот прозорец** за праг 1.5, и
- проширувањето на прашалникот со сличен збор, доколку содржи збор кој не се појавува во речникот од колекцијата, со вклучување на метриката базирана на **триаголниот прозорец** за праг 1.5.

Добиените резултати се детално анализирани, за да се оценат можностите за идно подобрување на точноста на системот. Во таа насока, направен е обид да се утврди големината на прозорецот w со која се добива највисока точност во одговарањето на одредена категорија прашања. Во тој случај, доколку системот во иднина се надгради со модул за класификација на прашањето, тогаш последниот модул адаптивно би можел да ги користи различните големини на w. Процентот на точно одговорени прашања од системот по категории, за различни вредности на w, е прикажан на сликата 16[54].

Анализата на резултатите прикажани на сликата 16 укажува на мошне интересно сознание во врска со фактовидните прашања. Имено, додека е очекувано да се зголемува бројот на точно одговорени описни прашања со зголемување на прозорецот, таквото однесување е особено забележливо и за прашањата кои припаѓаат

[54] Треба да се потенцира дека прашањата од категоријата **прашања со набројување**, се значително помалку застапени во однос на останатите две категории (табела 6).

во категоријата фактовидни прашања. При тоа, највисока точност за фактовидните прашања се добива за големина на прозорецот $w = 20$. За описните прашања, највисока точност се добива за $w = 10$. За оваа големина на прозорецот, системот успева да одговори едно описно прашање повеќе, отколку големината $w = 20$. Заради тоа, може да се заклучи дека дизајнираниот *QA*-систем успешно може да се користи за одговарање на прашањата (од тест-колекцијата по филозофија) од сите категории, со вредност на големината на прозорецот $w = 20$.



Слика 16. Процент на точно одговорени прашања по категории, за различни вредности на големината на прозорецот *w*

5. Компаративна анализа

„Многу често, повеќе може да се научи од неочекуваните прашања на детето, отколку разговорот со возрасен.“

John Locke

Изградениот систем во ова истражување, наменет пред сè за одговарање прашања поставени на македонски јазик, е искористен и за одговарање на прашањата од две дополнителни тест-колекции. Колекциите се исклучително блиски во однос на темите кои ги опфаќаат, но се напишани на два различни јазика: македонски и англиски. Во нив се вклучени документи и прашања со повеќекратен избор од областа на информатичките технологии. Испитувањето кое следи е направено со цел да се утврдат можностите за примена на дизајнираниот систем врз колекции од друга област, и да се оцени дали тоа побарува одредени надополнувања. Исто така, една од клучните цели е споредба на успешноста на системот доколку истиот се искористи за одговарање прашања на македонски и англиски јазик.

5.1. Опис на тест-колекциите на македонски и англиски јазик

Оваа секција дава опис на македонската и англиската тест-колекција од областа Информатички технологии (во натамошниот текст означени како МакИнфо и АнгИнфо, соодветно), чии прашања со повеќекратен избор се искористени за автоматско одговарање со системот за одговарање прашања, дизајниран во ова истражување.

Македонската тест-колекција содржи четири докуметни, насловени како: „Историјат на сметачите“ (47 прашања), „Хардвет“ (68 прашања), „Софтвер“ (18 прашања) и „Воведни поими“ (23 прашања). Документите од англиската колекција се: „Историја на сметачите“ (*History of computers*), „Компјутерски генерации“ (*Computer generation*), „Класификација на компјутерите“ (*Computer classification*), „Компјутерски хардвер“ (*Computer hardware*), „Компјутерски софтвер“ (*Computer software*), и „Компјутерски систем“ (*Computer system*). Вкупниот број на неединствени зборови (**токени**) во секој документ од двете колекции е даден во табелите 14 и 15.

Документ	# на токени во МакИнфо
Историјат на сметачите	4647
Хардвер	4159
Софтвер	1120
Воведни поими	2497
Вкупно	**12423**

Табела 14. Број на токени во документите од тест-колекцијата МакИнфо

Документ	# на токени во АнгИнфо
Историја на сметачите	3347
Компјутерски генерации	2729
Класификација на компјутерите	2022
Компјутерски хардвер	10940
Компјутерски софтвер	3891
Компјутерски систем	1590
Вкупно	**24519**

Табела 15. Број на токени во документите од тест-колекцијата АнгИнфо

Колекциите МакИнфо и АнгИнфо содржат 156, односно 90 прашања со повеќекратен избор, соодветно. Прашањата се извлечени од документите од соодветната колекција, и за секое прашање се понудени четири одговори, од кои само еден одговор е точен. Во согласност со дефинираната таксономија во секцијата 4.2, припадноста на прашањата во соодветната категорија за двете колекции, е дадена во табелата 16. Додатокот З и S содржат примери на прашања од секоја категорија, од колекциите МакИнфо и АнгИнфо, соодветно.

Груба/Фина категорија		МакИнфо			АнгИнфо		
		#		%	#		%
фактовидни прашања	личност	5	47	30.13	5	51	56.67
	ентитет	42			46		
описни прашања	опис	47	87	55.77	23	38	42.22
	исказ	0			0		
	дефиниција	40			15		
прашања со набројување	личност	1	22	14.10	0	1	1.11
	ентитет	21			1		
Вкупно:		**156**		**100**	**90**		**100**

Табела 16. Распределба на прашањата од колекциите МакИнфо и АнгИнфо, во согласност со нивната категорија

Креирање на Речникот_1. Креирањето на Речникот_1 за двете колекции (именувани Речник_1_МакИнфо и Речник_1_АнгИнфо) ја следи постапката опишана во секцијата 4.2.1. Прво, системот ги отстранува сите интерпункциски знаци од документите и врши означување на зборовната група за секој збор.

За означување на зборовите од колекцијата МакИнфо, применет е веќе спомнатиот анотиран речник за македонски јазик (секција 4.2.1). Во согласност со пристапот за креирање на Речникот_1 за тест-колекцијата „Филозофија“, Речникот_1_МакИнфо ги содржи само именките, глаголите, придавките, броевите и прилозите (означени како: *n*, *v*, *a*, *m*, и *r*). Зборовите кои не се вклучени во анотираниот речник, а се појавуваат во МакИнфо, повторно се означени рачно (особено сопствените именки) [178]. Исто така, бидејќи се користи речник за означување на зборовните групи, зборовите кои припаѓаат во неколку групи, се појавуваат само еднаш во Речникот_1_МакИнфо, и тоа со ознаката за зборовна група која има највисока тежина (равенки (4)). Колекцијата МакИнфо содржи и англиски зборови заради специфичноста на терминологијата. Во табелата И1 (додаток И) е дадена распределбата на англиските зборови по зборовни групи, означени со примена на речник за англискиот јазик.

За означување на зборовите од англиската колекција искористен е отворениот означувач на зборовни групи на *Stanford*[55]. Токму поради примената на креиран означувач, Речникот_1_АнгИнфо содржи и зборови кои имаат повеќе од една ознака. Анализата на ознаките на зборовите од Речникот_1_АнгИнфо потврди дека од вкупно 3539 единствени зборови, 170 зборови имаат повеќе од една ознака за зборовна група.

[55] https://nlp.stanford.edu/software/tagger.shtml

Распределбата на овие зборови по зборовни групи е дадена во табелата J1 (додаток J). Во процесот на обработка на прашалникот за прибирање пасуси (и селекција на точниот одговор), системот повторно го искористува означувачот на *Stanford*, со што се дефинира вистинската ознака на зборот, и во согласност со тоа се задава соодветна тежина (равенки (4)).

Ознака за зборовна група / Поспецифична ознака		**# во МакИнфо**		**# во АнгИнфо**	
np (лична именка)	np (лична именка без означен род)	3	**35**	**254**	
	npm (лична именка од машки род)	22			
	npf (лична именка од женски род)	10			
	npn (лична именка од среден род)	0			
n (општа именка)	ncm (општа именка од машки род)	514	**1212**	**1503**	
	ncf (општа именка од женски род)	487			
	ncn (општа именка од среден род)	133			
	ngn (глаголска именка од среден род)	78			
v (глагол)		**516**		**831**	
r (прилог)		**142**		**166**	
a (придавка)		**916**		**582**	
m (број)	m$ (број запишан со цифри)	169	**211**	187	**202**
	m(буква) (број запишан со букви)	42		15	
abb (скратеница)		**2**		**1**	
unknown (непознат збор)		**35**		/	
latin (латински збор)		**0**		/	
english (англиски збор во МакИнфо колекцијата)		**617**		/	
prefix (префикс)		**1**		/	
Вкупно:		**3687**		**3539**	

Табела 17. Распределба на термините од колекциите МакИнфо и АнгИнфо по зборовни групи

Во табелата 17 е даден детален преглед на распределбата на термините од Речникот_1_МакИнфо и Речникот_1_АнгИнфо, во согласност со нивната зборовна група. Иако колекцијата МакИнфо содржи скоро дупло помалку токени од колекцијата АнгИнфо (табела 14), табелата 17 покажува дека Речникот_1_МакИнфо содржи нешто повеќе термини од Речникот_1_АнгИнфо. Ова го потврдува фактот дека македонскиот јазик е побогат со зборформи во однос на англискиот јазик.

Креирање на Речник_2. Со цел да се креираат речниците кои содржат групи од зборформи, за двете колекции, искористен е пристапот кој даде најдобар резултат за тест-колекцијата „Филозофија“. Станува збор за примена на новодефинираната метрика за сличност на стрингови базирана на триаголниот прозорец, и тоа за праг 1.5. Речниците за македонската и англиската колекција се означени како Речник_2_МакИнфо и Речник_2_АнгИнфо, соодветно.

5.2. Експериментални резултати

Во оваа секција се презентирани резултатите добиени од процесот на одговарање прашања за двете опишани колекции, МакИнфо и АнгИнфо. Притоа, искористен е пристапот кој даде највисока точност за одговарање на прашањата од тест-колекцијата „Филозофија“ (секција 4.4) .

Колекција АнгИнфо. За колекцијата АнгИнфо се испитани следниве вредности за големина на прозорецот $w = \{4, 6, 10, 16, 20, 26, 30, 36, 46, 60\}$. Највисоката добиена точност изнесува **51.11%**, односно системот успева точно да одговори 46 од вкупно 90 прашања од колекцијата и тоа за две вредности на w, 36 и 46. Со цел да се постигне повисока точност, направена е детална анализа на целиот процес и добиените резултати. Анализата покажа дека причината за оваа „незадоволителна“ точност е малиот број термини во Речникот_1_АнгИнфо. Тоа придонесува исти термини често да се појавуваат во рамките на еден прозорец. До овој момент, **густината на распределба** (*DD*) базирана на *Hanning*-прозорската функција (секција 3.4.2.1) не ја зема предвид фреквенцијата на терминте кои се појавуваат во прозорецот. За разрешување на овој проблем, направена е нејзина модификација. Имено, за зборовите кои се повторуваат во рамките на еден прозорец, во DD сумата се вклучува само едно појавување и тоа она за кое *Hanning*-прозорската функција дава највисока вредност. Во додатокот К е даден пример на прашање со соодветен прозорец, каде зборовите од прашалникот се појавуваат повеќе пати, што од друга страна резултира со селекција на погрешен одговор како точен.

Направената модификација резултира со значително подобрување на точноста на *QA*-системот. Табелата 18 дава преглед на вкупниот број точно одговорени прашања од системот, за повеќе вредности на големината на прозорецот w. Како што може да се забележи од табелата, системот точно одговара **66.67%** од прашањата, и тоа за две вредности на w, **30** и **36**.

w	**Колекција АнгИнфо**	
	# на точно одговорени	*QA*-точност (%)
4	36	40.00
6	40	44.44
10	46	51.11
16	54	60.00
20	57	63.33
26	57	63.33
30	**60**	**66.67**
36	**60**	**66.67**
40	56	62.22

Табела 18. Постигната точност на *QA*-системот за колекцијата АнгИнфо, за различни вредности на прозорецот w

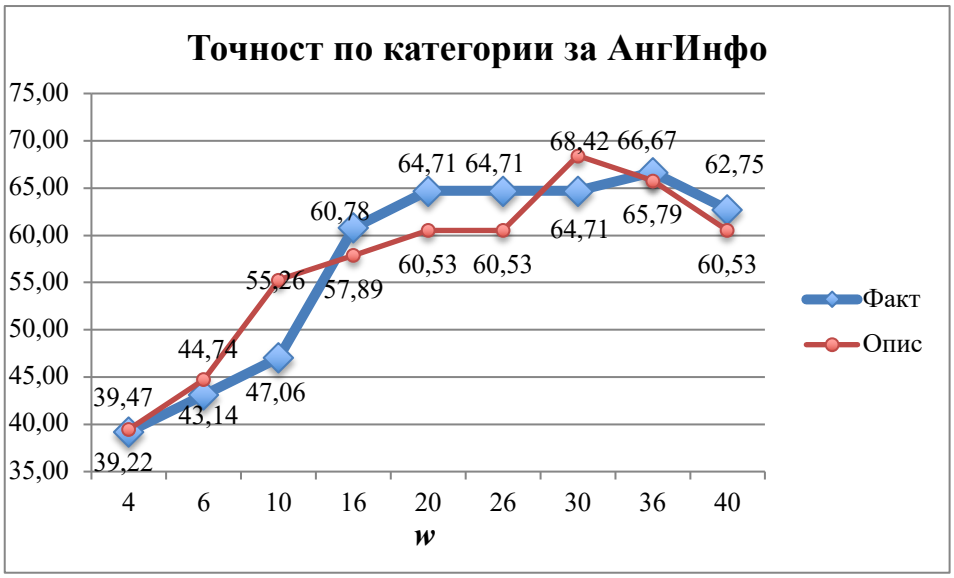
Колекција МакИнфо. Направената модификација на **густината на распределба** даде подобрување (иако значително помало) и на точноста при одговарање на прашањата од соодветната колекција на македонски јазик, МакИнфо. Релативното подобрување за оваа колекција изнесува 1.92%. Во табелата 19 се дадени вредностите за големина на прозорецот w, за кои е направено тестирање, како и соодветниот број на точно одговорени прашања од системот. Највисоката точност изнесува **61.54%**, и тоа за вредност **6** на големината на прозорецот w.

w	Колекција МакИнфо	
	# на точно одговорени	*QA*-точност (%)
4	92	58.97
6	**96**	**61.54**
10	93	59.62
16	93	59.62
20	92	58.97
26	93	59.62

Табела 19. Постигната точност на *QA*-системот за колекцијата МакИнфо, за различни вредности на прозорецот w

5.2.1. Детална анализа на добиените резултати

Добиените резултати за колекциите МакИнфо и АнгИнфо се детално анализирани, пред сè, за да се утврди која големина на прозорецот е најсоодветна за одговарање на прашањата од одредена категорија за секоја од двете колекции, и за да се направи споредба со заклучоците донесени за тест-колекцијата „Филозофија“. Исто така, направен е и обид да се утврдат главните причини за неуспех на системот при одговарање на прашањата од колекцијата МакИнфо, за која е постигната најмала точност.

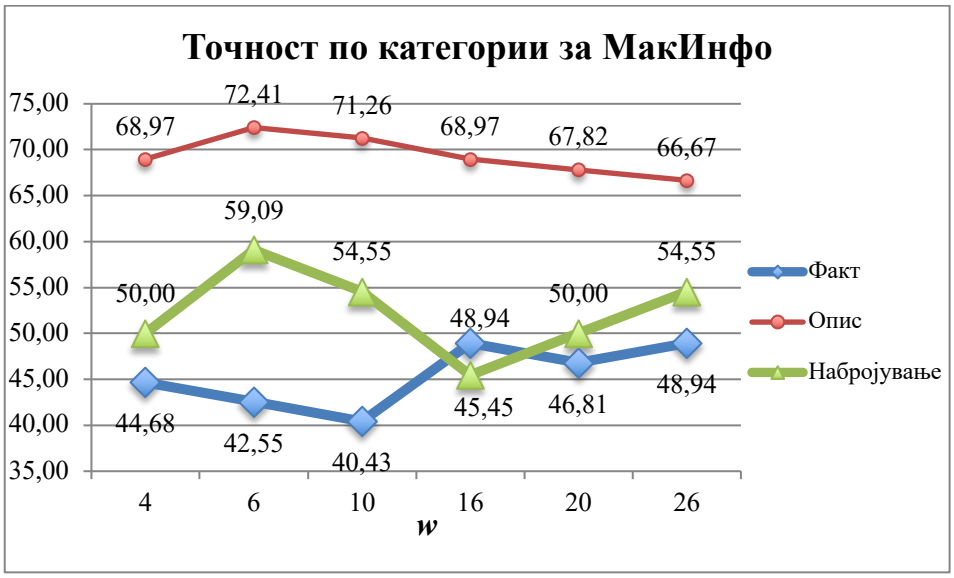


Слика 17. Процент на точно одговорени прашања за категориите фактовидни и описни прашања за колекцијата АнгИнфо, за различни вредности на големината на прозорецот w

Колекција АнгИнфо. Англиската колекција содржи само едно прашање со набројување, заради што не можат да се извлечат значајни заклучоци за оваа категорија. Токму затоа, на сликата 17 е даден процентот на точно одговорени прашања од системот само за останатите две категории: фактовидни и описни прашања (за различните вредности на w).

