НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»

Факультет прикладної математики

Кафедра прикладної математики

«На правах рукопису»
УДК 519.767.6

«До захисту допущено»

Завідувач кафедри
_____ О. Р. Чертов
(підпис)

«___» _____2015 р.

Магістерська дисертація

на здобуття ступеня магістра

зі спеціальності 8.04030101 «Прикладна математика»

на тему: Автоматизована система генерації відкритих запитань до природно-мовних текстів

Виконав: студент 2 курсу, групи КМ-31М

Кривоніс Богдан Юрійович _____
(підпис)

Науковий керівник доцент, канд. техн. наук Сирота С. В. _____
(підпис)

Консультант із нормоконтролю старший викладач Мальчиков В. В. _____
(підпис)

Рецензент професор, д-р техн. наук, проф. Жабін В. І. _____
(підпис)

Засвідчую, що у цій магістерській дисертації немає запозичень з праць інших авторів без відповідних посилань.
Студент _____
(підпис)

Київ – 2015 року

Національний технічний університет України
«Київський політехнічний інститут»

Факультет прикладної математики

Кафедра прикладної математики

Рівень вищої освіти – другий (магістерський)

Спеціальність 8.04030101 «Прикладна математика»

ЗАТВЕРДЖУЮ
Завідувач кафедри
_____ О. Р. Чертов
(підпис)
«___» _____2015 р.

ЗАВДАННЯ
на магістерську дисертацію студенту
Кривоносу Богдану Юрійовичу

1. Тема дисертації: «Автоматизована система генерації відкритих запитань до природомовних текстів»,
   науковий керівник дисертації Сирота Сергій Вікторович, канд. техн. наук,затверджені наказом по університету від «20» березня 2015 р. №785-С.
2. Термін подання студентом дисертації – «18» червня 2015 р.
3. Об'єкт дослідження: процеси, моделі та методи автоматизованого аналізу текстової інформації.
4. Предмет дослідження: методи та методології автоматизованого реферування та пошуку ключової інформації в тексті, методи генерації запитань до природо мовних текстів.
5. Перелік завдань, які потрібно розробити:
   - проаналізувати наявні методи та методики автоматизованого реферування та пошуку ключової інформації в тексті;
   - дослідити раніше розроблені методи генерації запитань до речень на природній мові;
   - розробити власну методику пошуку ключової інформації в тексті для подальшої генерації відкритих запитань;

- дослідити та розробити методи генерації відкритих запитань до науково-технічних текстів українською мовою;
- розробити математичне та програмне забезпечення для автоматизованої системи генерації відкритих запитань.

6. Орієнтовний перелік ілюстраційного матеріалу:
   - демонстраційні таблиці порівняння методів;
   - теоретичні аспекти роботи обраних методів;
   - структурна схема програмних модулів.
7. Орієнтовний перелік публікацій:
   - доповідь на факультетській науковій конференції «ПМК-2015»;
   - публікація статті на міжнародній науково-практичній конференції «Наука України: проблеми сьогодення та перспективи розвитку».
8. Дата видачі завдання «25» жовтня 2013 р

Календарний план

№ з/п	Назва етапів виконання магістерської дисертації	Термін виконання етапів магістерської дисертації	Примітка
1.	Вибір напряму дослідження та узгодження тематики МД з керівником.	15 вересня – 30 жовтня 2013	
2.	Ґрунтовне ознайомлення з предметною галуззю;	30 жовтня 2013 – 15 лютого 2014	
3.	Вивчення літератури, пошук додаткової інформації.	15 лютого – 1 вересня 2014	
4.	Проведення дослідження, розроблення програмного забезпечення	1 вересня 2014 – 1 березня 2015	
5.	Завершення роботи над основною частиною МД, переддипломна практика, робота над публікаціями.	1 березня – 1 травня 2015	
6.	Оформлення текстової і графічної частини МД.	1 травня – 1 червня 2015	
7.	Попередній захист МД.	1 червня – 15 червня 2015	
8.	Основний захист МД.	15 червня – 30 червня 2015	

Студент _____ Б.Ю. Кривоніс
(підпис) (ініціали, прізвище)

Науковий керівник дисертації _____ С. В. Сирота
(підпис) (ініціали, прізвище)

РЕФЕРАТ

Актуальність проблеми. З початком ери автоматизації кількість нових розробок та досліджень зростає з експоненціальною швидкістю. Завдяки Інтернету більшість новітніх досліджень і відкриттів стають доступними для широких мас населення й активно використовуються. Це призвело до необхідності постійного самонавчання як викладачів, так і фахівців інших галузей.

Самонавчання пов'язане з проблемою контролю засвоєння вивченої інформації, тому постає питання в створенні автоматизованих систем, котрі допомогли б у цьому, наприклад систем автоматизованої генерації запитань. На сьогодні не існує систем, які здатні повноцінно розв'язувати таку задачу, адже тут необхідно не тільки виділити найважливішу інформацію з тексту, а й згенерувати запитання з виділеної інформації. Методики, що розроблювались для вирішення кожної з цих двох задач, досить ефективно працюють окремо, але в поєднанні не показують хороших результатів. Наприклад, існують окремо системи автоматизованого реферування, які виділяють важливу інформацію з тексту, системи генерації запитань до конкретних речень, проте їх поєднання не дає бажаного результату. Крім того, ці системи створені для мов із чітким порядком слів у реченні (англійська, німецька мови тощо), для текстів українською мовою згадані вище системи не розроблялися. Тому створення методик та програмних модулів для генерації відкритих запитань до текстів українською мовою є актуальною і важливою задачею з наукової, так і з практичної точки зору.

Об'єкт дослідження – процеси, моделі та методи автоматизованого аналізу текстової інформації.

Предмет дослідження – методи та методики автоматизованого реферування та пошуку ключової інформації в тексті, методи генерації запитань до текстів українською мовою.

Мета роботи – розробка та удосконалення методики пошуку ключової інформації в науково-технічних текстах та методів генерації відкритих запитань до науково-технічних текстів українською мовою.

Методи дослідження:

— аналіз, синтез, індукція, дедукція, абстрагування, конкретизація, класифікація, систематизація, схематизація;

— методи аналізу текстової інформації, методи кластеризації даних та методи автоматизованого реферування.

Наукова новизна роботи полягає в тому, що:

1. Розроблено методику пошуку ключових речень у науково-технічних текстах, яка відрізняється від наявних тим, що дозволяє враховувати тематичне спрямування і смислове навантаження речень аналізованого тексту.

2. Вперше запропоновано використовувати метод нейронних мереж для генерації запитань до речень, що дало змогу класифікувати речення текстів українською мовою (де немає чіткого синтаксичного порядку слів у реченні) і генерації запитань до них.

3. Розроблено принципово нові класи для класифікації речень, що дозволило трансформувати вхідні речення та отримувати в результаті запитання.

Практична цінність отриманих в роботі результатів полягає в тому, що запропоновані варіанти розв'язання проблем дають хороші результати, а методи та програмні модулі можуть використовуватись під час створення системи генерації запитань, яка спрощує перевірку рівня засвоєння інформації.

Апробація роботи. Основні положення і результати роботи були представлені на VII науковій конференції магістрантів та аспірантів «Прикладна математика та комп'ютинг» ПМК-2015 (Київ, 15–16 квітня 2015 р.) та опубліковані у збірнику міжнародної науково-практичної конференції «Наука України: проблеми сьогодення та перспективи розвитку».

Структура та обсяг роботи. Магістерська дисертація складається зі вступу, чотирьох розділів, висновків та додатків.

У вступі викладень загальну характеристику роботи, подано оцінку

сучасного стану досліджуваної проблеми, обґрунтовано актуальність дослідження, сформульовано мету і завдання дослідження, наукову новизну отриманих результатів і практичну цінність роботи, наведено відомості про апробацію результатів і їх впровадження.

У першому розділі розглянуто основні методи автоматизованого реферування та методи кластеризації, подано результати їхнього порівняльного аналізу; на основі отриманих результатів обрано методи і методики, які будуть корисними для проведення дисертаційного дослідження.

У другому розділі проаналізовано принцип моделювання інформації і виявлення знань в текстах; удосконалено методику пошуку найбільш інформативних речень у тексті з урахуванням недоліків, які були виявлені в процесі аналізу методів.

У третьому розділі запропоновано власну класифікацію речень, котра буде застосовуватись для генерації запитань; висвітлено теоретичні аспекти методу, що використовується.

У четвертому розділі подано інформацію про розроблені програмні модулі; наведено приклади та результати тестування розроблених методик пошуку ключової інформації та методу генерації запитань.

У висновках викладено найбільш значущі наукові та практичні результати проведеного наукового дослідження та програмної реалізації, обґрунтовано достовірність отриманих результатів.

Загальний обсяг роботи становить 95 сторінок, основний зміст викладено на 59 сторінках. Робота містить 2 додатки, список використаних літературних джерел з 38 найменувань, 8 рисунків та 2 таблиці.

Ключові слова: метод автоматичного реферування, методи пошуку ключової інформації, метод класифікації, генерація запитань.

РЕФЕРАТ

Актуальность проблемы. С началом эры автоматизации количество новых разработок и исследований растет с экспоненциальной скоростью. Благодаря Интернету большинство новейших исследований и открытий становятся доступными для широких масс населения и активно используются. Это привело к необходимости постоянного самообучения как преподавателей, так и специалистов разных отраслей.