Анализата прикажана на сликата 17 ја потврдува забелешката дадена за тест-колекцијата „Филозофија“, а која се однесува на одговарањето прашања од категоријата фактовидни прашања. И за колекцијата АнгИнф, зголемувањето на прозорецот w резултира со зголемување на бројот на точно одговорени прашања од оваа категорија. Трендот на растење важи за двете категории прашања, сè до големина на прозорецот еднаква на 30 зборови. Ова потврдува дека истата големина на прозорецот од 30 (или 36) зборови, може да се искористи за успешно одговарање на сите прашања од колекцијата, независно во која категоријата припаѓаат.

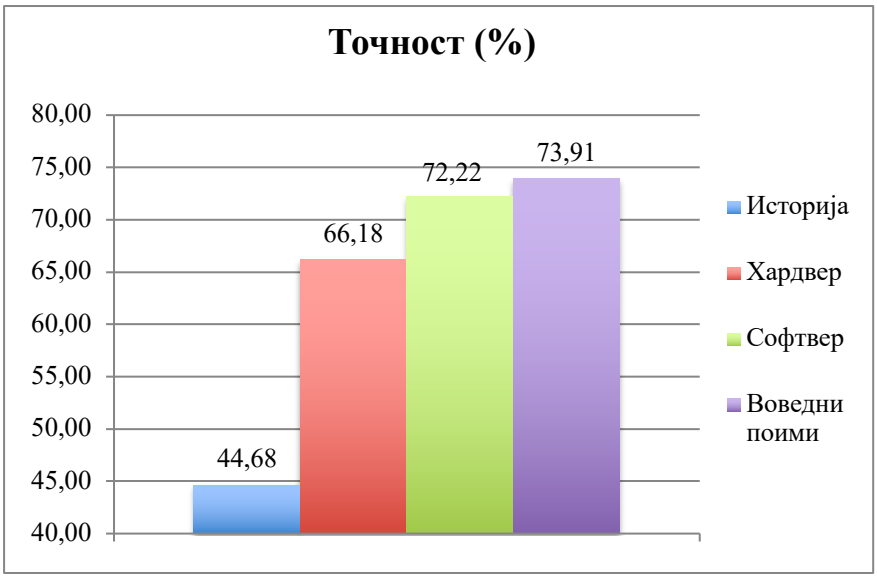
Колекција МакИнфо. Сликата 18 го дава процентот на точно одговорени прашања по категории за тест-колекцијата МакИнфо.



Слика 18. Процент на точно одговорени прашања по категории за колекцијата МакИнфо, за различни вредности на големината на прозорецот w

Она што се забележува од досега изложеното е дека *QA*-системот постигнува најмала точност за прашањата од колекцијата МакИнфо. Исто така, исклучително интересна забелешка за оваа колекција е дека процентот на точно одговорени описни прашања е значително поголем отколку точноста за останатите две категории. Токму ова однесување е причина детално да се анализираат сите 60 погрешно одговорени прашања за големина на прозорецот $w = 6$, за која системот постигнува највисока вкупна точност. Анализата открива неколку суштински проблеми кои влијаат на намалување на точноста во одговарањето прашања, и кои треба да се земат во предвид при одговарање на прашања од било која колекција. Меѓу нив, најзначајнни се следниве:

1. Дел од погрешно одговорените прашања (вкупно осум) побаруваат факт (личност) како одговор, со тоа што записот на името на личноста во документот (каде се наоѓа точниот одговор) е на македонски јазик, додека во одговорот е на англиски јазик (или обратно). Во додатокот Л (во делот 1) е наведен пример на такво прашање, заедно со сегментот од документот каде се наоѓа точниот одговор на тоа прашање. Седум од овие прашања се извлечени од документот „Историјат на сметачите“. Сликата 19 ја прикажува точноста на системот при одговарање на прашањата од четирите документи од колекцијата МакИнфо. Како што може да се забележи, за документот „Историјат на сметачите“ системот постигнува најниска точност.



Слика 19. Процент на точно одговорени прашања по документи за колекцијата МакИнфо

2. Дел од погрешно одговорените прашања побаруваат број како одговор. Различните можности за запишување на броевите е причина за неуспешно одговарање на дел од прашањата од тест-колекцијата. Во додатокот Л (во делот 2) е наведен пример на такво прашање, заедно со сегментот од документот каде се наоѓа точниот одговор на тоа прашање.

3. Постојат прашања во тест-колекцијата кои не ги задоволуваат стандардните протоколи за превалидација. Во додатокот Л (во делот 3) се наведени два примера на такви прашања. Првото прашање има коренот кој е искажан со негација, додека второто прашање содржи различни типови алтернативи (со набројување и описни), кои од друга страна, се со значително различна должина.

4. Дел од неодговорените прашања содржат само еден клучен збор, што најверојатно придонесува да се зголеми можноста за прибирање на погрешни пасуси. Пример на такво прашање е даден во додатокот Л (во делот 4). Еден од начините за резрешување на овој проблем е примена на **дистрибутивните методи** (заради отсуство на лексички извори како *WordNet* и *MeSH*), кои претпоставуваат дека значењето на зборот е поврзано со дистрибуцијата на зборовите околу него [8]. Дистрибутивните методи овозможуваат извлекување на синоними или други релации меѓу зборовите.

За потребите на своето истражување, *Jovanovska et al*. [178] имплементираат дистрибутивен метод кој би овозможил проширување на прашалниците што содржат само еден збор, со зборови кои често се појавуваат во неговата околина. Во таа насока, искористен е *Dice*-методот, како мерка за сличност на вектори (кои ги претставуваат зборовите од речникот), и $t-test$ статистиката за задавање тежини на компонентите на векторите (кои ги претставуваат останатите зборови од речникот) [8]. Зборовите кои имаат највисока сличност според *Dice*-методот, се смета дека се зборови кои често се појавуваат заедно во текстуални сегменти (*co-occurring words*). Вакавата комбинација се покажува како најуспешна за оваа задача [21]. За да се добијат значајни резултати, колекцијата МакИнфо е проширена со дополнителни документи од областа на информатичките технологии, со што вкупниот број на уникатни зборови (термини во Речникот_1) изнесува 9300. Во табелата Љ1 (додаток Љ) се дадени 10 зборови, за кои дистрибутивниот метод дава највисока сличност со друг збор. Исто така, дадени се останатите два збора кои имаат највисока сличност со дадениот збор.

Сепак, направените анализи потврдуваат дека за да се утврди поголем број значајни зборови кои се појавуваат заедно во текстуални сегметни (и кои би имале влијание во процесот на прибирање информации и одговарање прашања), неопходно е да се располага со значително поголема колекција од документи.

5. Анализата потврди дека за едно од неодговорените прашања, одговорот не може да се пронајде во документите од колекција (додаток Л, дел 5).

6. Заклучок и натамошна работа

„Сè се одгатнува, освен како да се живее.“

Jean-Paul Sartre

Технологијата за одговарање прашања сè уште е мошне активно поле за истражување, најмногу поради различните побарувања на реалните корисници [8]. Во насока на задоволување на овие барања, градењето модерен систем за одговарање прашања подразбира здружување на различни инженерски решенија од повеќе области, меѓу кои најзначајни се:

- **прибирањето информации** (*IR*) од текстуални документи или од бази со знаење,
- **обработката на природните јазици** (*NLP*), со различни техники како означувањето на зборовните групи (*tagging*), површното разложување (*chunking*), синтаксичкото разложување (*parsing*), препознавањето на именуваните ентитети (*named entity recognition*), разрешувањето на кореференците (*coreference resolution*), и други,
- **вештачката интелигенција** (*AI*), и
- **машинското учење** (*ML*).

Комбинирањето на техниките од овие области во насока на креирање „идеален“ систем за одговарање прашања, претставува исклучително предизвикувачка задача. Во овој контекст, треба да се потенцира дека дел од потешкотиите произлегуваат и од спецификите на природниот јазик на кој е поставено прашањето, како и јазикот на кој се напишани документите подложени на пребарување, со цел да се најде одговорот на истото. Градењето на „идеален“ *QA*-систем за одреден јазик побарува и постоење на анотиран корпус од кој може да се извлечат одредени знаења за успешна реализација на овој процес.

Овој истражувачки труд дава детален преглед на најзначајните постигнувања во областа на одговарањето прашања. Големо внимание е посветено на развојот на неопходните лексички бази на податоци, различните алатки и пристапи за одговарање прашања поставени на англиски јазик. Иако во помала мера, дадени се и белешки кои се однесуваат и на постигнувањата во другите јазици. Сепак, овој труд се фокусира на процесот на градење систем за одговарање прашања со повеќекратен избор за македонски јазик и ги изложува предизвиците кои произлегоа од овој процес. Во него е потенцирана и неопходноста од постоење на тест-колекција со прашања и документи на македонски јазик, која ги задоволува стандардните протоколи за превалидација и поствалидација. Само ваква колекција може да даде реална слика за резултатите добиени од направените тестирања и анализи на веќе постојните, како и новите методи за утврдување на точниот одговор, од неколку понудени одговори. Со цел да се обезбеди непристрасност во истражувањето, комплетното испитување во него е направено врз множество од прашања со повеќекратен избор вклучени во државната матура во Р. Македонија. Прашањата се од областа филозофија, бидејќи учебникот

„Филозофија“ (за четврта година гимназиско образование) е единствениот достапен извор со дозвола за преземање. Овие прашања се подложени на анализа заради утврдување на нивната многустраност и веродостојност. Множеството содржи вкупно 231 прашање, чии одговори се содржат во учебникот „Филозофија“.

Акцентот во овој истражувачки труд е ставен на откривањето на морфолошките белези на македонскиот јазик кои имаат силно влијание во процесот на прибирање информации. Поточно, се прави обид да се утврди важноста на информацијата за припадност на зборовите во одредена зборовна група и како оваа информација може да се искористи за подобрување на резултатите во пребарувањето. Исто така, направен е обид да се утврди квалитетот на прибраните резултати, доколку во пребарувањето се искористат:

- само зборовите од прашалникот,
- сите зборформи на збор од прашалникот кои се јавуваат во речникот генериран од учебникот „Филозофија“ (документ-колекцијата), и
- сите зборови од речникот кои имаат ист основен збор (стем) со збор од прашалникот.

Со цел да се разрешат овие прашања, истражувањето имплементира три стратегии за прибирање пасуси за дадено прашање, од кои петте најдобро рангирани се проследени на понатамошна анализа, за да се утврди точниот одговор на прашањето. Поточно, во системот се имплементирани:

- **моделот на векторски простор со косинус мерката** кој во ова истражување претставува модел за споредба (важноста на терминот се определува во согласност со $tf - idf$ шемата за задавање тежини),
- **совпаѓањето на координати** (важноста на терминот е 1 или 0, во согласност со тоа дали терминот се појавува во прашалникот или не), и
- **совпаѓањето на координати засилено со информацијата за зборовната група** (важноста на терминот е во согласност со неговата припадност во одредена зборовна група).

Во согласност со истражувањата во другите јазици (како и претпоставките за значајноста на зборовните групи во македонскиот јазик), дизајнираниот *QA*-систем вклучува пет зборовни групи во севкупниот процес на обработка, и тоа: именките, глаголите, придавките, прилозите и броевите. Најголема важност им е доделена на сопствените именки, следени од општите именки, потоа придавките и глаголите, додека прилозите и броевите се третирани како зборови со најмала важност. Направените испитувања за различни вредности на тежините на различните зборовни групи потврдуваат дека информацијата за зборовната група не е доминантната карактеристика во процесот на прибирање информации за тест-колекција „Филозофија“ (највисока точност се добива за моделот за споредба). Притоа, забележано е дека зголемувањето на разликата во тежините меѓу различните зборовни групи, доведува до намалување на квалитетот на прибраните резултати. За подобрување на резултатите во процесот на прибирање пасуси, направено е

комбинирање на информацијата за зборовната група на терминот, со неговата фреквенција и инверзната документ фреквенција. Ваквата комбинација се покажува како најуспешна за опишаната тест-колекција на македонски јазик. Уште повеќе, со вклучување на зборформите дополнително се зголемува точноста на системот.

Зборовите од македонската колекција во ова истражување се означени со примена на анотиран речник за македонскиот јазик, кој ги содржи најчесто користените македонски зборови, заедно со множеството зборформи за секој од нив. Од друга страна, заради непостоењето на лематизатор (*lemmatizer*) за македонскиот јазик, утврдувањето на зборовите (од речникот на колекцијата) кои припаѓаат во иста лексема е направено со примена на *Dice*-метриката за сличност на стрингови и кластерирањето со комплетна врска (за праг 0.5). Анализата на генерираните групи открива дека *Dice*-метриката може да се искористи единствено како помош во овој процес. За точното дефинирање на потребните групи, неопходно е истите да бидат подоложени на рачна корекција. Сепак, извршувањето на рачните корекции е исклучително макотрпен процес кој побарува и длабинско познавање на самиот јазик. За таа цел, за да се утврди метрика која ефективно би можела да се користи за автоматско групирање на зборовите кои припаѓаат во иста лексема, ова истражување имплементира две дополнителни метрики за сличност на стрингови, и тоа: *Positional Dice* и *Jaccard*. Истражувањето дефинира и нова метрика за автоматско групирање базирана на **триаголниот прозорец**, како една од познатите прозорски функции. Идејата да се примени оваа метрика произлезе од потребата да се зададе **награда** на секој карактер до кој се совпаѓаат два термина, сè до првиот карактер во кој тие се разликуваат. За тој карактер, како и за секој нареден, метриката задава **казна**. Имплементацијата на генерираните групи од четирите спомнати метрики (заедно со кластерирањето со комплетна врска за различни вредности на прагот) во процесот на прибирањето пасуси, потврди дека новодефинираната метрика постигнува највисоката точност (дури и во однос на примената на точните кластери од зборови, креирани рачно). Имено, системот успева да прибере точни пасуси за 215 прашања од вкупно 231 прашање (односно, точноста на системот изнесува 93.1%).

Анализата на добиените групи со примена на новата метрика покажа дека бројот на несоодветните групи (групи кои вклучуваат зборови што припаѓаат на различни лексеми) се зголемува со зголемувањето на вредноста на прагот. Таков пример претставува следнава група од зборови, добиена за вредност на прагот 2.5:

- {интегритет, интензитет, интерес, интереси, интересот, интересите, интернет}

Од друга страна, повисоките вредности на прагот дефинираат и групи од зборови кои имаат ист стем. Следниве групи претставуваат такви примери, добиени за вредност на прагот 2.5:

- {интелект, интелектот, интелектуалец, интелектуалецот, интелектуалци, интелектулизмот},
- {информатиката, информациите, информација, информацијата, информираноста}, и
- {настава, наставата, наставници, наставниците}.

Токму ваквото **делумно** групирање на зборовите од речникот во согласност со нивниот стем (групирањето е **делумно** бидејќи овие примери ги вклучуваат само зборовите кои имаат иста ознака за зборовната група), се претпоставува дека ја зголемува точноста на системот, во однос на примената на точните групи од зборови кои припаѓаат на иста лексема (креирани рачно).

Метриката базирана на **триаголниот прозорец** е искористена и за проширување на прашалникот за пребарување со **сличен** збор, во случај да содржи збор кој го нема во речникот од колекцијата (наречен **непознат** збор). Притоа, се смета дека метриката правилно го проширува прашалникот, доколку како сличен утврдува збор кој има ист стем како и непознатиот збор. Проширувањето на прашалникот е направено и со примена на претходно спомнатите метрики за сличност на стрингови: *Dice*, *Positional Dice* и *Jaccard*, и тоа за повеќе вредности на прагот. Сепак, анализите потврдуваат дека за оваа задача најуспешна е метриката базирана на **триаголниот прозорец**. Со имплементација на проширувањето на прашалникот со оваа метрика во процесот на прибирање пасуси, се добива највисока точност, и тоа за повеќе вредности на прагот. Во овој случај, точноста на системот изнесува 93.9% (односно, системот утврдува точен пасус за 217 прашања, од вкупно 231 прашање од тест-колекцијата „Филозофија“).