Самообучение связано с проблемой контроля усвоения изученной информации, поэтому возникает вопрос в создании автоматизированных систем, которые помогли бы в этом, например систем автоматизированной генерации вопросов. На сегодняшний день не существует систем, которые способны полноценно решать такую задачу, ведь здесь необходимо не только выделить важнейшую информацию из текста, но и сгенерировать вопрос с выбранной информации. Методики, разрабатаные для решения каждой из этих двух задач, достаточно эффективно работают отдельно, но в сочетании не показывают хороших результатов. Например, существуют отдельно системы автоматизированного реферирования, которые выделяют важную информацию из текста, системы генерации вопросов к конкретным предложений, однако их сочетание не дает желаемого результата. Кроме того, эти системы созданы для языков с четким порядком слов в предложении (английский, немецкий языки и т.д.), для текстов на украинском языке упомянутые выше системы не разрабатывались. Поэтому создание методик и программных модулей для генерации открытых вопросов к текстам на украинском языке является актуальной и важной задачей как с научной, так и с практической точки зрения.

Объект исследования - Процессы, модели и методы автоматизированного анализа текстовой информации.

Предмет исследования - методы и методики автоматизированного реферирования и поиска ключевой информации в тексте, методы генерации вопросов к текстам на украинском языке.

Цель работы - разработка и усовершенствование методики поиска

ключевой информации в научно-технических текстах и методов генерации открытых вопросов к научно-техническим текстам на украинском языке.

Методы исследования:

- Анализ, синтез, индукция, дедукция, абстрагирование, конкретизация, классификация, систематизация, схематизация;

- Методы анализа текстовой информации, методы кластеризации данных и методы автоматизированного реферирования.

Научная новизна работы заключается в том, что:

1. Разработана методика поиска ключевых предложений в научно-технических текстах, которая отличается от имеющихся тем, что позволяет учитывать тематическое направление и смысловую нагрузку предложений анализируемого текста.

2. Впервые предложено использовать метод нейронных сетей для генерации вопросов к предложениям, что позволило классифицировать предложения текстов на украинском языке (где нет четкого синтаксического порядка слов в предложении) и генерации вопросов к ним.

3. Разработаны принципиально новые классы для классификации предложений, что позволило трансформировать входные предложения и получать в результате вопрос.

Практическая ценность полученных в работе результатов заключается в том, что предложенные варианты решения проблем дают хорошие результаты, а методы и программные модули могут использоваться при создании системы генерации вопросов, которая упрощает проверку уровня усвоения информации.

Апробация работы. Основные положения и результаты работы были представлены на VII научной конференции магистрантов и аспирантов «Прикладная математика и компьютинг» ПМК-2015 (Киев, 15-16 апреля 2015) и опубликованы в сборнике международной научно-практической конференции «Наука Украины: проблемы и перспективы развития ».

Структура и объем работы. Магистерская диссертация состоит из введения, четырех глав, заключения и приложений.

Во введении представлено общую характеристику работы, дана оценка

современного состояния исследуемой проблемы, обоснована актуальность исследования, сформулированы цели и задачи исследования, научная новизна полученных результатов и практическую ценность работы, приведены сведения об апробации результатов и их внедрения.

В первом разделе рассмотрены основные методы автоматизированного реферирования и методы кластеризации, представлены результаты их сравнительного анализа; на основе полученных результатов избраны методы и методики, которые будут полезными для проведения диссертационного исследования.

Во втором разделе проанализированы принцип моделирования информации и выявления знаний в текстах; усовершенствована методика поиска наиболее информативных предложений в тексте с учетом недостатков, которые были выявлены в процессе анализа методов.

В третьем разделе предложено собственную классификацию предложений, которая будет применяться для генерации вопросов; освещены теоретические аспекты используемого метода.

В четвертом разделе представлена информация о разработанных программных модуля; приведены примеры и результаты тестирования разработанных методик поиска ключевой информации и метода генерации вопросов.

В выводах изложены наиболее значимые научные и практические результаты проведенного научного исследования и программной реализации, обоснованно достоверность полученных результатов.

Общий объем работы составляет 95 страниц, основное содержание изложено на 59 страницах. Работа содержит 2 приложения, список использованных литературных источников из 38 наименований, 8 рисунков и 2 таблицы.

Ключевые слова: метод автоматического реферирования, методы поиска ключевой информации, метод классификации, генерация вопросов.

ABSTRACT

Background. Since the beginning of the era of automation the number of new research and development is growing at an exponential rate. Thanks to the Internet most of the latest research and discoveries are made available to the general population and actively used. This led to the need for continuous learning as teachers and professionals in other industries.