Истражувањето ја потенцира и важноста на близината меѓу термините од прашањето и термините од точниот одговор, во процесот на креирање стратегии за селекција на еден од четирите понудени одговори за дадено прашање. Во таа насока, *Hanning*-прозорската функција е имплементирана во последната фаза на системот за одговарање прашања. Во оваа фаза, системот ги анализира петте најдобро рангирани пасуси, определени во фазата за прибирање пасуси. *Hanning*-прозорската функција се вклучува во утврдување на густината на термините од прашалникот и еден од понудените одговори во одреден прозорец, со тоа што задава повисока тежина на термините, доколку тие се појавуваат поблиску еден до друг во прозорецот (утврдувањето важи за сите четири понудени одговори поодделно). Се претпоставува дека прозорецот каде густината е најголема, го содржи точниот одговор на даденото прашање. Во таа насока, истражувањето тестира повеќе вредности за големина на прозорецот w. Резултатите потврдуваат дека системот постигнува највисока точност за $w = 20$ (односно, прозорец кој содржи вкупно 21 збор од петте зборовни групи кои се вклучени во пребарувањето). Точноста на системот за селекција на точниот одговор изнесува 83.12% за тест-колекцијата „Филозофија“.

Конечните резултати потврдуваат дека во насока на успешно одговарање прашања на македонски јазик, исклучително важна улога има вклучувањето на зборформите на зборовите од прашањето (прашалникот) во процесот на прибирањето информации (и одговарањето прашања), како и земањето предвид на близината меѓу зборовите од прашањето и понудениот (генерираниот) одговор. За тест-колекцијата „Филозофија“, големината 20 за прозорецот w дава најдобри резултати при одговарање на прашањата од сите категории.

Изградениот систем во ова истражување е наменет пред сè за одговарање прашања поставени на македонски јазик (особено од областа филозофија). Сепак, за да се оценат можностите за негова примена врз колекции од друга област и напишани на друг природен јазик, системот е искористен за одговарање прашања со повеќекратен избор од две дополнителни тест-колекции. Колекциите се од иста област, информатички технологии, и се напишани на два различни јазика: македонски и англиски. Првите резултати добиени за двете колекции, потврдија дека е неопходно да се направат модификации на **густината на распределба** (*DD*), во која е вклучена *Hanning*-прозорската функција, со цел да се подобри точноста на системот. Причината за оваа потреба е лоцирана во малиот број термини кои се содржат во речникот на секоја од двете колекции (т.е. колекцијата документи од двете тест-колекции содржи мал број, кратки документи). Направената модификација покажа дека системот успешно може да се користи и за одговарање прашања од областа информатички технологии, и тоа на двата јазика. Помалата точност постигната за македонската колекција, се покажа како индикатор на низа недостатоци во нејзиното креирање. Најзабележителен проблем претставува незадоволувањето на стандардните протоколи за превалидација на дел од прашањата. Вкупните резултати, добиени за трите тестирани тест-колекции, потврдуваат дека во одредена мера, успешноста на дизајнираниот систем зависи и од карактеристиките на самата колекција (пред сè, нејзината големина, исполнувањето на стандардите за превалидација и поствалидација, застапеноста на зборови од други јазици, и слично).

Идните активности се во насока на дополнително продлабочување на истражувањата во повеќе сегменти, со цел да се подобри успешноста на системот за одговарање прашања на македонски јазик. Во овој момент, најзабележителен проблем претставува непостоењето на стандардизирана тест-колекција со прашања со повеќекратен избор на македонски јазик, која е неопходна за да се добие реална слика за постигнувањата на еден *QA*-систем. Една ваква колекција би овозможила извлекување на зборови кои се појавуваат заедно во текстуални сегменти од колекцијата документи (*co-occurring words*), како и извлекување на поврзаните зборови (*related words*). Следната важна цел е дефинирањето на означувач за македонскиот јазик за општ домен документи, како и искористувањето на синтаксичките белези кои позитивно би влијаеле врз процесот на прибирање информации (како и одговарање прашања поставени на македонски јазик). Со тоа, може да се очекува дека ефектот од примената на создадениот систем за одговарање прашања со повеќекратен избор ќе добие на квалитет и ќе ги достигне границите кои се поставени за јазиците со многу пообемна лингвистичка традиција.

РЕФЕРЕНЦИ:

[1] Bert F. Green, Jr., Alice K. Wolf, Carol Chomsky, and Kenneth Laughery. 1961. Baseball: An automatic question answerer. In Proceedings of Western Joint IRE-AIEE-ACM Computing Conference, volume 19, pp. 219–224.
DOI: 10.1145/1460690.1460714

[2] William A. Woods. 1973. Progress in Natural Language Understanding – An Application to Lunar Geology. In Proceedings of AFIPS Conference, volume 42, pp. 441–450.
DOI: 10.1145/1499586.1499695

[3] D. Ferrucci, E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. A. Kalyanpur, A. Lally, J. W. Murdock, E. Nyberg, J. Prager, N. Schlaefer, and C. Welty. 2010. Building Watson: An overview of the DeepQA project. *AI Magazine*, volume 31, issue 3, pp. 59–79.
DOI: 10.1609/aimag.v31i3.2303

[4] Nitin Indurkhya and Fred J. Damerau (Eds.). 2010. *Handbook of Natural Language Processing*, second edition (Chapman & Hall/CRC, Boca Raton).

[5] Jaime G. Carbonell, Donna Harman, Eduard Hovy, Steve Maiorano, John Prange, and Karen Sparck-Jones. 2000. Vision statement to guide research in question & answering (Q&A) and text summarization. *Report Technique NIST*.
http://www-nlpir.nist.gov/projects/duc/papers/Final-Vision-Paper-v1a.doc

[6] John Burger, Clarie Cardie, Vinay Chaudhri, Robert Gaizauskas, Sanda Harabagiu, David Israel, Christian Jacquemin, Chin-Yew Lin, Steve Maiorano, George Miller, Dan Moldovan, Bill Ogden, John Prager, Ellen Riloff, Amit Singhal, Rohini Shrihari, Tomek Strzalkowski, Ellen Voorhees, and Ralph Weishedel. 2000. Issues, tasks and program structures to roadmap research in question & answering (Q&A).
http://www-nlpir.nist.gov/projects/duc/papers/qa.Roadmap- paper_v2.doc

[7] Oleksandr Kolomiyets and Marie-Francine Moens. 2011. A survey on question answering technology from an information retrieval perspective. *Information Science*, volume 181, issue 24, pp. 5412– 5434.
DOI: 10.1016/j.ins.2011.07.047

[8] Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing* (2nd Edition). Prentice-Hall, Inc. Upper Saddle River, NJ, USA.

[9] John Prager, Jennifer Chu-Carroll, Krzysztof Czuba, Christopher Welty, Abraham Ittycheriah, and Richi Mahindru. 2003. IBM's PIQUANT in TREC2003. In *TREC 2003 Conference Notebook*, pp. 36-45.

[10] Dan Moldovan, Marius Pasca, Sanda Harabagiu, and Mihai Surdeanu. 2003. Performance issues and error analysis in an open-domain question answering system. ACM Transactions on Information Systems, volume 21, issue 2, pp. 133-154.
DOI: 10.1145/763693.763694

[11] Joao Silva, Luısa Coheur, Ana Mendes, and Andreas Wichert. 2011. From symbolic to sub-symbolic information in question classification. *Artificial Intelligence Review*, volume 35, issue 2, pp. 137–154.
DOI: 10.1007/s10462-010-9188-4

[12] John Prager, Dragomir Radev, Eric Brown, and Anni Coden. 1999. The use of predictive annotation for question answering in TREC8. In *NIST Special Publication 500-246: The Eighth Text REtrieval Conference (TREC 8)*, pp. 399–411.

[13] Xin Li and Dan Roth. Learning question classifiers: The role of semantic information. *Natural Language Engineering*, volume 12, issue 3, pp. 229-249.
DOI: 10.1007/s10462-010-9188-4

[14] Tianyong Hao, Wenxiu Xie, Chun Chen, and Yuming Shen. 2015. Systematic Comparison of Question Target Classification Taxonomies Towards Question Answering. *Social Media Processing*. *Communications in Computer and Information Science*, volume 568. Springer, Singapore.
DOI: 10.1007/978-981-10-0080-5_12

[15] Renu Mudgal, Rosy Madaan, A. K. Sharma, and Ashutosh Dixit. 2013. A Novel Architecture for Question Classification based Indexing Scheme for Efficient Question Answering. *International Journal of Computer Engineering and Applications*, volume 2, issue 2.

[16] Dan Moldovan, Sanda Harabagiu, Marius Pasca, Rada Mihalcea, Roxana Girju, Richard Goodrum, Vasile Rus. 2000. The Structure and Performance of an Open-Domain Question Answering System. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics (ACL-2000)*, pp. 563-570.
DOI: 10.3115/1075218.1075289

[17] Arthur C. Graesser, Natalie K. Person. 1994. Question asking during tutoring. *American Educational Research Journal*, volume 31, no. 1, pp. 104-137.

[18] Li Xin and Dan Roth. 2002. Learning question classifiers. In *Proceedings of the 19th International Conference on Computational Linguistics*, volume 1, pp. 1-7.
DOI: 10.3115/1072228.1072378

[19] Benamara Farah. 2004. Cooperative question answering in restricted domains: The WEBCOOP Experiments. In *Workshop on Question Answering in Restricted Domains*. 42nd Annual Meeting of the Association for Computational Linguistics, Barcelona, Spain, pp. 31–38.
http://www.aclweb.org/antology/W04-0506

[20] Kirk Roberts, Kate Masterton, Marcelo Fiszman, Halil Kilicoglu, and Dina Demner-Fushman. 2014. Annotating question types for consumer health questions. In *Proceedings of the Fourth LREC Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing*.

[21] James Richard Curran. 2003. From Distributional to Semantic Similarity, PhD thesis, University of Edinburgh.

https://www.era.lib.ed.ac.uk/bitstream/handle/1842/563/IP030023.pdf;jsessionid=79FAEDF2B885A2BD7CC512986B580C79?sequence=2

[22] Jochen L. Leidner. 2004. Open-Domain Question Answering from Large Text Collections, M. Pasca. *Journal of Logic, Language and Information*, volume 13, issue 3, pp. 373-376.
DOI: 10.1023/B:JLLI.0000028422.67488.ec

[23] Cristof Monz. 2004. Minimal span weighting retrieval for question answering. In *Proceedings of SIGIR Workshop on Information Retrieval for Question Answering*, pp. 23–30.

[24] Eric Brill, Susan Dumais, and Michele Banko. 2002. An analysis of the AskMSR question-answering system. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, volume, 10, pp. 257–264.
DOI: 10.3115/1118693.1118726

[25] James P. Callan. 1994. Passage-level evidence in document retrieval. In Proceedings of the 17th annual international ACM SIGIR conference on Research and Development in Information Retrieval, pp. 302–310.

[26] Ian Roberts and Robert Gaizauskas. 2004. Evaluating passage retrieval approaches for question answering. In *Proceedings 26th European Conference on IR Research (ECIR 2004)*, pages 72–84, Springer, Berlin, Heidelberg.
DOI: 10.1007/978-3-540-24752-4_6

[27] George A. Miller. 1995. WordNet: A Lexical Database for English. *Communications of the ACM*, volume 38, issue 11, pp. 39-41. ACM, New York, NY, USA.
DOI: 10.1145/219717.219748

[28] Bernardo Magnini, Matteo Negri, Roberto Prevete, and Hristo Tanev. 2002. Is It the Right Answer? Exploiting Web Redundancy for Answer Validation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (ACL-2002), pp. 425-432.

[29] Hui Yang and Tat-Seng Chua. 2002. The integration of lexical knowledge and external resources for question answering. In *Proceedings of the 11th Text REtrieval Conference*, Gaithersburg, Maryland, USA. NIST Special Publications 500-251 (2003).

[30] C. L. Clarke, G. V. Cormack, G. Kemkes, M. Laszlo, T. R. Lynam, E. L. Terra, and P. L. Tilker. 2002. Statistical selection of exact answers (MultiText Experiments for TREC 2002). In *Proceedings of the 11th Text REtrieval Conference*, Gaithersburg, Maryland, USA. NIST Special Publications 500-251 (2003), pp. 823-831.

[31] James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning*. O'Reilly Media, Inc., Sebastopol, CA.

[32] Diego Molla Aliod and Mary Gardiner 2004. Answerfinder-Question answering by combining lexical, syntactic and semantic information. In *Proceedings of the Australasian Language Technology Workshop 2004*, pp. 9-16.

[33] Amit Mishra and Sanjay Kumar Jain. 2016. A survey on question answering systems with classification. *Journal of King Saud University – Computer and Information Sciences*, volume 28, issue 3, pp. 345-361, Elsevier Science Inc., New York, USA. DOI: 10.1016/j.jksuci.2014.10.007

[34] Michael Minock. 2005. Where are the 'killer applications' of restricted domain question answering? In *Proceedings of the IJCAI Workshop on Knowledge Reasoning in Question Answering*, pp. 4. Edinburg, Scotland.

[35] Vanessa Lopez, Victoria Uren, Marta Sabou, and Enrico Motta. 2011. Is question answering fit for the semantic web?: a survey. Semantic Web, volume 2, issue 2, pp. 125–155.
DOI: 10.3233/SW-2011-0041

[36] Donald Metzler and W. Bruce Croft. 2005. Analysis of statistical question classification for fact-based questions. *Information Retrieval*, volume 8, issue 3, pp. 481–504.
DOI: 10.1007/s10791-005-6995-3

[37] Bernard J. Jansen and Danielle Booth. 2010. Classifying web queries by topic and user intent. In *Proceedings of the 28th international conference on human factors in computing systems*, pp. 4285-4290.
DOI: 10.1145/1753846.1754140

[38] Fan Bu, Xingwei Zhu, Yu Hao, and Xiaoyan Zhu. 2010. Function-based question classification for general QA. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, pp. 1119–1128.

[39] Sanjay K. Dwivedi and Vaishali Singh. 2013. Research and reviews in question answering system. *International Conference on Computational Intelligence: Modeling Techniques and Applications*. Procedia Technology, volume 10, pp. 417-424. DOI: 10.1016/j.protcy.2013.12.378

[40] Ellen M. Voorhees. 1999. The TREC-8 Question Answering Track Report. *Trec 99*, pp. 77-82.

[41] Arnaud Grappy and Brigitte Grau. 2010. Answer type validation in question answering systems. In *Proceedings of RIAO '10 Adaptivity, Personalization and Fusion of Heterogeneous Information*, pp. 9-15.

[42] Suzan Verberne, Lou Boves, Nelleke Oostdijk, and Peter-Arno Coppen. 2008. Using syntactic information for improving *why*-question answering. In *Proceedings of the 22nd International Conference on Computational Linguistics*, volume 1, pp. 953–960.

[43] Suzan Verberne, Lou Boves, Nelleke Oostdijk, and Peter-Arno Coppen. 2010. What is not in the bag of words for why-QA? *Computational Linguistics*, volume 32, issue 2, pp. 229-245.

[44] Malik Muhammad Saad Missen, Mohand Boughanem, and Guillaume Cabanac. 2010. Opinion finding in blogs: a passage-based language modeling approach. In *Proceedings of the 10th International Conference on Adaptively, Personalization and Fusion of Heterogeneous Information*, pp. 148–152.

[45] Soujanya Poria, Alexander Gelbukh, Erik Cambria, Peipei Yang, Amir Hussain, and Tariq Durrani. 2012. Merging SenticNet and WordNet-Affect emotion lists for sentiment analysis. *IEEE 11th International Conference on Signal Processing*, volume 2, pp. 1251–1255.
DOI: 10.1109/ICoSP.2012.6491803

[46] Soujanya Poria, Erik Cambria, Gregoire Winterstein, and Guang-Bin Huang. 2014. Sentic patterns: dependency-based rules for concept-level sentiment analysis. Knowledge-Based Systems, volume 69, pp. 45–63.
DOI: 10.1016/j.knosys.2014.05.005

[47] Khairullah Khan, Baharum Baharudin, Aurnagzeb Khan, and Ashraf Ullah. 2014. Mining opinion components from unstructured reviews: A review. *Journal of King Saud University – Computer and Information Sciences*, volume 26, Issue 3, pp. 258-275, Elsevier Science Inc., New York, USA.
DOI: 10.1016/j.jksuci.2014.03.009

[48] Khairullah Khan, Baharum Baharudin, Aurnagzeb Khan, and Asharf Ullah. 2014. Mining opinion components from unstructured reviews: A review. *Journal of King Saud University – Computer and Information Sciences*, volume 23, issue 3, pp. 258-275.
DOI: 10.1016/j.knosys.2014.03.009

[49] Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, volume 28, issue 3, pp. 245-288.
DOI: 10.1162/089120102760275983

[50] Elizabeth D. Liddy. 1998. Enhanced Text Retrieval Using Natural Language Processing. *Bulletin of the American Society for Information Science*, volume 24, issue 4, pp. 14-16.
http://www.asis.org/Bulletin/Apr-98/liddy.html.

[51] Inderjeet Mani and Mark T. Maybury. 1999. *Advances in Automatic Text Summarization*. MIT Press Cambridge, MA, USA.

[52] Frederik Hogenboom, Flavius Frasincar, and Uzay Kaymak. 2010. An Overview of Approaches to Extract Information from Natural Language Corpora. In *10th Dutch-Beligian Information Retrieval Workshop*, pp. 69-70.
https://www.frederikhogenboom.nl/work/papers/dir10-nlp.pdf

[53] Li Cai, Guangyou Zhou, Kang Liu, Jun Zhao. 2011. Large-scale question classification in cQA by leveraging Wikipedia semantic knowledge. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pp. 1321–1330.
DOI: 10.1145/2063576.2063768

[54] Y. Q. Niu. 2011. Study on Question Classification in Chinese Question Answering System (in Chinese), master thesis.

[55] Natsuda Laokulrat. 2013. A survey on question classification techniques for question answering. *KMITL Information Technology Journal*, volume 2, issue 1.
http://journal.it.kmitl.ac.th

[56] Hakan Sundblad. 2007. Question classification in question answering systems, diploma thesis. Department of Computer and Information Science at Linköping University.

[57] Dell Zhang and Wee Sun Lee. 2003. Question classification using support vector machines. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 26–32.
DOI: 10.1145/860435.860443

[58] Zhiheng Huang, Marcus Thint, and Zengchang Qin. 2008. Question classification using head words and their hypernyms. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 927–936.

[59] Zhiheng Huang, Marcus Thint, and Asli Celikyilmaz. 2009. Investigation of question classifier in question answering. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, volume 2, pp. 543– 550.
DOI: 10.1145/1148170.1148282

[60] Phil Blunsom, Krystle Kocik, and James R. Curran. 2006. Question classification with log-linear models. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 615–616.

[61] Babak Loni, Gijs van Tulder, Pascal Wiggers, Marco Loog, and David Tax. 2011. Question classification by weighted combination of lexical, syntactical and semantic features. In *Proceedings of the 14th international conference on Text, speech and dialogue*, pp. 243-250.

[62] Guohua Chen, Yong Tang, Yan Pan, and Qiang Deng. 2011. Question Classification using Multiple Kernel Learning and Semantic Information. *Journal of Computers* volume 6, no. 11.

[63] H. Hardy and Yu-N Cheah. 2013. Question Classifcation Using Extreme Learning Machine on Semantic Features. Journal of ICT Research and Applications, volume 7, no. 1, pp. 36-58.
DOI: 10.5614%2Fitbj.ict.res.appl.2013.7.1.3

[64] Nguyen Van-Tu and Le Anh-Cuong. 2016. Improving Question Classification by Feature Extraction and Selection. *Indian Journal of Science and Technology*, volume 9, issue 17.
DOI: 10.17485/ijst/2016/v9i17/93160

[65] Kepei Zhang and Jieyu Zhao. 2010. A Chinese question answering system with question classification and answer clustering. In *Proceedings of IEEE International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, volume 6, pp. 2692-2696.
DOI: 10.1162/089120102760275983

[66] Ali Mollaei, Saeed Rahati-Quchani, and Azam Estaji. 2012. Question classification in Persian language based on conditional random fields. In *Proceedings of 2nd International eConference on Computer and Knowledge Engineering*.
DOI:10.1109/ICCKE.2012.6395395

[67] Nouha Othman and Rim Faiz. 2016. Question Answering Passage Retrieval and Re-ranking Using N-grams and SVM. *Computación y Sistemas*, volume 20, no. 3, pp. 483–494.
DOI: 10.13053/CyS-20-3-2470

[68] Di Wang and Eric Nyberg. 2015. CMU OAQA at TREC 2015 LiveQA: Discovering the Right Answer with Clues. In *Proceedings of the 24th Text Retrieval Conference*. http://trec.nist.gov/pubs/trec24/papers/oaqa-QA.pdf

[69] Di Wang and Eric Nyberg. 2015. A long short-term memory model for answer sentence selection in question answering. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), pp. 707–712, ACL.

[70] Di Wang and Eric Nyberg. 2015. A recurrent neural network based answer ranking model for web question answering. In *SIGIR Workshop on Web Question Answering: Beyond Factoids*.

[71] Joel Mackenzie, Ruey-Cheng Chen, and J. Shane Culpepper. 2016. RMIT at the TREC 2016 LiveQA Track. In *Proceedings of the 25th Text Retrieval Conference*. http://jmmackenzie.io/pdf/mcc16-liveqa.pdf

[72] Vanessa Murdock and W. Bruce Croft. 2004. Simple Translation Models for Sentence Retrieval in Factoid Question Answering. In Proceedings of the SIGIR Workshop on Information Retrieval for Question Answering, pp. 31-35.

[73] Jose Manuel Gomez-Soriano, Manuel Montes-y-Gomez, Emilio Sanchis-Arnal, Luis Viallasenor-Pineda, and Paolo Rosso. 2005. Language Independent Passage Retrieval for Question Answering. In *MICAI 2005: Advances in Artificial Intelligence*, pp. 816-823.
DOI: 10.1007/11579427_83

[74] Abraham Ittycheriah, Martin Franz, Wei-Jing Zhu, Adwait Ratnaparkhi, and Richard J. Mammone. 2000. IBM's statistical question answering system. In *Proceedings of the 9th Text Retrieval Conference (TREC-9)*, pp. 231-236.

[75] Anselmo Penas, Pamela Forner, Alvaro Rodrigo, Richard F. E. Sutcliffe, Corina Forascu, and Cristina Mota. 2010. Overview of ResPubliQA 2010: Question answering evaluation over European legislation. *CLEF 2010 LABs and Workshops, Notebook Papers*.

[76] Eugene Agichtein, David Carmel, Donna Harman, Dan Pelleg, and Yuval Pinter. Overview of the TREC 2015 LiveQA Track. In *Preceedings of the 24th Text Retrieval Conference*, Gaithersburg, Maryland, USA.
http://trec.nist.gov/pubs/trec24/papers/Overview-QA.pdf

[77] Alexandru Marius Pasca. 2001. High performance, open-domain question answering from large text collections, doctorial dissertation. Southern Methodist University Dallas, TX, USA.

[78] Abraham P. Ittycheriah. 2001. Trainable question-answering systems, doctorial dissertation. Rutgers University New Brunswick, NJ, USA.

[79] Jun Suzuki, Yutaka Sasaki, and Eisaku Maeda. 2002. SVM answer selection for open-domain question answering. In *Proceedings of l9th International Conference on Computational Linguistics*, volume 1, pp. 1-7.
DOI: 10.3115/1072228.1072347

[80] Peng Li, Yi Guan, Xiao-Iong Wang. 2006. Answer extraction based on system similarity model and stratified sampling logistic regression in rare data. *International Journal of Computer Science and Network Security*, volume 6, no. 3, pp.189-196.

[81] Adam Berger, Rich Caruana, David Cohn, Dayne Freitag, and Vibhu Mittal. 2000. Bridging the lexical chasm: statistical approaches to answer-finding. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 192-199.
DOI: 10.1145/345508.345576

[82] Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. Computational Linguistics, volume 16, issue 2, pp. 79–85.

[83] Radu Soricut and Eric Brill. 2006. Automatic question answering using the web: Beyond the factoid. In *Journal of Information Retrieval-Special Issue on Web Information Retrieval*, volume 9, issue 2, pp. 191-206.
DOI: 10.1007/s10791-006-7149-y

[84] Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. Computational Linguistics – Special issue on using large corpora, volume 19, issue 2, pp. 263-312.

[85] Dongfen Cai, Yanju Dong, Dexin Lv, Guiping Zhang, and Xuelei Miao. 2005. A Web-based Chinese question answering with answer validation. In *Proceedings of IEEE International Conference on Natural Language Processing and Knowledge Engineering*, pp. 499-502.
DOI: 10.1109/NLPKE.2005.1598788

[86] Suzan Verberne, Hans van Halteren, Daphne Theijssen, Stephan Raaijmakers, and Lou Boves. 2011. Learning to rank for why-questions answering. *Information Retrieval*, volume 14, issue 2, pp. 107-132, Springer.
DOI: 10.1007/s10791-010-9136-6

[87] Ludovid Denoyer and Patrick Gallinari. 2006. The Wikipedia XML corpus. ACM SIGIR Forum, volume 40, issue 1, pp. 64–69. ACM, New York, USA.
DOI: 10.1145/1147197.1147210

[88] Eduard Hovy, Ulf Hermjakob, and Deepak Ravichandran. 2002. A Question/Answer Typology with Surface Text Patterns. In *Proceedings of the 2nd international conference on Human Language Technology Research*, pp. 247–251. San Diego, CA, USA.

[89] Arvind Agarwal, Hema Raghavan, Karthik Subbian, Prem Melville, Richard D. Lawrence, David C. Gondek, and James Fan. 2012. Learning to rank for robust question answering. In *Proceedings of the 21st ACM international conference on Information and knowledge* management, pp. 833-842.
DOI: 10.1145/2396761.2396867

[90] Yoav Freund, Raj Iyer, Robert E. Schapire, and Yoram Singer. 2003. An efficient boosting algorithm for combining preferences. *The Journal of Machine Learning Research*, volume 4, pp. 933–969.

[91] Jun Xu and Hang Li. 2007. AdaRank: a boosting algorithm for information retrieval. In Proceedings of the International 30th annual international ACM SIGIR conference on research and development in information retrieval, pp. 391–398.
DOI: 10.1145/1277741.1277809

[92] Donald Metzler and W. Bruce Croft. 2007. Linear feature-based models for information retrieval. *Information Retrieval*, volume 10, issue 3, pp. 257–2
DOI: 10.1007/s10791-006-9019-z

[93] Cristopher J. C. Burges, Roberto Rango, and Quoc Viet Le. 2006. Learning to rank with nonsmooth cost functions. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, pp. 193–200.

[94] Minwei Feng, Bing Xiang, Michael R. Glass, Lidan Wang, and Bowen Zhou. 2015.Applying deep learning to answer selection: A study and an open task. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU),* IEEE. DOI: 10.1109/ASRU.2015.7404872

[95] Joseph Weizenbaum. 1966. ELIZA – a computer program for the study of natural language communication between man and machine. In *Communications of the ACM*,

volume 9, issue 1, pp. 36-45.
DOI: 10.1145/365153.365168

[96] Daniel G. Bobrow, Ronald M. Kaplan, Matin Kay, Donald A. Norman, Henry Thompson, and Terry Winograd. 1977. Gus, a frame-driven dialog system. *Artificial Intelligence*, volume 8, issue 2, pp. 155-173.
DOI: 10.1016/0004-3702(77)90018-2

[97] Boris Katz. 1997. Annotating the World Wide Web using natural language. In *Proceedings of the 5th RIAO conference on Computer Assisted Information Searching on the Internet*, pp. 136-159.

[98] Hoojung Chung, Young-In Song, Kyoung-Soo Han, Do-Sang Yoon, Joo-Young Lee, Hae-Chang Rim, and Soo-Hong Kim. 2004. A Practical QA System in Restricted Domains. In *Workshop on Question Answering in Restricted Domains. 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 39-45.

[99] Amit Mishra, Nidhi Mishra, and Anupam Agrawal. 2010. Context-aware restricted geographical domain question answering system. In *Proceedings of IEEE International Conference on Computational Intelligence and Communication Networks (CICN)*, pp. 548-553.
DOI: 10.1109/CICN.2010.108

[100] Slav Petrov and Dan Klein. Improved inference for unlexicalized parsing. 2007. In *Proceedings of NAACL HLT 2007*, pp. 404-411, ACL.
http://aclweb.org/anthology/N07-1051

[101] Payal Biswas, Aditi Sharan, and Rakesh Kumar. 2014. Question Classification using syntactic and rule based approach. *International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, IEEE.
DOI: 10.1109/ICACCI.2014.6968434

[102] Martin M. Soubbotin and Sergei M. Soubbotin. 2002. Use of patterns for detection of likely answer strings: A systematic approach. In *Proceedings of the 11th Text Retrieval Conference*, Gaithersburg, MD.

[103] Jennifer Chu-Carroll, John Prager, Christopher Welty, Krzysztof Czuba, and David Ferrucci. 2003. A multi-strategy and multi-source approach to question answering. In *Proceedings of TREC 2002*, pp. 281-288.

[104] D. A. Ferrucci. 2012. Introduction to “This is Watson”. *IBM Journal of Research and Development*, volume 56, issue 3.4, pp. 1-15, IBM.
DOI: 10.1147/JRD.2012.2184356

[105] A. Lally, J. M. Prager, M. C. McCord, B. K. Boguraev, S. Patwardhan, J. Fan, P. Fodor, and J. Chu-Carroll. 2012. Question analysis: How Watson reads a clue. *IBM Journal of Research and Development*, volume 56, issue 3.4, pp. 2:1-2:14.
DOI: 10.1147/JRD.2012.2184637

[106] M. C. McCord, J. W. Murdock, and B. K. Boguraev. 2012. Deep parsing in Watson. *IBM Journal of Research and Development*, volume 56, issue 3, pp. 264-278, Riverton, NJ, USA.
DOI: 10.1147/JRD.2012.2185409

[107] C. Wang, A. Kalyanpur, J. Fan, B. K. Boguraev, and D. C. Gondek. Relation extraction and scoring in DeepQA. 2012. *IBM Journal of Research and Development*, volume 56, issue, 3, pp. 339-350. Riverton, NJ, USA.
DOI: 10.1147/JRD.2012.2187239

[108] J. Chu-Carroll, J. Fan, B. K. Boguraev, D. Carmel, D. Sheinwald, and C. Welty. 2012. Finding needles in the haystack: Search and candidate generation. *IBM Journal of Research and Development*, volume 56, issue 3, pp. 300-311. Riverton, NJ, USA. DOI: 10.1147/JRD.2012.2186682

[109] Donald Metzler and W. Bruce Croft. 2004. Combining the language model and inference network approaches to retrieval. International Journal of *Information Processing and Management - Special Issue: Bayesian networks and information retrieval*, volume 40, issue 5, pp. 735–750. Tarrytown, NY, USA.
DOI: 10.1016/j.ipm.2004.05.001

[110] Erik Hatcher and Otis Gospodnetic. 2005. Lucene in Action. Manning Publication Co., Greenwich, CT.

[111] J. Fan, A. Kalyanpur, D. C. Gondek, and D. Ferrucci. 2012. Automatic knowledge extraction from documents. *IBM Journal of Research and Development,* volume 56, issue 3.4, pp. 290-299.
DOI: 10.1147/JRD.2012.2186519

[112] J. W. Murdock, J. Fan, A. Lally, H. Shima, and B. K. Boguraev. 2012. Textual evidence gathering and analysis. *IBM Journal of Research and Development*, volume 56, issue 3, pp. 325-338. Riverton, NJ, USA.
DOI: 10.1147/JRD.2012.2187249

[113] Temple F. Smith and Michael S. Waterman. 1981. Identification of common molecular subsequences. *Journal of molecular biology*, volume 147, issue 1, pp. 195–197. London.
DOI: 10.1016/0022-2836(81)90087-5

[114] D. C. Gondek, A. Lally, A. Kalyanpur, J. W. Murdock, P. A. Duboue, L. Zhang, Y. Pan, Z. M. Qiu, and C. Welty. 2012. A framework for merging and ranking of answers in deepqa. *IBM Journal of Research and Development*, volume 56, issue 3, pp. 399-410. Riverton, NJ, USA.

[115] Huan Liu and Rudy Setiono. 1996. A probabilistic approach to feature selection - a filter solution. In *Proceedings of the 13th International Conference on International Conference on Machine Learning*, pp. 319-327. Bari, Italy.