Self-control problems associated with the assimilation of learned information that's way the question to create automated systems that would help in this, such as automated generation of questions, interesting. At present there are systems that are able to fully solve such a problem, because there should not only highlight the most important information from the text, but also to generate questions from the selected information. Methods witch are developed to address each of these two tasks quite effectively operate separately but in conjunction did not show good results. For example, there are separate systems of automatic summarization which distinguish important information from text generation system to specific questions sentences, but their combination gives the desired result. In addition, these systems are designed for languages with strict word order in a sentence (English, German, etc.) for Ukrainian Texts aforementioned systems are not developed. The creation of methods and software modules for generation of open questions for Ukrainian Texts is an urgent and important task of the scientific and practical point of view.

The object of research - the processes, models and methods of automated analysis of textual information.

Subject of research - Methods and techniques of automated summarization and retrieval of key information in the text, methods of generating questions to the text in Ukrainian.

Purpose of research - development and improvement of methods of finding key information in scientific texts and methods of open-ended questions to generate scientific texts in Ukrainian.

Methods of research:

- Analysis, synthesis, induction, deduction, abstraction, specification,

classification, systematization, schematization;

- Methods of analysis of textual information, data clustering methods and techniques of automatic summarization.

Scientific novelty lies in the fact that:

1. A method of finding the key sentences in scientific texts that differs from existing that takes into account thematic orientation and meaning of sentences analyzed text.

2. The first time the use of neural networks method for generating questions to the sentences, which allowed the sentence to classify texts Ukrainian (where there is no clear syntactic order of words in a sentence) and generate questions for them.

3. A fundamentally new classes for the classification of sentences that allowed to transform the input sentence and receive as a result of questions.

The practical value obtained in the results is that the proposed alternative solutions to problems give good results, methods and software modules can be used when creating a system of generating questions that simplifies the verification of the assimilation of information.

Testing of work. Substantive provisions and results were presented at the VII scientific conference of graduate and post-graduate "Applied mathematics and computing 'MVP-2015 (Kyiv, 15-16 April 2015) and published in the Proceedings of the international scientific conference" Science of Ukraine: problems Present and Prospects for Development ".

The structure and scope of work. Master's thesis consists of introduction, four chapters, conclusions and applications.

In the introduction the general characteristics of assumptions, evaluates the current state of research problem, the urgency of research, formulated the purpose and objectives of the study, scientific novelty of the results and practical value of work, provides information on testing results and their implementation.

In the first chapter the basic methods of automatic summarization and clustering techniques, presented the results of their comparative analysis; Based on the results of selected methods and techniques that will be useful for dissertation research.

The second section analyzes the principle of information modeling and

knowledge discovery in texts; improved method of finding the most informative sentences in the text and the costs that were found in the analysis methods.

The third section offered its own classification sentences, which will be used to generate questions; highlights the theoretical aspects of the method used.

The fourth section presents information on developed software modules; are examples of test results and developed methods of search key information and method of generating questions.

The conclusions set out the most important scientific and practical results of scientific research and program implementation, proved the reliability of the results.

The total amount of work is 88 pages, the essence contained 60 pages. Work includes 2 applications, the list of used literature 38 items, 8 figures and 2 tables.

Keywords: method of automatic summarization, search methods key information, the method of classification, the generation of questions.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, СКОРОЧЕНЬ І ТЕРМІНІВ

АР	Автоматизоване реферування.
SVM	Support vector machine – метод опорних векторів.
МФЗ	Синтаксичний міжфразовий зв’язок.
ШНМ	Штучні нейронні мережі.
ШІ	Штучний інтелект.

ВСТУП

Сучасні комп'ютерні технології призвели до суттєвої трансформації змісту та організаційних форм освіти. Зокрема, усе більшої актуальності набуває самонавчання. Адже кількість наукових розробок невпинно зростає, а новітні комунікації за допомогою мережі Інтернет дозволили користувачам усього світу отримати швидкий доступ до необхідних знань. Технічні засоби нині стали доступними для застосування їх під час самоосвіти, тому й виникла потреба на науковому рівні розглянути їх можливості для створення нових систем автоматизованого аналізу наукової літератури.

Однією з важливих проблем самоосвіти є питання контролю отриманих знань. Застосування сучасних комп'ютерних технологій під час контролю засвоєної інформації дасть змогу суттєво підвищити якість самостійного навчання. Проте, аналіз наукової та прикладної літератури з теми дав змогу констатувати, що на сьогодні ще не розроблено автоматизованих систем, здатних ефективно генерувати запитання для самоконтролю. Зокрема, це пов'язано з певною складністю реалізації цього процесу. Адже автоматизована генерація запитань має охоплювати такі етапи: виділення ключової інформації з тексту та генерацію до неї запитань, не кажучи вже про попередній синтаксично-семантичний розбір тексту. Крім того, системи, що вирішують окремо кожну задачу, створюються для мов із чіткою структурою речення (англійська, німецька),

в той час як для української мови, де порядок слів чітко не визначений, подібних методик не розроблялося. Ті методи, що існують на сьогодні, здатні вирішувати окремо кожну з наведених задач, проте у поєднанні вони показують незадовільні результати.