[116] Ellen M. Voorhees and Dawn M. Tice. 2000. Building a question answering test collection. In *Proceedings of the 23rd Annual International ACM SIGIR Conference*

on Research and Development in Information Retrieval, pp. 200–207, ACM, New Jork, USA.
DOI: 10.1145/345508.345577

[117] Lynette Hirschman, Marc Light, Eric Breck, and John D. Burger. 1999. Deep read: A reading comprehension system. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics* (*ACL-99*), pp. 325–332, Stroudsburg, PA, USA.
DOI: 10.3115/1034678.1034731

[118] Eugene Charniak, Yasemin Altun, Rodrigo de Salvo Braz, Benjamin Garrett, Margaret Kosmala, Tomer Moscovich, Lixin Pang, Changhee Pyo, Ye Sun, Wei Wy, Zhongfa Yang, Shawn Zeller, and Lisa Zorn. 2000. Reading comprehension programs in a statistical-language-processing class. In *Proceedings of the 2000 ANLP/NAACL Workshop on Reading comprehension tests as evaluation for computer-based language understanding sytems*, volume 6, pp. 1–5, Stroudsburg, PA, USA.
DOI: 10.3115/1117595.1117596

[119] Eric Breck, John D. Burger, L. Ferro, Lisa Ferro, Lynette Hirschman, David House, Marc Light, and Inderjeet Mani. 2000. How to evaluate your question answering system every day and still get real work done. In *Proceedings of 2nd International Conference on Language Resources and Evaluation (LREC-2000)*, pp. 1495-1500, Athens, Greece.

[120] Јасмина Арменска. 2011. Примена на методите за пребарување информации кај системите за одговарање прашања, магистерска тема. Природно-математички факултет – Скопје, УКИМ.

[121] Eugene Agichtein, David Carmel, Dan Pelleg, Yuval Pinter, and Donna Harman. 2016. Overview of the TREC 2016 LiveQA Track. In *Proceedings of the 25th Text Retrieval Conference (TREC 2016)*.
http://trec.nist.gov/pubs/trec25/papers/Overview-QA.pdf

[122] Di Wang and Eric Nyberg. 2016. CMU OAQA at TREC 2016 LiveQA: An Attentional Neural Encoder-Decoder Approach for Answer Ranking. In *Proceedings of the 25th Text Retrieval Conference (TREC 2016)*.
http://trec.nist.gov/pubs/trec25/papers/CMU-OAQA-QA.pdf

[123] Стојка Бојковска, Лилјана Минова-Гуркова, Димитар Пандев, Живко Цветковски. 2008. *Општа граматика на македонскиот јазик*. Просветно дело, Скопје.

[124] Mari Vallez, Rafaela Pedraza-Jimenez. 2007. Natural Language Processing in Textual Information Retrieval and Related Topics. *Hipertext.net*, num. 5.
https://www.upf.edu/hipertextnet/en/numero-5/pln.html

[125] Ayman Farahat, Francine Chen, and Thorsten Brants. 2003. Optimizing story link detection is not equivalent to optimizing new event detection. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, volume 1,

pp. 232-239, ACL, Stroudsburg, PA, USA.
DOI: 10.3115/1075096.1075126

[126] Anne Cutler. 1986. Forbear is a homophone: Lexical prosody does not constrain lexical access. *Language and Speech*, volume 29, issue 3, pp. 201–220.
DOI: 10.1177/002383098602900302

[127] Barbara B. Greene and Gerald M. Rubin. 1971. Automatic grammatical tagging of English. *Technical Report*, Department of Linguistics, Brown University, Providence, Rhode Island.

[128] Roger Garside and Nicholas Smith. 1997. A hybrid grammatical tagger: CLAWS4. In *Corpus Annotation: Linguistic Information from Computer Text Corpora*, pp. 102–121, Longman, London.

[129] Francis, W. Nelson. 1979. A tagged corpus – problems and prospects. In *Studies in English linguistics for Randolph Quirk*, pp. 192–209, Longman, London.

[130] Francis, W. N. and Kucera, H. (1982). *Frequency Analysis of English Usage*. Houghton Mifflin, Boston.

[131] Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics – Special issue on using large corpora: II*, volume 19, issue 2, pp. 313–330, MIT Press Cambridge, MA, USA.

[132] Steven J. DeRose. 1988. Grammatical category disambiguation by statistical optimization. *Computational Linguistics*, volume 14, issue 1, pp. 31-39, MIT Press Cambridge, MA, USA.

[133] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schutze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, England.

[134] M. Bonchanoski, and K. Zdravkova. 2017. Machine Learning-based approach to automatic POS tagging of Macedonian language. In *Proceedings of 8th Balkan Conference in Informatics, (BCI2017)*, article no. 11, ACM New York, NY, USA.
DOI: 10.1145/3136273.3136275

[135] John Lyons. 1977. *Semantics: Volume 2*. Cambridge University Press

[136] Otto Jespersen. 1929. *The Philosophy of Grammar.* Allen and Unwin, London.

[137] Wessel Kraaij and Renee Pohlmann. 1996. Viewing stemming as recall enhancement. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 40–48, ACM, New York, USA.
DOI: 10.1145/243199.243209

[138] Chirag Shah and Pushpak Bhattacharyya. 2002. A Study for Evaluating the Importance of Various Parts of Speech (POS) for Information Retrieval (IR). In *Proceedings of International Conference on Universal Knowledge and Languages (ICUKL)*, Goa, India.

[139] Judith Klavans and Min-Yen Kan. 1998. Role of Verbs in Document Analysis. In *Proceedings of the 17th International Conference on Computational Linguistics*, volume 1, pp. 680–686, ACL, Stroudsburg, PA, USA.
DOI: 10.3115/980451.980959

[140] Peter D. Turney and Michael L. Littman. 2002. Unsupervised Learning of Semantic Orientation from a Hundred-Billion-Word Corpus. *National Research Council, Institute for Information Technology, Technical report ERB-1094*, 11 pages.

[141] Ruilin Xu. 2014. POS weighted TF-IDF algorithm and its application for an MOOC search engine. *International Conference on Audio, Language and Image Processing (ICALIP)*, pp. 868-873, IEEE.
DOI: 10.1109/ICALIP.2014.7009919

[142] Christina Lioma and Roi Blanco. 2009. Part of Speech Based Term Weighting for Information Retrieval. In *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, pp. 412-423, Springer-Verlag, Berlin, Heidelberg.
DOI: 10.1007/978-3-642-00958-7_37

[143] Yanshan Wang, Digcheng Li, Stephen Wu, and Hongfang Liu. 2015. Improving Clinical Information Retrieval by Incorporating Part-Of-Speech Tagging. Delivery Science Summit 2015.
DOI: 10.13140/RG.2.1.2907.7363

[144] Chengxiang Zhai and John D. Lafferty. 2001. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 334–342, ACM, New York, USA.
DOI: 10.1145/383952.384019

[145] Thorsten Brants. 2000. TnT: a statistical part-of-speech tagger. In *Proceedings of the 6th conference on Applied natural language processing*, pp. 224–231, ACL, Stroudsburg, PA, USA.

[146] Csaba Oravecz and Pter Dienes. 2002. Efficient stochastic part-of-speech tagging for Hungarian. In *Proceedings of the 3th International Conference on Language Resources and Evaluation*, pp. 710–717.

[147] Jan Hajic. 2000. Morphological tagging: data vs. dictionaries. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pp. 94-101, ACL, Stroudsburg, PA, USA.

[148] Tomaz Erjavec. 2004. Multext-east version 3: Multilingual morphosyntactic specifications, lexicons and corpora. In *4th International Conference on Language Resources and Evaluation, LREC'04,* pp. 1535–1538, ELRA, Paris.
http://nl.ijs.si/et/Bib/LREC04/

[149] Reza Karimpour, Amineh Ghorbani, Azadeh Pishdad, Mitra Mohtarami, Abolfazl AleAhmad, Hadi Amiri, and Farhad Oroumchian. 2009. Improving Persian

Information Retrieval Systems Using Stemming and Part of Speech Tagging. In *CLEF'08 Proceedings of the 9th Cross-language evaluation forum conference on evaluating systems for multilingual and multimodal information access*, pp. 89-96, Springer-Verlag Berlin, Heidelberg.
DOI: 10.1007/978-3-642-04447-2_10

[150] Jon Dehdari and Deryle Lonsdale. 2008. A Link Grammar Parser for Persian. *Aspects of Iranian Linguistics*, volume 1, Cambridge Scholars Press.

[151] Ghassan Kanaan, Riyad al-Shalabi, and Majdi Sawalha. 2005. Improving Arabic information retrieval systems using part of speech tagging. *International Technology Journal*, volume 4, issue 1, pp. 32-37.
DOI: 10.3923/itj.2005.32.37

[152] Abdur Chowdhury and M. Catherine McCabe. 1998. *Improving Information Retrieval Systems using Part of Speech Tagging*. Technical report, ISR, Institute for Systems Research.

[153] Yue Zhang and Stephen Clark. 2008. Joint word segmentation and POS tagging using a single perceptron. In *Proceedings of ACL-08: HLT*, pp. 888–896, ACL.

[154] Michael Collins. 2002. Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, volume 8, pp. 1–8, ACL, Stroudsburg, PA, USA.
DOI: 10.3115/1118693.1118694

[155] Hui Ning, Hua Yang, and Zhihuo Li. 2007. A method integrating rule and HMM for Chinese part-of-speech tagging. In *2nd IEEE Conference on Industrial Electronics and Applications*, pp. 723–725, IEEE.
DOI: 10.1109/ICIEA.2007.4318501

[156] Anjali Ganesh Jivani. 2011. A Comparative Study of Stemming Algorithms. *International Journal of Computer Technology and Application*, volume 2, issue 6, pp. 1930-1938.

[157] Ari Pirkola. 2001. Morphological typology of languages for IR. Journal of Documentation, volume 57, no. 3, pp. 330-348.

[158] Mirko Popovic, Peter Willett. 1992. The effectiveness of stemming for natural-language access to Slovene textual data. *Journal of the American Society for Information Science*, volume 43, no. 5, pp. 384-390.

[159] Dalwadi Bijal and Suthar Sanket. 2014. Overview of Stemming Algorithms for Indian and Non-Indian Languages. International Journal of Computer Science and Information Technologies, volume 5, issue 2, pp. 1144-1146.

[160] Donna Harman. 1991. How effective is suffixing? *Journal of the American Society for Information Science*, volume 42, issue 1, pp. 7-15.

[161] Martin F Porter. 1980. An algorithm for suffix stripping. *Program electronic library and information systems*, volume 14, no. 3, pp.130–137.

[162] Julie Beth Lovins. 1968. Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, volume 11, no. 1-2, pp. 22–31.

[163] Chris D. Paice. 1990. Another stemmer. *ACM SIGIR Forum*, volume 24, issue 3, pp. 56-61, ACM New York, USA.
DOI: 10.1145/101306.101310

[164] J. L. Dawson. 1974. Suffix removal for word conflation. *Bulletin of the Association for Literary & Linguistic Computing*, volume 2, issue 3, pp. 33-46. (161) [164]

[165] Felix Naumann. 2013. Similarity measures.
https://hpi.de/fileadmin/user_upload/fachgebiete/naumann/folien/SS13/DPDC/DPDC_12_Similarity.pdf

[166] Massimo Melucci and Nicola Orio. 2003. A novel method for stemmer generation based on hidden Markov models. In *Proceedings of the 12th international conference on Information and knowledge management*, pp. 131-138, ACM, New York, USA.
DOI: 10.1145/956863.956889

[167] Prasenjit Majumder, Mandar Mitra, Swapan K. Parui, Gobinda Kole, Pabitra Mitra, and Kalyankumar Datta. 2007. YASS: Yet another suffix stripper. *ACM Transactions on Information Systems*, volume 25, issue 4, article no. 18. ACM, New York, USA.
DOI: 10.1145/1281485.1281489

[168] Robert Krovetz. 1993. Viewing morphology as an inference process. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 191-202, ACM, New York, USA.
DOI: 10.1145/160688.160718

[169] David A. Hull and Gregory Grefenstette. 1996. A detailed analysis of English Stemming Algorithms. *XEROX Technical Report*.

[170] Jinxi Xu and W. Bruce Croft. 1998. Corpus-based stemming using co-occurrence of word variants. *ACM Transactions on Information Systems*, volume 16, issue 1, pp. 61-81. ACM, New York, USA.
DOI: 10.1145/267954.267957

[171] Funchun Peng, Nawaaz Ahmed, Xin Li and Yumao Lu. 2007. Context sensitive stemming for web search. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 639-646, ACM, New York, USA.
DOI: 10.1145/1277741.1277851

[172] George W. Adamson and Jillian Boreham. 1974. The use of an association measure based on character structure to identify semantically related pairs of words and document titles. *Information Storage and Retrieval*, volume 10, issues 7-8, pp. 253-

260, Elsevier.
DOI: 10.1016/0020-0271(74)90020-5

[173] Alexander M. Robertson, Peter Willett. 1998. Applications of n-grams in textual information systems. *Journal of Documentation*, volume, 54, issue 1, pp. 48 – 67. DOI: 10.1108/EUM0000000007161

[174] F. Cuna Ekmekcioglu, Michael F. Lynch, Peter Willett. 1996. Stemming and n-gram matching for term conflation in Turkish texts. *Information Research*, volume 2, no. 2, pp. 2-6.

[175] Jan Snajder and Bojana Dalbelo Basic. String Distance-Based Stemming of the Highly Inflected Croatian Language. In *Proceedings of Recent Advances in Natural Language Processing (RANLP-2009)*, pp. 411-415.
https://www.aclweb.org/anthology/R09-1074

[176] V. I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. Soviet Physics Doklady, volume 10, issue 8, pp. 707–710.
https://nymity.ch/sybilhunting/pdf/Levenshtein1966a.pdf

[177] Prasenjit Majumder, Mandar Mitra, and Dispasree Pal. Hungarian and Czech stemming using YASS. 2007. In *Working Notes for the CLEF 2007 Workshop*.
http://ceur-ws.org/Vol-1173/CLEF2007wn-adhoc-MajumderEt2007.pdf

[178] Jasmina Jovanovska, Ivana Bozhinova, and Katerina Zdravkova. 2016. Using NLP Methods to Improve the Effectiveness of a Macedonian Question Answering System. In *Proceedings of ICT Innovations 2015. Advances in Intelligent Systems and Computing*, pp. 205-214, Springer, Cham.
DOI: 10.1007/978-3-319-25733-4_21

[179] Tao Tao and ChengXiang Zhai. 2007. An exploration of proximity measures in information retrieval. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 295–302, ACM New York, NY, USA.
DOI: 10.1145/1277741.1277794

[180] Doug Beeferman, Adam Berger, and John Lafferty. 1997. A model of lexical attraction and repulsion. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th conference on European chapter of the Association for Computational Linguistics, pages 373–380, ACL, Stroudsburg, PA, USA.
DOI: 10.3115/976909.979665

[181] Ruihua Song, Michael J. Taylor, Ji-Rong Wen, Hsiao-Wuen Hon, and Yong Yu. 2008. Viewing term proximity from a different perspective. In *Proceedings of the IR research, 30th European conference on Advances in information retrieval*, pp. 346-357, Springer-Verlang, Berlin, Heidelberg.

[182] Ben He, Jimmy Xiangji Huang, Xiaofeng Zhou. 2011. Modeling term proximity for probabilistic information retrieval models. *International Journal of Information*

Sciences, volume 181, issue 14, pp. 3017-3031, Elsevier Science Inc. New York, NY, USA.
DOI: 10.1016/j.ins.2011.03.007

[183] Yuanhua Lv and ChengXiang Zhai. 2009. Positional language models for information retrieval. In *Proceedings 32nd international ACM SIGIR conference on Research and development in information retrieval*, pp. 299–306, ACM, New York, USA.
DOI: 10.1145/1571941.1571994

[184] Jinglei Zhao and Yeogirl Yun. 2009. A proximity language model for information retrieval. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pp. 291–298, ACM, New York, USA.
DOI: 10.1145/1571941.1571993

[185] Ronan Cummins and Colm O'Riordan. 2009. Learning in a pairwise term-term proximity framework for information retrieval. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 251–258, ACM, New York, USA.
DOI: 10.1145/1571941.1571986

[186] Olga Vechtomova and Ying Wang. 2006. A study of the effect of term proximity on query expansion. *Journal of Information Science*, volume 32, issue 4, pp. 324–333.

[187] Jun Miao, Jimmy Xiangji Huang, and Zheng Ye. 2012. Proximity-based rocchio's model for pseudo relevance. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pp. 535–544, ACM, New York, USA.
DOI: 10.1145/2348283.2348356

[188] T. Takaki and T. Kitani. 1999. Relevance Ranking of Documents Using Query Word Co-occurences (in Japanese). *Transactions of Information Processing Society of Japan*, volume 40, no. SIG 8, pp. 74–84.

[189] John Prager, Eric Brown, Anni Coden, and Dragomir Radev. 2000. Question Answering by Predictive Annotation. In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in the information retrieval, pp. 184–191, ACM New York, USA.
DOI: 10.1145/345508.345574

[190] Hideki Kozima and Teiji Furugori. 1994. Segmenting narrative text into coherent scenes. *Literary and Linguistic Computing*, volume 9, issue 1, pp. 13–19.
https://doi.org/10.1093/llc/9.1.13

[191] Owen de Kretser and Alistair Moffat. 1999. Effective document presentation with a locality-based similarity heuristic. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 113–120, ACM New York, USA.
DOI: 10.1145/312624.312664

[192] Koichi Kise, Markus Junker, Andreas Dengel, and Keinosuke Matsumoto. 2004. Passage Retrieval Based on Density Distributions of Terms and Its Applications to Document Retrieval and Question Answering. In *Reading and Learning. Lecture Notes in Computer Science*, volume 2956, Springer, Berlin, Heidelberg.
DOI: 10.1007/978-3-540-24642-8_17

[193] Cynthia J. Brame. 2013. Writing good multiple choice test questions. https://cft.vanderbilt.edu/guides-sub-pages/writing-good-multiple-choice-test-questions/

[194] Don R. Bacon. 2003. Assessing learning outcomes: A comparison of multiple-choice and short-answer questions in a marketing context. *Journal of Marketing Education,* volume 25, issue 1, pp. 31-36.
DOI: 10.1177/0273475302250570

[195] Catherin Walsh and Lisa Seldomridge. 2006. Critical thinking: Back to square two. *Journal of Nursing Education,* volume 45, issue 6, pp. 212-219.

[196] Benjamin S. Bloom. 1977. *Taxonomy of Educational Objectives*. New York: David McKay Company Inc.

[197] Edward J. Palmer and Peter G. Devitt. 2007. Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple-choice questions? *BMC Medical Education*, volume 7, pp. 49-55.
DOI: 10.1186/1472-6920-7-49

[198] Thomas M. Haladyna and Steven M. Downing. 1989. Validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, volume 2, issue 1, pp. 51-78.

[199] Michael Rodriguez. 1997. The art and science of item-writing: A meta-analysis of multiple-choice item format effects. *Paper presented at the annual meeting of the American Educational Research Association*, Chicago, IL.

[200] Stella Statman. 1988. Ask a clear question and get a clear answer: An enquiry into the question/answer and the sentence completion formats of multiple-choice items. *System,* volume 16, issue 3, pp. 367–376.

[201] Minu Ramakrishnan, Aditya B. Sathe, and Vinayak A. 2017. Item analysis: A tool to increase MCQ validity. *Indian Journal of Basic and Applied Medical Research*, volume 6, issue 3, pp. 67-71.
http://ijbamr.com/pdf/June 2017 67-71.pdf.pdf

[202] Swapnagandha S. Halikar, Veerendra Godbole, and Saurabh Chaudhari. Item Analysis to Assess Quality of MCQs. 2016. *Medical Science*, volume 6, issue 3, pp. 123-125.

[203] David DiBattista and Laura Kurzawa. 2011. Examination of the quality of multiple-choice items on classroom tests. *The Canadian Journal for the Scholarship of*

Teaching and Learning, volume 2, issue 2, article 4.
http://ir.lib.uwo.ca/cjsotl_rcacea/vol2/iss2/4

[204] Ralph F. Jozefowicz, Bruce M. Koeppen, Susan Case, Robert Galbraith, David Swanson, and Robert H. Glew. 2002. The quality of in-house medical examinations. *Academic Medicine*, volume 77, issue 2, pp. 156-161.

[205] Стефан Сидовски и Кирил Темков. 2005. Филозофија (Учебник за IV година на реформираното гимназиско образование). Просветно дело, Скопје.

[206] William B. Frakes and Ricardo Baeza Yates. 1992. *Information retrieval: Data structures and algorithms*. Prentice-Hall, Inc. Upper Saddle River, NJ, USA.

ДОДАТОК А – Таксономии дефинирани во различни истражувања, за различни тест-колекции

Во табелите А1 до А7 се дадени таксономии дефинирани во неколку истражувања, за различни тест-колекции.

Категорија на прашањето	Тип на одговор
кој, која, кое, кои (*who*)	личност, организација
каде (*where*)	локација
што (*what*)	пари, број, дефиниција, процедура, скратеница, организација, личност, година, месец, ден, време, локација
кога (*when*)	време, година, ден, месец
чиј, чија, чие, чии (*which*)	личност, локација, месец, време, година, ден
зошто (*why*)	причина
како (*how*)	процес

Табела А1. Таксономија на *Mudgal et al.* [15]

Категорија на прашањето	Тип на одговор
што (*what*)	пари, број, дефиниција, наслов, недефинирано
	личност, организација
	датум
	локација
кој, која, кое, кои (*who*)	личност, организација
како (*how*)	начин
	број
	време/растојание
	пари/цена
	недефинирано
каде (*where*)	локација
кога (*when*)	датум
чиј, чија, чие, чии (*which*)	личност
	локација
	датум
	организација
име	личност/организација
	локација
	титула
зошто (*why*)	причина
кому (*whom*)	личност/организација

Табела А2. Таксономија на *Moldovan et al*. [16]

Категорија на прашањето		Пример
Краток одговор	верификација	Дали одговорот е 5?
	дисјунктивност	Дали X или Y е променлива?
	завршување на концепт	Кој го изведе овој експеримент?
	спецификација на карактеристики	Кои се својствата на бар графикот?
	квантификација	Колку степени на слобода има оваа променлива?
Долг одговор	дефиниција	Што претставува t – тестот?
	пример	Наведете пример од факториел.
	споредба	Која е споредбата меѓу t – тестот и F – тестот?
	интерпретација	Што се случува на овој граф?
	причина	Зошто овој експеримент е неуспешен?
	последица	Што ќе се случи доколку ова ниво се намали?
	цел	Зошто ја поставивте латентноста на y – оската?
	инструмент/процедура	Како го презентирате стимулот за секој обид?
	овозможување	Кои уред овозможуваат мерење на стресот?
	очекување	Зошто нема интеракција?
	мислење	Што мислите за оваа оперативна дефиниција?
	изјава	Не ги разбирам главните ефекти.
	барање/наредба	Би ги собрал овие броеви?

Табела А3. Таксономија на *Graesser et al.* [17]

Груба категорија	Фина категорија
скратеница	акроним, скратен израз
ентитет	животно, тело, боја, креативност, валута, болест и медицина, настан, храна, инструмент, јазик, буква, растение, производ, религија, спорт, супстанција, симбол, техника, поим, возило, збор, друго
опис	дефиниција, опис, начин, исказ
човек	група, индивидуа, титула, состојба (на човекот)
локација	град, земја, планина, држава, друго
нумерички	код, број, датум, растојание, пари, редослед, период, процент, брзина, температура, големина, тежина, друго

Табела А4. Таксономија на *Xin et al.* [18]

Категорија	Примери
фактовидни информации	имиња, места, одделаченост, итн.
описни информации	распоред на летање, трошоци за транспорт
општопознато знаење	за секое патување, времето на пристигнување е поголемо од времето на тргнување
хиерархиско знаење	ресторанот се карактеризира по храна, локација, категорија, итн.
процедури и инструкции	како да се резревира соба во хотел
дефиниции	дефиниција на поими
регулативи и откажувања	наплата за задоцнето откажување
класификација	сортирање на хотелите по нивната категорија
интерпретација	скап хотел, далеку од плажа

Табела А5. Таксономија на *Benamara* [19]

Груба категорија	Прашањата се однесуваат на:
анатомија	одреден дел од телото кој е под влијание на одредена болест
причина	причини за болеста
компликација	ризици од одредено заболување
дијагноза	поставување на дијагноза, вклучувајќи тестови и методи
информација	општи информации за одредена болест
менаџмент	менаџмент, третман, лечење и превенција од болест
манифестација	симптоми од одредена болест
други ефекти	ефекти од болеста, исклучувајќи ги манифестацијата и компликациите
личност/организација	личности и организации запознаени со одредена болест
прогноза	животниот век, квалитетот на живот и веројатноста за успех на одреден третман
подложност	ширење и распространување на болеста
друго	нешто што не припаѓа на претходните категории
не е болест	нешто што не е болест
истражување	најнови истражувања и клинички тестирања

Табела А6. Таксономија на *Roberts et al*. [20]

Категорија	Опис
дефиниција	прашање кое содржи дефиниција на одговорот
релација	одговорот е во семантичка релација со прашањето, пришто релацијата е специфицирана од областа во која припаѓа прашањето
дополнување	прашање кое побарува дополнување на фраза
скратеница	одговорот е во форма на кратенка во прашањето
puzzle	одговорот побарува изведување, синтеза, заклучување, итн.
етимологија	прашање кое побарува англиски збор изведен од странски збор даден со значење
глагол	прашање кое побарува глагол како одговор
превод	прашање кое побарува превод на збор или фраза од еден во друг јазик
број	одговорот е број
поврзување	прашање кое побарува што е заедничко меѓу множество ентитети
прашање со повеќекратен избор	прашање кое содржи повеќе дадени одговори од кои треба да се избере точниот
датум	прашање кое побарува датум

Табела А7. Таксономија на *Watson* (утврдена на множество за тестирање од 3500 прашања) [105]

ДОДАТОК Б – Примери на прашања од македонската тест-колекција „Филозофија“, кои не ги задоволуваат стандардните протоколи за превалидација

Следниве примери на прашања од македонската тест-колекција „Филозофија“ не ги задоволуваат стандардните протоколи за превалидација. За секое прашање е наведен протоколот кој не е задоволен, како и неговата подобрена формулација. Притоа, со црвена боја е означен точниот одговор, од четирите понудени одговори.

Незначаен корен	***Подобрена формулација***
Која од следниве тези е вистинита:	Феноменологијата претставува наука за?
А. феноменологија е наука за состојбите во природата	А. состојбите во природата
Б. феноменологија е наука за убавите нешта	Б. убавите нешта
В. феноменологија е наука за појавите во свеста	**В. појавите во свеста**
Г. феноменологија е наука за принципите на добрата волја	Г. принципите на добрата волја

Пример Б1. Пример на незначаен корен и неговата подобрена формулација

Корен со ирелевантни информации	***Подобрена формулација***
Аристотел дошол на студии во Атина во Академијата на Платон како 18-годишно момче. Истакнувајќи се како најдобар студент, соработник и предавач, подоцна станува учител на:	Чиј учител бил Аристотел?
А. Александар Македонски	**А. на Александар Македонски**
Б. Филип Втори	Б. на Филип Втори
В. Сократ	В. на Сократ
Г. Ал-Фараби	Г. на Ал-Фараби

Пример Б2. Пример на корен со ирелевантни информации и неговата подобрена формулација

Корен со негација	***Подобрена формулација***
Кој НЕ е претставник на арапската филозофија?	Сите наведени се претставници на арапската филозофија освен:
А. Авицена	А. Авицена
Б. Ал-Фараби	Б. Ал-Фараби
В. Аверос	В. Аверос
Г. Ериугена	**Г. Ериугена**

Пример Б3. Пример на корен со негација и неговата подобрена формулација

Корен со внатрешно празно место

Заедно со Холбах, _____ е еден од најзначајните француски енциклопедисти.

А. Хелвециус

Б. Ками

В. Прудон

Г. Дерида

Подобрена формулација

Заедно со Холбах, кој е еден од најзначајните француски енциклопедисти?

А. Хелвециус

Б. Ками

В. Прудон

Г. Дерида

Пример Б4. Пример на корен со внатрешно празно место и неговата подобрена формулација

Неуверливи алтернативи (Б и Г)

Кој филозоф е највлијателен претставник на современиот интуиционизам?

А. Анри Бергсон

Б. Снупи

В. Норберт Винер

Г. Џон Кенеди

Пример Б5. Пример на неуверливи алтернативи за даден корен

ДОДАТОК В – Примери на прашања од македонската тест-колекција „Филозофија“ од секоја категорија, во согласност со ново-дефинираната таксономија

Во овој додаток се дадени примери на прашања од седумте различни фини категории, дефинирани во согласност со новата таксономија. Со црвена боја е означен точниот одговор, од четирите понудени одговори.

Прашање од категорија „Факт – личност“

Ал-Фараби своите истражувања од областа на логиката ги темели врз учењето на:

А. Горгија
Б. Платон
В. Сократ
Г. Аристотел

Прашање од категорија „Факт – ентитет“

Физиката на Епикур го следи учењето на**:**

А. стоиците
Б. питагорејците
В. атомизмот
Г. Елејската школа

Прашање од категорија „Опис“

Иморализмот на Ниче подразбира:

А. возвишување на егалитаристичките вредности
Б. почитување на актуелните морални вредности на општеството
В. зајакнување на демократските вредности
Г. отфрлање на наметнатите вредности

Прашање од категорија „Исказ“

Најпознатата етичка порака на Свети Климент гласи:

А. биди позитивен

Б. да се надминеме во добрина

В. прави добрини
Г. тргај се од злото и прави добро

Прашање од категорија „Дефиниција“

Како се вика филозофскиот правец според кој сето познание произлегува од искуството?

А. рационализам
Б. експресионизам
В. позитивизам
Г. емпиризам

Прашање од категорија „Набројување на факти – личности“

Кои се првите словенски просветители?

А. Светите Климент и Константин
Б. Светите Кирил и Методиј
В. Светите Климент и Методиј
Г. Светите Наум и Константин

Прашање од категорија „Набројување на факти - ентитети“

Според егзистенцијализмот, за човекот се карактеристични:

А. мудроста и безгрижноста
Б. стравот и неизвесноста
В. веселоста и генијалноста
Г. несвесноста и загубеноста

ДОДАТОК Г – Примери на соодветни/несоодветни групи (кластери) од збороформи, генерирани со примена на *Dice*-метриката

Во следниов дел се дадени неколку примери на соодветни/несоодветни кластери, генерирани со вкучување на *Dice*-метриката со праг 0.5, со цел автоматско групирање на збороформите од Речникот_1 од тест-колекцијата „Филозофија“.

Соодветни кластери:

ncm: ['компјутер', 'компјутери', 'компјутерите', 'компјутерот']
ngn: ['страдања', 'страдањата', 'страдање', 'страдањето']
ncf: ['манипулации', 'манипулациите', 'манипулација']
v: ['зборува', 'зборуваш', 'зборувај', 'зборуваа', 'зборуваат', 'зборувал', 'зборувале', 'зборувало', 'зборуваме']
a: ['дотогашен', 'дотогашната', 'дотогашни', 'дотогашниот', 'дотогашните']
r: ['максимално']
m: ['две', 'двеве', 'двете']

Несоодветни кластери:

ncm: ['прозорци', 'прозорците', 'пророци', 'пророците']
ngn: ['продолжување', 'пролевање', 'промислување', 'појавување', 'простување', 'простувањето', 'прочистување', 'просудување', 'просудувањето', 'проучување', 'проучувањето']
ncf: ['страна', 'страната', 'странка', 'страни', 'страните']
v: ['избрал', 'изброкале', 'изброи']
a: ['конституирано', 'континуиран', 'континуирани', 'континуирано']
r: ['разбудувајќи', 'разликувајќи', 'размислувајќи', 'развивајќи']

ДОДАТОК Д – Примери на прашања од македонската тест-колекција „Филозофија“ за кои *QA*-системот не прибира точен пасус

Следниов дел содржи две прашања од тест-колекцијата „Филозофија“ за кои дизајнираниот *QA*-систем не прибира точен пасус (со примена на точните кластери од збороформи), заради незастапеноста на одредени клучни зборови од прашањата во Речникот_1 за тест-колекцијата „Филозофија“.

Прашање од категорија „Факт – личност“

Кој од филозофите емпирист ги застапувал скептистичките ставови?
А. Џорџ Беркли
Б. Дејвид Хјум
В. Џон Лок
Г. Френсис Бекон

Зборовите „емпирист“ и „скептистиките“ не се содржат во Речникот_1 од тест-колекцијата „Филозофија“.
Системот прибира пасуси користејќи прашалник кој ги содржи само зборовите „филозофите“, „застапувал“ и „ставови“ (вклучително и нивните збороформи).

Прашање од категорија „Опис – исказ“

Според егзистенцијалистичката филозофија на Жан-Пол Сартр, човекот е:
А. осуден да биде роб
Б. осуден да биде среќен
В. осуден да биде слободен
Г. осуден да биде сакан

Зборот „егзистенцијалистичката“ не е вклучен во Речникот_1 од тест-колекцијата „Филозофија“.
Системот прибира пасуси користејќи прашалник кој ги содржи само зборовите „филозофија“, „Жан-Пол Сартр“ и „човекот“ (вклучително и нивните збороформи).