Отже, нині постає потреба розробити нові ефективні методики та програмні модулі для автоматизованої генерації запитань до наукового тексту, котрі дозволять значно покращити якість самостійного навчання населення.

Детальний аналіз літератури з теми автоматизованого генерування запитань до тексту показав відсутність наукових та прикладних розробок у цьому напрямі. Пов'язана з досліджуваною темою проблема автоматизованого реферування текстів досліджена досить добре і представлена науковими працями як українських, так і зарубіжних авторів. Зокрема, питання автоматизованого реферування знайшло відображення у працях В. Берзона, В. Горькової, О. Трішук та ін. Створення моделей автоматизованого реферування викладено, зокрема, у працях О. Лазаренко та А. Яковенко. Учені сходяться в думці, що основним завданням автоматизованого реферування текстів учені вважають виокремлення найбільш важливих понять, фраз та речень із тексту.

Проблема адекватного автоматизованого реферування документів на сьогодні особливо актуальна для тих досліджень, в рамках яких розробляються відповідні алгоритми і програмні засоби. В більшості випадків системи і методи, що розроблюються не є універсальними, і потребують доопрацювань для їх використання в інших системах.

Питання генерації запитань у наш час є найбільш досліджуваною серед закордонних учених. Нею займаються такі науковці, як Michael Heilman, Noah A. Smith, V. Rus, J. Lester та багато інших. Майже всі роботи цих авторів присвячені генеруванню запитань на англійській мові або на іншій мові з чітким порядком слів. Також слід додати, що для генерування

запитань до тексту англійською мовою зазвичай використовують представлення речень у вигляді зв'язних графів. Такий підхід є досить простим і ефективним для мов на кшталт англійської, але не показує задовільних результатів під час роботи з мовами без чіткого порядку слів.

Також одним з способів генерації запитань є спосіб, пов'язаний з класифікацією речення. Такий спосіб зустрічається у роботі «Automatic Question Generation from Sentences» [1]. Проблемами класифікації речень та текстів розглядаються в роботах «Experiments with Sentence Classification» [2], «Convolutional Neural Networks for Sentence Classification» [3]. Однак тут висвітлюються лише загальні принципи класифікації речень у текстах англійською мовою. Тоді як для генерації запитань потрібно розробити інші класи, а отже і методи, які дозволять легко генерувати запитання до речення після його класифікації.

Отже, для розробки системи автоматизованої генерації запитань до текстів українською мовою необхідно: по-перше, удосконалити існуючі методи та методики виділення ключової інформації, яка дозволить достатньо якісно виявляти найбільш важливі речення в наукових текстах; а по-друге – розробити метод класифікації обраних речень, що дозволить ефективно трансформувати вхідні речення в запитання.

Для розв'язання вищезазначених проблем об'єктом дослідження візьмемо процеси, моделі та методи автоматизованого аналізу текстової інформації. Предметом дослідження є методи та методики автоматизованого реферування та пошуку ключової інформації в тексті, методи генерації запитань до текстів українською мовою.

Метою роботи є розробка та удосконалення методик пошуку ключової інформації в науково-технічних текстах та методів генерації відкритих запитань до науково-технічних текстів українською мовою. Виходячи з мети магістерської дисертації сформулюємо завдання: порівняти методи автоматизованого реферування і методи класифікації, дослідити модель

представлення знань у тексті, удосконалити методику виділення ключових речень в тексті, створити власну класифікацію речень за допомогою нейронних мереж, а також розробити та провести тестування програмного продукту для реалізації завдання роботи.

Серед методів дослідження зазначимо такі: аналіз, синтез, індукція, дедукція, абстрагування, конкретизація, класифікація, систематизація, схематизація – з метою узагальнення наукових розробок провідних вітчизняних та зарубіжних учених із проблеми автоматизованої обробки текстової інформації; методи аналізу текстової інформації, методи кластеризації даних та методи автоматизованого реферування – з метою створення нового програмного модуля для реалізації завдання дослідження.