ДОДАТОК Ѓ – Проширување на прашалникот во случај на непознат збор

Во табелите Ѓ1 и Ѓ2 се дадени примери на **непознати** зборови (зборови кои не се застапени во Речникот_1 од тест-колекцијата „Филозофија“) за кои е точно/погрешно утврден **сличен** збор. Дадените резултати се добиени со примена на метриките базирани на коефициентите *Dice*, *Positional Dice* и *Jaccard* за прагот 0.35 (табела Ѓ1) и метриката базирана на триаголниот прозорец за прагот 1.5 (табела Ѓ2). Со црвена боја се означени погрешно утврдените **слични** зборови, додека *NONE* означува дека не е утврден сличен збор за наведените прагови.

Збор	*Dice*	*Pos_Dice*	*Jaccard*
австриски	**австралиска**	австрија	NONE
бесплатно	NONE	NONE	NONE
демократска	демократски	демократски	демократски
елитна	елита	елита	NONE
исихистите	**истите**	NONE	NONE
лавот	лав	лав	NONE
мизантропијата	мизантроп	мизантроп	NONE
поплавите	**појавите**	NONE	NONE
прекрасне	прекрасна	прекрасна	прекрасна
ураганите	NONE	NONE	NONE

Табела Ѓ1. Дел од непознатите зборови со најсличниот збор добиен со метриките базирани на коефициентите *Dice*, *Positional Dice* и *Jaccard*, за прагот 0.35

Збор	*Triangle*
австриски	австрија
бесплатно	**бесплодна**
демократска	демократски
елитна	елита
исихистите	исихизам
лавот	лав
мизантропијата	мизантроп
поплавите	NONE
прекрасне	прекрасна
ураганите	NONE

Табела Ѓ2. Дел од непознатите зборови со најсличниот збор добиен со метриката базирана на триаголниот прозорец, за прагот 1.5

ДОДАТОК Е – Точност во *IR*-фазата со инкорпорирање на метрики за сличност на зборови со цел проширување на прашалниците

Табелите Е1, Е2 и Е3 ги презентираат резултатите добиени во процесот на прибирање пасуси со примена на точните кластери и кластерите генерирани со метриката базирана на триаголниот прозорец, за праговите 0.5, 1.5 и 2.5, како и проширување на прашалникот со метриките базирани на коефициентите *Dice*, *Positional Dice* и *Jaccard*, за повеќе вредности на праговите, соодветно. Со црвена боја е означена највисоката постигната точност на системот, за секоја од метриките.

*Dice*_Праг	Точни кластери	*Triangle_0.5*	*Triangle_1.5*	*Triangle_2.5*
0.10	0.918	0.926	0.926	0.926
0.15	0.918	0.926	0.926	0.926
0.20	0.922	0.931	0.931	0.931
0.25	0.926	**0.935**	**0.935**	0.931
0.30	0.926	**0.935**	**0.935**	0.931
0.35	0.931	**0.935**	**0.935**	0.931
0.40	0.931	**0.935**	**0.935**	0.931
0.45	0.931	**0.935**	**0.935**	0.931

Табела Е1. Постигната точност во фазата за прибирање пасуси, со примена на точните кластери и кластерите генерирани со метриката за сличност базирана на триаголниот прозорец, и проширување на прашалникот со *Dice*-метриката

*Pos_Dice*_Праг	Точни кластери	*Triangle_0.5*	*Triangle_1.5*	*Triangle_2.5*
0.10	0.918	0.926	0.926	0.926
0.15	0.918	0.926	0.926	0.926
0.20	0.922	0.931	0.931	0.931
0.25	0.926	0.935	0.935	0.931
0.30	0.926	0.935	0.935	0.931
0.35	0.931	0.935	0.935	0.931
0.40	0.931	0.935	**0.939**	0.935
0.45	0.931	0.935	**0.939**	0.935

Табела Е2. Постигната точност во фазата за прибирање пасуси, со примена на точните кластери и кластерите генерирани со метриката за сличност базирана на триаголниот прозорец, и проширување на прашалникот со *Positional Dice*-метриката

*Jaccard*_Праг	Точни кластери	*Triangle_0.5*	*Triangle_1.5*	*Triangle_2.5*
0.10	0.918	0.926	0.926	0.926
0.15	0.918	0.926	0.926	0.926
0.20	0.918	0.926	0.926	0.926
0.25	0.918	0.926	0.926	0.922
0.30	0.922	0.931	0.931	0.926
0.35	0.922	0.931	0.931	0.926
0.40	0.926	0.935	**0.935**	0.926
0.45	0.926	0.935	**0.935**	0.926

Табела Е3. Постигната точност во фазата за прибирање пасуси, со примена на точните кластери и кластерите генерирани со метриката за сличност базирана на триаголниот прозорец, и проширување на прашалникот со *Jaccard*-метриката

ДОДАТОК Ж – Точност на *QA*-системот со примена на *Hanning*-прозорската функција во фазата за селекција на точниот одговор

Во табелата Ж1 се дадени постигнувањата на *Hanning*-прозорецот применет врз петте најдобро рангирани пасуси од *IR*-фазата. Пасусите се добиени со вклучување на кластерите од збороформи генерирани со метриката за сличност базирана на **триаголниот прозорец** за праговите 0.5 и 2.5, и проширување на прашалникот со истата метрика за прагот 1.5. Притоа, при имплементација на *Hanning*-прозорецот не се искористени збороформите на зборовите од четирите прашалници и не е направено проширување на прашалникот во случај на непознат збор. Со црвена боја е означена највисоката постигната точност за соодветната комбинација.

w	**Пасуси добиени во *IR* со примена на кластери од збороформи генерирани со *Triangle_2.5* и проширување на прашалнник со *Triangle_1.5***		**Пасуси добиени во *IR* со примена на кластери од збороформи генерирани со *Triangle_0.5* и проширување на прашалнник со *Triangle_1.5***	
	# на точно одговорени	*QA*-точност (%)	# на точно одговорени	*QA*-точност (%)
4	169	73.16	164	71.00
6	176	76.19	174	75.32
10	**180**	**77.92**	177	76.62
16	178	77.06	**180**	**77.92**
20	178	77.06	**180**	**77.92**
26	176	76.19	177	76.62

Табела Ж1. Постигната точност во фазата за селекција на точниот одговор, со примена на *Hanning*-прозорецот, без збороформи и без проширување на прашалникот

Во табелата Ж2 се дадени постигнувањата на *Hanning*-прозорецот применет врз петте најдобро рангирани пасуси од *IR*-фазата, прибрани со примена на кластерите од збороформи генерирани со метриката за сличност базирана на **триаголниот прозорец** за прагот 1.5 и проширување на прашалникот со истата метрика за прагот 1.5. Притоа, при имплементација на *Hanning*-прозорецот се искористени збороформите на зборовите од четирите прашалници (генерирани со *Triangle*-метриката за праг 1.5), а не е направено проширување на прашалникот во случај на непознат збор.

w	***Hanning*-прозорец со примена на кластери од збороформи генерирани со *Triangle_1.5*, без проширување на прашалнникот**	
	# на точно одговорени	*QA*-точност (%)
4	167	72.29
6	185	80.09
10	187	80.95
16	185	80.09
20	**190**	**82.25**
26	185	80.09

Табела Ж2. Постигната точност во фазата за селекција на точен одговор, со примена на *Hanning*-прозорецот, со вклучување на збороформи, а без проширување на прашалникот

ДОДАТОК 3 – Примери на прашања од македонската тест-колекција по „Информатички технологии“

Во продолжение се дадени примери на прашања од македонската тест-колекција по „Информатички технологии“ од секоја категорија, во согласност со новодефинираната таксономија. Со црвена боја е означен точниот одговор.

Прашање од категоријата „Факт – личност“

Кој ги поставил основите на пишаното сметање?

А. Архимед
Б. Питагора
В. Боетије
Г. Ал Хорезми

Прашање од категоријата „Факт – ентитет“

Кој е прв општонаменски, автоматски, електро-механички калкулатор?

А. Марк-I
Б. IBM 360
В. ABC
Г. ENIAC

Прашање од категоријата „Опис“

Што е основната разлика меѓу компјутерите и калкулаторите?

А. компјутерите се поголеми и имаат екран и тастатура
Б. компјутерите се повеќенаменски и можат да извршуваат истовремено повеќе задачи
В. компјутерите можат да се програмираат
Г. не постои разлика, но компјутерите се поскапи

Прашање од категоријата „Дефиниција“

Што е вештачка интелигенција?

А. постапка на замена на природниот интелект со вештачки
Б. наука што создава системи коишто извршуваат умствени задачи
В. наука што им помага на луѓето да си го зголемат интелектот
Г. способност на сметачите да се однесуваат интелигентно

Прашање од категоријата „Набројување на факти - личности“

Кој го измислил интегралното коло?

А. Џек Килби и Роберт Нојс
Б. Вилијам Шокли
В. Грејс Мари Хопер
Г. Бил Гејтс

Прашање од категоријата „Набројување на факти - ентитети“

Најновите трендови на информатиката се:

А. мрежите, мултимедијалноста и вештачката интелигенција
Б. мрежите, интерактивноста и говорното избирање
В. Интернет и мултимедијалноста
Г. Интернет, експертните системи и образовниот софтвер

ДОДАТОК S – Примери на прашања од англиската тест-колекција по „Информатички технологии“

Во продолжение се дадени примери на прашања од англиската тест-колекција по „Информатички технологии“ од секоја категорија, во согласност со новодефинираната таксономија. Со црвена боја е означен точниот одговор.

Прашање од категоријата „Факт – личност“

Who invented the vacuum tubes triode?

A. John Bardeen
B. William Shockley
C. Lee de Forest
D. John Bardeen

Прашање од категоријата „Факт – ентитет“

What was the expected feature of the fifth generation computers when Japan started FGCS?

A. Operating Systems
B. Parallel Processing
C. ULSI
D. Bio-Chips

Прашање од категоријата „Опис“

Why ABC computer is called so?

A. Because it was developed by Atanasoff and Berry
B. Because it was named with first alphabets of English
C. The inventor chose that name randomly
D. No reason

Прашање од категоријата „Дефиниција“

What is an interpreter?

A. Language processor that converts one statement of a program at a time
Б. Representation of the system being designed
B. General-purpose language providing very efficient execution
Г. Language that allows the user to tell the computer what to do

Прашање од категоријата „Набројување на факти - ентитети“

Which are classified as low-level programming languages?

A. Basic, COBOL, FORTRAN
B. Prolog 2, Expert Systems
C. Knowledge based Systems
D. Assembly, Machine language

ДОДАТОК И – Распределба на англиските термини од македонската тест-колекција по „Информатички технологии“ по зборовни групи

Во табелата И1 е дадена распределбата на англиските термини од колекцијата МакИнфо по зборовни групи.

Ознака за зборовна група	Број	Значење на ознаката
np	155	лична именка без означен род
npm	62	лична именка од машки род
npf	4	лична именка од женски род
n	240	општа именка
v	3	глагол
a	90	придавка
c	1	сврзник
p	7	предлог
d	1	член
m$	5	број запишан со цифри
m(буква)	3	број запишан со букви
unknown	23	непознат збор
abb	23	скратеница
Вкупно	**617**	

Табела И1. Распределба на англиските термините од колекцијата МакИнфо по зборовни групи

ДОДАТОК Ј – Распределба по зборовни групи на зборовите од англиската тест-колекција по „Информатички технологии“, кои имаат повеќе од една ознака

Во табелата Ј1 е дадена распределбата по зборовни групи на зборовите од Речникот_1_АнгИнфо, кои имаат повеќе од една ознака.

Ознака за зборовна група	Број	Пример
n, m	1	1990s
n, v	90	study
np, n	4	mark
n, a	35	complex
v, a	16	open
n, r	3	back
a, r	16	longer
np, a	3	basic
n, v, a	1	control
v, a, r	1	close
Вкупно	**170**	

Табела Ј1. Распределба на англиските зборови со повеќе ознаки од колекцијата АнгИнфо по зборовни групи

ДОДАТОК К – Сегмент од резултатот добиен со примена на *Hanning*-прозорската функција, за англиската тест-колекција по „Информатички технологии“

Во овој додаток е даден дел од резултатот добиен за прашањето 2 од тест-колекцијата АнгИнфо, со примена на *Hanning*-прозорската функција за $w = 36$.

Прашање 2 од тест-колекцијата АнгИнфо

Programs designed to perform specific tasks are known as

A. system software

B. application software

C. utility programs

D. operating system

Четвртиот понуден одговор го креира следниов прашалник:

***QUERY* 4:** ['programs', 'design', 'perform', 'special', 'tasks', 'is', 'know', 'operating', 'system']

Следи параграф од колекцијата документи означен како 107, кој системот го прибира во *IR*-фазата како еден од петте најдобро рангирани пасуси.

***PARAGRAPH*: 107**

provides the basic non-task-specific functions of the computer and application software which is used by users to accomplish specific tasks wikipedia in other words software is a set of programs procedures algorithms and its documentation concerned with the operation of a data processing system thus the software contains the instructions that tell a computer what to do and how to do to solve a specific problem in general use by software we understand a group of programs to make a system run a program contains instructions or commands to perform a task the term package or suite is used to describe a group of related software types of software software is generally classified into two groups – system software and application software some people prefer three types system application and utilities system software system software is responsible for controlling integrating and managing the individual hardware components of a computer system so that other software and the users of the system see it as a functional unit without having to be concerned with the low-level details such as transferring data from memory to disk or rendering text onto a display generally system software consists of an operating system and some fundamental utilities such as disk formatters file managers display managers text editors user authentication login and management tools and networking and device control software there are three type of software under system software operating systems language language processors and utilities operating system an operating system os is a set of software that manages computer hardware resources and provides common services for computer programs the operating system is a vital component of the system software in a computer system application programs require an operating system to function the operating system is the most important program that runs on a computer every general-purpose

Прашалникот ***QUERY* 4** и ***PARAGRAPH*: 107** ги имаат следниве заеднички зборови:

***Intersection with the answer*:** ['operating', 'system']

Зборот *„operating“* се наоѓа на следниве позиции во ***PARAGRAPH*: 107**:

> ***WORD*:** operating
>
> ***POSITIONS*:** [130, 159, 165, 168, 183, 197, 200]
>
> ***WORDS IN THE WINDOW [179 - 207]*:** ['common', 'services', 'computer', 'programs', 'operating', 'system', 'is', 'vital', 'component', 'system', 'software', 'in', 'computer', 'system', 'application', 'programs', 'require', 'an', 'operating', 'system', 'function', 'operating', 'system', 'is', 'most', 'important', 'program', 'runs']
>
> ***INTERSECTION WINDOW - QUERY*:** ['programs', 'operating', 'system', 'is', 'system', 'system', 'programs', 'operating', 'system', 'operating', 'system', 'is']

Зборот *'system'* се појавува петпати во пресекот на прозорецот [179 – 207] и прашалникот ***QUERY* 4**. Тежините на овие зборови добиени со примена на *Hanning*-прозорецот се следни: 0.17861, 0.5, 0.82139, 0.99240, и 0.88302. Направената модификација на **густината на распределба** (*DD*) базирана на *Hanning*-прозорската функција, се однесува на вклучување на зборот *'system'* само еднаш и тоа со највисоката добиена вредност, во случајов 0.99240. Истата модификација се прави за секој збор кој се повторува во ***INTERSECTION WINDOW – QUERY***.

ДОДАТОК Л – Причини за намалување на точноста на *QA*-системот, при одговарање на прашањата од македонската тест-колекција по „Информатички технологии“

1. Пример на прашање каде одговорот е запишан на македонски јазик, додека во изворниот документ, одговорот е запишан на англиски јазик.

Прашање од категоријата „Факт – личност“

Кој е татко на електронските сметачи?

А. Чарлс Бебиџ

Б. Херман Холерит

В. Џон Атанасоф

Г. Конрад Цузе

Во 1939 година, во САД, John Vincent Atanasoff, познат како татко на електронските сметачи, го развива системот наречен ABC, наменет за решавање системи линеарни равенки. Тоа е првиот електричен калкулатор со електронска аритметичка единица.

Слика Л1. Сегмент од документот „Историјат на сметачите“ каде се наоѓа одговорот на наведеното прашање

2. Пример на прашање каде како одговор се побарува факт (време). Во понудените одговори времето е наведено со векови, додека во изворниот документ со години.