РОЗДІЛ 1. ПОРІВНЯННЯ МЕТОДІВ АВТОМАТИЗОВАНОГО РЕФЕРУВАННЯ І МЕТОДІВ КЛАСИФІКАЦІЇ

Методи автоматизованого аналізу текстової інформації почали досліджувати ще в 60-х роках 20-го сторіччя. Це пов'язано з появою великої кількості текстових документів і необхідності їх швидкого аналізу та обробки. В наш час існує багато перспективних напрямків обробки текстової інформації: починаючи з індексування документів в Інтернеті, закінчуючи виділення ключових знань з текстів. Автоматизованою обробкою інформації займалися такі вчені як H. Kunichika, A. Takeuchi, S. Otsuki, Є. Борохова, Х. Лун, М. Жинкіна А. Новикова, В. Рубашкіна, Г. Жданова, та багато інших. Зокрема, питання автоматизованого реферування знайшло відображення у працях В. Берзона, В. Горькової, О. Тріщук та ін. Створення моделей автоматизованого реферування викладено, зокрема, у працях О. Лазаренко та А. Яковенко. Завданнями генерації текстів займались такі вчені як J. Wolfe, L. Baptist, S. Seneff, Yushi Xu, Anna Goldie. Одними з найперспективніших напрямів дослідження автоматизованого аналізу тексту, на наш погляд є питання виявлення ключової інформації та проблема генерації нового тексту.

Класичною задачею виявлення ключової інформації в тексті є задача автоматизованого реферування. Ця задача включає в себе виділення найбільш інформативних речень в тексті та адаптацію їх до застосування в рефераті. Для нашого дослідження питання автоматизованого реферування

буде цікавим для нас зі сторони виявлення ключових речень в тексті, до яких потім будуть генеруватись запитання.

Однією задач в галузі генерації тексту є задача генерації запитань. Існує декілька способів генерації запитань: з ключових слів, за заданими критеріями чи генерація запитання до існуючого речення. В цій роботі будуть представлені методи генерації запитань до речень, методом класифікації типу речення.

1.1. Аналіз методів автоматизованого реферування

Нині розроблено досить велику кількість різноманітних методів автоматизованого реферування. Що призвело до різних способів класифікації методів. Однією з таких класифікацій є поділ методів на поверхневі та глибинні. Значний аналітичний огляд також містить монографія В. І. Горькової і Є. А. Борохова [4] про теоретичні і прикладні дослідження лінгвістичних і структурних характеристик реферату та їх реалізацію в автоматизованій обробці.

Так, у роботі [5] В. Є. Берзона можна знайти цілком прийнятну та докладну класифікацію методів реферування. Дослідник розрізняє наступні методи:

1) статистичні

2) позиційні

3) дескрипторні

4) семантичні та синтаксичні

5) засновані на дослідженні структури зв'язного тексту.

Родоначальником статистичних методів є Х. П. Лун, який вважається основоположником автоматичного реферування взагалі. За Х. П. Луном, основний смисловий зміст реферованого джерела можна розкрити у вигляді переліку речень, найбільш значущих для даного документа. Значущими реченнями вважаються ті, що містять у своєму складі «скупчення» значущих для даного документа слів. Значущість слів обумовлюється частотою їх вживання в тексті. Більшість сучасних програм з АР працюють на основі саме статистичних методів.

Ці методи по своїй реалізації є одними з найпростіших, як в реалізації, так в розумінні. Самі методи не потребують складних алгоритмів чи розрахунків. Простота в реалізації і легкість в розумінні стали ключовим аспектом в популярності цих методів. Очевидно що простота методів тягне за собою і ряд недоліків. Методи не є достатньо ефективними і зазвичай використовуються в поєднанні з позиційними.

У позиційних методах реферування для ідентифікації найбільш значущих речень використовують розташування речень у тексті. Існує думка, що основний зміст первинного документа відбивається в рефераті, який складається з перших речень усіх абзаців або з першого, другого та останнього речень тексту. [6]

По своїй структурі позиційні методи дуже схожі на статистичні, мають схожі переваги і недоліки. Очевидно, що позиційний метод використовується в сучасних системах АР у поєднанні зі статистичним.

Дескрипторні методи полягають у виділенні інформації на основі певного дескриптора. Тобто виділяються речення або абзаци, які підпадають під певні критерії. Ці методи схожі на поєднання статистичних і позиційних, але потребують втручання ззовні, за для визначення правил

відбору (дескрипторів). Вони трохи складніші в реалізації ніж попередні, але так само не враховують смислу тексту.

Синтаксичні та семантичні методи базуються на синтаксичних характеристиках та семантичних відношеннях в тексті. Досить цікавою вважається також ідея Е. Ф. Скороходько про вибір оптимальної процедури реферування в залежності від типу семантичної структури тексту і числових характеристик семантичних відношень у цьому тексті. У роботі Н. І. Гендіної розглядається один із підходів до укладання рефератів на основі формально-змістового аналізу текстів первинних документів. Відмінність цього підходу від інших відомих полягає в тому, що він ураховує формальні текстові ознаки, так звані маркери – стійкі словесні звороти, що характеризують конкретні аспекти змісту [6].