Прашање од категоријата „Факт – ентитет“

Кога се измислени машините на Бебиџ?

А. на почетокот на XVII век

Б. на почетокот на XVIII век

В. на крајот на XVII век

Г. на почетокот на XIX век

(accumulators), се собирале меѓурезултатите. Операциите кај диференцната машина било замислено да се извршуваат паралелно, што е еден од најмодерните трендови во современата информатика. Во 1824 Бебиџ направил негов прототип (сл. 7.) што пресметувал квадрати и кубови од шестцифрените броеви и квадратни полиноми. За жал, овој генијален изум не бил направен ниту во повторниот обид на Бебиџ, во периодот меѓу 1847 и 1849 година, пред сè заради недоволно развиените технички можности во тоа време.

Слика Л2. Сегмент од документот „Историјат на сметачите“ каде се наоѓа одговорот на наведеното прашање

3. Примери прашања кои не ги задоволуваат стандардните протоколи за превалидација.

Прашање од категоријата „Опис“, во кое коренот е искажан со негација (како и алтернативите)

Зошто не биле конструирани машините на Бебиџ?

А. струјата не била откриена

Б. не постоела документација

В. не постоеле технички услови

Г. не била користена бинарната аритметика

Прашање од категоријата „Опис“, во кое алтернативите се со значително различна должина

Од кои бои се формира една боја на екранот?

А. црвена, зелена, сина

Б. нијансите на трите основни бои и црната, таканаречено CMYK

В. стотици нијанси на RGB (црвена, зелена и сина)

Г. зависи од резолуцијата на екранот

4. Пример на прашање кое содржи само еден клучен збор.

Прашање од категоријата „Дефиниција“

Што е меморија?

А. склад за податоците и програмските инструкции

Б. место за сместување на датотеките

В. мемориски медиум со хиерархиска структура

Г. внатрешна (ROM и RAM) и надворешна (диск, дискета, CD и DVD)

5. Пример на прашање чиј одговор не може да се пронајде во документите од колекцијата.

Прашање од категоријата „Опис“

Кој е најзначајниот механички изум што го овозможил ширењето на информациите?

А. појавата на писмото пред 50 века

Б. печатарската преса на Гутенберг од средината на XV век

В. појавата на сметачите од средината на XX век

Г. појавата на Интернет од крајот на XX век

ДОДАТОК ЉЉ – Дел од примената на дистрибутивен метод за утврдување на зборови кои се наоѓаат заедно во текстуални сегменти

Во табелата ЉЉ1 е даден мал дел од резултатите добиени со примена на дистрибутивен метод, со цел наоѓање на зборови кои се појавуваат заедно во текстуални сегменти (*co-occurring words*). Методот е применет врз проширената тест-колекција на македонски јазик МакИнфо. При тоа, во табелата се означени само највисоките десет добиени вредности за **сличност**, за сите парови зборови од речникот на колекцијата.

Збор	*Co-occuring word*	Сличност
науки	факултет	**10.32816049**
	физиката	1.70214005
	традиционалниот	1.392233701
оперативниот	систем	**9.193583735**
	системи	4.766984363
	може	4.336874896
системи	оперативни	**8.329859332**
	систем	6.797438775
	користат	4.27534502
број	голем	**7.077940279**
	инструкции	5.176443824
	имаат	3.676242609
бидат	можат	**6.523685999**
	може	4.233127184
	многу	3.625924802
извршување	времето	**6.249618224**
	програми	2.883346701
	систем	2.685219783
технологии	апликации	**5.871888528**
	поглавје	5.8707256
	специјални	1.938080239
компјутери	персонални	**5.810344828**
	персоналните	4.084151236
	користат	3.882182845
плоча	матичната	**5.173324498**
	матична	4.222966269
	има	2.933812351
инструкции	бројот	**5.10366281**
	компјутерот	3.64406423
	извршени	3.235048777

Табела ЉЉ1. Десет збора со соодветните зборови кои веројатно се појавуваат во исти сегменти во проширената колекција МакИнфо

Пример: Со цел проширување на прашалникот генериран за прашањето дадено во додатокот Л (во делот 4), може да се искористат некои од зборовите со кои зборот „меморија“ (или неговите зборoформи) има највисока сличност, применувајќи го диструбуцискиот метод (табела Љ2).

Збор	*Co-occuring word*	Сличност
меморија	податоците	**4.12918239**
	единици	3.05295969
	главна	2.71495243
меморијата	меморија	**4.66234249**
	податоците	3.56755637
	пристап	3.28656648
мемориите	рам	**1.40952453**
	флеш	1.40600775
	пазарот	1.40249097
мемории	меморија	**3.65026968**
	единици	3.24032223
	податоците	3.1145842

Табела Љ2. Проширување на прашалникот со примена на дистрибутивен метод

ДОДАТОК М – Делови од кодот со кои се реализирани најкарактеристичните функции на системот за одговарање прашања

Во додатокот М се дадени делови од кодот со кои се извршуваат најзначајните функции на системот за одговарање прашања.

```
Def cosine_similarity(doc1, doc2):
    """

    :param doc1: a dictionary where the keys are the words in the first document;
the values are the corresponding
    tf-idf calculations
    :param doc2: a dictionary where the keys are the words in the second document;
the values are the corresponding
    tf-idf calculations
    :return: the cosine similarity between the two documents
    """
    intersection = set(doc1.keys()) & set(doc2.keys())
    numerator = sum([doc1[x] * doc2[x] for x in intersection])
    sum1 = sum([doc1[x] ** 2 for x in doc1.keys()])
    sum2 = sum([doc2[x] ** 2 for x in doc2.keys()])
    denominator = math.sqrt(sum1) * math.sqrt(sum2)
    if not denominator == 0:
        return float(numerator) / denominator
    return 0
```

Слика М1. Пресметување на косинус сличноста меѓу два документа

```
def tf_idf_clusters(query, documents_list, n, clusters):
    weights = {}
    for word in query:
        tf = count_appearances(word, query, clusters)
        if not tf == 0:
            tf = 1 + math.log(tf)
        value = 1 + n / (1 + documents_containing_word_from_cluster(word,
documents_list, clusters))

        idf = math.log(value)
        result = tf * idf
        if result != 0:
            weights[str(word)] = result
    return weights
```

Слика М2. Пресметување на косинус сличноста меѓу два документа

```
def binary(query, document):
    weights = {}
    for word in query:
        if word in document:
            weights[str(word)] = 1
        else:
            weights[str(word)] = 0
    return weights
```

Слика М3. Пресметување на сличноста меѓу прашалникот и документот користејќи го совпаѓањето на координатите

```
class Dice(WordDistance):
    def distance(self, word1, word2, n):
        """

        Calculates the dice similarity between two words
        :param word1: the first word
        :param word2: the second word
        :param n: the length of n-grams used to calculate the coefficient
        :return:
        """

        if not helpers.same_ngram_prefix(word1, word2, n):
            return 0

        n_grams_word1 = helpers.generate_n_grams(word1, n)
        n_grams_word2 = helpers.generate_n_grams(word2, n)
        intersection_n_grams = get_n_gram_intersection(n_grams_word1,
n_grams_word2)

        if not (len(n_grams_word1) + len(n_grams_word2)) == 0:
            return 1 - (2 * (len(intersection_n_grams)) / (len(n_grams_word1) +
len(n_grams_word2)))

        return 0
```

Слика М4. Метрика за сличност на стрингови базирана на *Dice*-коефициентот

```
class PositionalDice(WordDistance):
    def distance(self, word1, word2, n):
        """

        Positional dice metric for word similarity
        :param word1: the first word
        :param word2: the second word
        :param n: the value for the n-grams
        :return: the similarity between word1 and word2
        Example:
            Positional Dice: (па,1), (ап,2), (пи,3), (ит,4), (те,5);
            патепис - (па,1), (ат,2), (те,3), (еп,4), (пи,5), (ис,6);
            Pos_Dice=1-(1/11)=0.91.
            Се поклопуваат само во првиот подреден пар.
        """
        if not helpers.same_ngram_prefix(word1, word2, n):
            return 0

        n_grams_word1 = helpers.generate_positional_n_grams(word1, n)
        n_grams_word2 = helpers.generate_positional_n_grams(word2, n)
        intersection_n_grams = get_n_gram_intersection(n_grams_word1,
n_grams_word2)

        if not (len(n_grams_word1) + len(n_grams_word2)) == 0:
            return 1 - (2 * (len(intersection_n_grams)) / (len(n_grams_word1) +
len(n_grams_word2)))

        return 0
```

Слика М5. Метрика за сличност на стрингови базирана на *Positional Dice*-коефициентот

```
def generate_n_grams(word, n):
    """

    Generates all substrings of length n
    :param word: the input word
    :param n: the length of the substrings
    :return:
    """
    n_grams = []
    for i in range(len(word) - 1):
        n_grams.append(word[0:n])
        word = word[1:]

    return n_grams
```

Слика М6. Помошен метод за генерирање на n – грамите на даден збор

```
def generate_positional_n_grams(word, n):
    """

    Generates all substrings of length n with their position in the word
    :param word: the input word
    :param n: the length of the substrings
    :return:
    """
    n_grams = []
    for positional_n_gram in range(len(word) - 1):
        n_grams.append((word[0:n], positional_n_gram))
        word = word[1:]

    return n_grams
```

Слика М7. Помошен метод за генерирање на позициските n – грами на даден збор

```
class Jaccard(WordDistance):
    def distance(self, word1, word2, n):
        """

        Calculates the dice similarity between two words
        :param word1: the first word
        :param word2: the second word
        :param n: the length of n-grams used to calculate the coefficient
        :return:
        """
        if not helpers.same_ngram_prefix(word1, word2, n):
            return 0

        n_grams_word1 = helpers.generate_n_grams(word1, n)
        n_grams_word2 = helpers.generate_n_grams(word2, n)
        intersection_n_grams = get_n_gram_intersection(n_grams_word1,
n_grams_word2)

        intersection_n_grams_len = len(intersection_n_grams)

        if not intersection_n_grams_len == 0:
            return 1 - (intersection_n_grams_len /
                        (len(n_grams_word1) + len(n_grams_word2) -
intersection_n_grams_len))
        return 0
```

Слика М8. Метрика за сличност на стрингови базирана на *Jaccard*-коефициентот

```
class Triangle(WordDistance):
    def distance(self, word1, word2, n):
        """

        Calculates the triangle similarity between two strings
        - make sure that word1 is always the longer word
        :param n:
        :param word1: the first word
        :param word2: the second word
        :return:
        """
        if not helpers.same_ngram_prefix(word1, word2, n):
            return 0

        difference_index = self.index_of_first_different_letter(word1, word2)
        return (len(word1) - difference_index) / (difference_index - 1)
```

Слика М9. Метрика за сличност на стрингови базирана на триаголниот прозорец

```
def expand_query(question_query, similarity_metric):
    unknown = query_expansion.find_unknown_words_in_question(question_query)
    expanded = [query_expansion.match(similarity_metric, w) for w in unknown]
    question_query += expanded

    return question_query
```

Слика М10. Проширување на прашалникот во случај на непознат збор со примена на метриката базирана на триаголниот прозорец

```
WINDOW = 26
X_VALUES = [i for i in range(int(math.floor(- WINDOW / 2)), int(math.ceil(WINDOW /
2) + 1))]
WORD_DISTANCE = int(WINDOW / 2)
X = 0.5
FACTOR = 0.05
N = 2


def question_answering():
    correct = 0
    for question_id in all_questions.keys():

        question = all_questions[question_id]
        answers = all_answers[question_id]

        top_values = [0] * 4
        top_paragraphs = [0] * 4

        for answer_idx in range(len(answers)):
            answer = answers[answer_idx]
            query = question + answer

            for paragraph_id in paragraphs_for_question[question_id]:
                paragraph = all_paragraphs[paragraph_id]
                intersection_answer_paragraph = [w for w in answer if w in
paragraph]
                for word in intersection_answer_paragraph:
                    word_positions = inverted_index[paragraph_id][word]
                    for position in word_positions:
                        window_start = max(position + X_VALUES[0], 0)
                        window_end = min(position - X_VALUES[0] + 1,
len(paragraph))
                        words_in_window = [paragraph[j] for j in
range(window_start, window_end)]
                        intersection_window_query = [w for w in words_in_window if
w in query]
                        result = 0
                        for x in X_VALUES:
                            if x + WORD_DISTANCE < len(words_in_window):
```

```
                                if words_in_window[x + WORD_DISTANCE] in
intersection_window_query:
                                    f = hanning_window_function(x, WINDOW)
                                else:
                                    f = 0
                                b =
metrics.get_pos_specific_weight(words_in_window[x + WORD_DISTANCE], X, FACTOR)
                                result += f * b
                            if result >= top_values[answer_idx]:
                                top_values[answer_idx] = result
                                top_paragraphs[answer_idx] = paragraph_id

        max_index, max_value = max(enumerate(top_values),
key=operator.itemgetter(1))
        if max_index == correct_answers[question_id]:
            correct = correct + 1


def hanning_window_function(x, w):
    if abs(x) <= (w / 2):
        return (1 + math.cos(2 * math.pi * x / w)) / 2
    return 0



def get_pos_specific_weight(word, x, factor):
    pos_tag = pos_tagging.get_pos_tag(word)[:N]

    """

    Personal Noun -                PN=x
    Noun -                         N=(x-a)
    Number -                       M=x-3a
    Adjective -                    A=x-2a
    Adverb -                       R=x-3a
    Verb -                         V=x-2a

    """
    if pos_tag == dictionary.constants.NOUN_PERSONAL:
        return x
    pos_tag = pos_tag[:1]
    if pos_tag == dictionary.constants.NOUN:
        return x - factor
    if pos_tag == dictionary.constants.NUMBER:
        return x - (3 * factor)
    if pos_tag == dictionary.constants.ADJECTIVE:
        return x - (2 * factor)
    if pos_tag == dictionary.constants.ADVERB:
        return x - (3 * factor)
    if pos_tag == dictionary.constants.VERB:
        return x - (2 * factor)
    return 0
```

Слика М11. Селекција на точниот одговор за дадено прашање со примена на *Hanning*-прозорската функција

ОБЈАВЕНИ ТРУДОВИ:

1. Jasmina Armenska, Aleksandar Tomovski, Katerina Zdravkova, and Jovan Pehcevski. 2011. Information retrieval using a Macedonian test collection for question answering. In *International Conference on ICT Innovations*, volume 83, pp. 205-214, Springer, Berlin, Heidelberg.
   DOI: 10.1007/978-3-642-19325-5_21
2. Jasmina Armenska and Katerina Zdravkova. 2012. Comparison of Information Retrieval Models for Question Answering. In *Proceedings of the Fifth Balkan Conference in Informatics*, pp. 162-167, ACM.
   DOI: 10.1145/2371316.2371348
3. Jasmina Armenska, Nace Stojanov, and Goce Armenski. 2014. Students' opportunities to use ICT during the teaching process and their computer skills. In *Proceedings of the 9th International Balkans Education and Science Congress*, pp. 312-317. Trakya University - Edirne, Turkey.
4. Stevica Bozhinoski, Ivana Bozhinova, and Jasmina Armenska. Information retrieval over Macedonian test collection using word forms and transliteration. *The 12th International Conference on Informatics and Information Technologies*, 23-26 April 2015, Bitola.
5. Jasmina Jovanovska, Ivana Bozhinova, and Katerina Zdravkova. 2016. Using NLP methods to improve the effectiveness of a Macedonian question answering system. *ICT Innovations 2015. Advances in Intelligent Systems and Computing*, volume 399, pp. 205-214, Springer, Cham.
   DOI: 10.1007/978-3-319-25733-4_21
6. Jasmina Jovanovska and Goce Armenski. Implementing a recommendation system in an e-commerce web portal. *The 3th International Conference — Education across Borders, Education and Research across Time and Space*, 6-7 October 2016, Bitola, Macedonia.
7. Jasmina Jovanovska and Goce Armenski. Improving student knowledge with interactive online courses. *The 11th International Balkan Education and Science Congress "The Future of Education and Education for the Future"*, Poreč, Croatia, October 2016.
8. Jasmina Jovanovska, Ivana Bozhinovska, and Katerina Zdravkova. 2017. Information retrieval with reinforced word classes. In *Proceedings of 8th Balkan Conference in Informatics,* Skopje, 21 – 23 September 2017 (BCI2017), 8 pages.
   DOI: 10.1145/3136273.3136292
9. Jasmina Jovanovska and Goce Armenski. Analyzing multiple-choice question validity. *International scientific conference "Education at the crossroads - conditions, challenges, solutions and perspectives"*, Bitola, 10-11 November 2017, in press.