Ці методи є набагато ефективнішими в порівнянні з дескрипторними чи статистичними. Звісно, вони набагато складніші в реалізації. Але в наш час вже створені програмні продукти здатні виконувати синтаксичний та семантичний аналіз речень. Одним з головних недоліків даних методів є те, що вони не враховують зміст тексту і здатні створювати лише квазіреферати.

Методи, засновані на дослідженні структури зв'язного тексту базуються на вивченні й систематизації глибинних механізмів зв'язного тексту. При цьому треба враховувати, що текст являє собою єдиний механізм, який складається з трьох різних за своєю природою механізмів: семіотичної системи, мовної системи і системи знань про світ. Лише при такому підході з'явиться можливість перейти від «псевдосмислу» до смислу, від квазірефератів до рефератів [6].

Вище було розглянуто основні класи методів авто реферування. Спробуємо провести аналіз цих класів. Проведемо порівняння за наступними критеріями:

1) Простота в використанні
2) Імітує інтелектуальний процес
3) Враховує зміст тексту
4) Враховує тему та тематику тексту

Порівняємо методи відповідно до вищезазначених критеріїв (табл. 1.1).

Таблиця 1.1 – Порівняння методів автоматизованого реферування

Метод/критерії	Простота в використанні	Імітує інтелектуальний процес	Враховує зміст тексту	Враховує тему та тематику тексту
Статистичні методи	+	–	–	+
Позиційні методи	+	–	–	–
Дескрипторні методи	+	–	–	–
Синтаксичні та семантичні методи	+	+	–	–
Методи що засновані на дослідженні структури зв'язного тексту	+	+	+	–

Порівнявши характеристики підходів у кожної групи методів було запропоновано поєднати простий статистичний метод з методом, що заснованій на дослідженні структури зв'язного тексту. Для такого поєднання було обрано метод міжфразових зв'язків та метод ключових слів.

1.2 Аналіз математичних методів класифікації

Найбільш поширеним методом генерації запитань до речення є метод через класифікацію самого речення. Якщо стає зрозумілим тип речення то запитання генерується методом заміни декількох слів та адаптацією закінчень в деяких словах. Розробкою алгоритмів класифікації інформації займається розділ математики DataMining. Цей розділ налічує десятки різних математичних методів, більшість з яких є універсальними і здатні розв'язувати широкий спектр задач. Однією з таких є задача класифікації і кластеризації.

Далі розглянемо використовувані математичні методи, призначені для класифікації, та перелічимо їх позитивні сторони й недоліки.

Метод опорних векторів (SVM) належить до групи граничних методів; визначає класи за допомогою меж просторів. Опорними векторами вважаються об'єкти множини, що лежать на цих межах. Класифікація вважається вдалою, якщо простір між межами – пустий. Машина опорних векторів — дуже потужний класифікатор. Крім того, після навчання класифікація нових результатів вибір відбувається дуже швидко, оскільки потрібно лише визначити, по яку сторону від роздільника виявилася крапка. Перетворивши дискретні величини в числа, ви зможете змусити SVM працювати з дискретними і числовими даними одночасно. Недолік полягає в тому, що оптимальна ядерна функція і її параметри залежать від конкретного набору даних, так що всякий раз доводиться підбирати їх заново. Перебір можливих значень у циклі частково вирішує проблему, але для цього потрібно мати достатньо великий набір даних, щоб результатами перехресної перевірки можна було

довіряти. У загальному випадку метод опорних векторів краще пристосований для таких завдань, де доступний великий обсяг даних, тоді як інші методи, скажімо дерева рішень, дають цікаву інформацію вже на досить скромних наборах даних. Як і нейронні мережі, SVM є «чорний ящик»; зрозуміти хід міркувань тут навіть складніше через трансформації в багатомірному простір. SVM може дати правильну відповідь, але ви ніколи не дізнаєтеся, як вона отримана [7].

Кластерний аналіз (англ. *data clustering*) — задача розбиття заданої вибірки об'єктів (ситуацій) на підмножини, що називаються кластерами, так, щоб кожен кластер складався з схожих об'єктів, а об'єкти різних кластерів істотно відрізнялися. Завдання кластеризації відноситься до статистичної обробки, а також до широкого класу завдань навчання без вчителя. Кластерний аналіз — це багатовимірна статистична процедура, яка виконує збір даних, що містять інформацію про вибірку об'єктів і потім упорядковує об'єкти в порівняно однорідні групи — кластери.

Основна мета кластерного аналізу — знаходження груп схожих об'єктів у вибірці. Спектр застосувань кластерного аналізу дуже широкий: його використовують в археології, антропології, медицині, психології, хімії, біології, державному управлінні, філології, маркетингу, соціології та інших дисциплінах. Однак універсальність застосування привела до появи великої кількості несумісних термінів, методів і підходів, що утруднюють однозначне використання і несуперечливу інтерпретацію кластерного аналізу. Розглянемо один з таких методів.

Метод K-найближчих сусідів — метод автоматичної класифікації об'єктів. Основним принципом методу є те, що об'єкт присвоюється тому класу, який є найбільш поширеним серед сусідів цього елемента. Сусіди беруться виходячи з безлічі об'єктів, класи яких вже відомі, і, виходячи з ключового для цього методу значення k вираховується, який клас найбільш

численний серед них. Одним з позитивних моментів цього методу є те, що процес міркування легко зрозуміти, а також можна побачити, які сусіди брали участь в обчисленнях. Наприклад нейронні мережі, які здатні розв'язувати ті ж задачі, не можуть показати схожі зразки, щоб можна було зрозуміти, як був отриманий результат. Далі процедура визначення підходящих коефіцієнтів масштабуванням дозволяє не тільки поліпшити якість прогнозу, та пошуку але і підказати, які змінні істотні для цього. Алгоритм відноситься до числа оперативних методів, тобто дані можна додавати в будь-який момент, на відміну, скажімо, від машинно-опорних векторів, яку потрібно перенавчати при кожному даних. Більше того, при додаванні нових даних не потрібні ніякі обчислення; досить просто включити дані в набір. Основний недолік цього алгоритму полягає в тому, що для пошуку йому потрібні всі дані, на яких проводилося навчання. Якщо в наборі мільйони зразків, то на це витрачається не тільки пам'яті, але і час. Ще один недолік — складний пошук підходящих коефіцієнтів масштабування. Якщо доводиться випробовувати багато різних змінних, то для знаходження відповідного поєднання масштабних коефіцієнтів, можливо, буде потрібно переглянути мільйони комбінацій [7].

Наївний байєсовський класифікатор — простий імовірнісний класифікатор, заснований на застосуванні Теореми Байєса зі строгими (наївними) припущеннями про незалежність. Залежно від точної природи ймовірнісної моделі, наївні байєсовські класифікатори можуть навчатися дуже ефективно. У багатьох практичних додатках для оцінки параметрів для наївних Байесових моделей використовують метод максимальної правдоподібності; іншими словами, можна працювати з наївною байесовской моделлю, не використовуючи байєсовські методи. Незважаючи на наївний вигляд і, безсумнівно, дуже спрощені умови, наївні байєсовські класифікатори часто працюють набагато краще в

багатьох складних життєвих ситуаціях. Мабуть, найістотніша перевага наївних байєсівських класифікаторів в порівнянні з іншими методами полягає в тому, що їх можна навчати і потім опитувати на великих наборах даних. Навіть якщо навчальний набір дуже великий, зазвичай для кожного зразка є лише невелика кількість ознак, а навчання і класифікація зводяться до простих математичних операцій над ймовірностями ознаками. Це особливо важливо, коли навчання проводиться інкрементно, — кожний новий пред'явлений зразок можна використовувати для оновлення ймовірностей без використання старих навчальних даних. Ще одна перевага наївних байєсівських класифікаторів — відносна простота інтерпретації того, чого класифікатор навчився. Оскільки вірогідність використання кожної ознаки зберігається, можна в будь який момент подивитись в базі даних, які ознаки оптимальні. Основний недолік наївних байєсівських класифікаторів — їх нездатність враховувати залежність результату від поєднання ознак [7].

Дерево прийняття рішень (також можуть називатися деревами класифікацій або регресійними деревами) — використовується в галузі статистики та аналізу даних для прогнозних моделей. Структура дерева містить такі елементи: «листя» і «гілки». На ребрах («гілках») дерева прийняття рішення записані атрибути, від яких залежить цільова функція, в «листі» записані значення цільової функції, а в інших вузлах — атрибути, за якими розрізняються випадки. Щоб класифікувати новий випадок, треба спуститися по дереву до листа і видати відповідне значення. Подібні дерева рішень широко використовуються в інтелектуальному аналізі даних. Мета полягає в тому, щоб створити модель, яка прогнозує значення цільової змінної на основі декількох змінних на вході.

Впадає в око простота інтерпретації навченої моделі і те, як добре алгоритм поміщає найбільш важливі фактори ближче до кореня дерева. Це

означає, що дерево рішень корисно не тільки для класифікації, а й для інтерпретації результатів. Як і у випадку байєсовського класифікатора, можна прослідкувати процес прийняття рішення і зрозуміти, чому вийшов саме такий, а не інший результат. Це може полегшити прийняття рішень поза процесом класифікації. Дерева рішення можуть працювати не тільки з дискретними, але і з числовими даними, оскільки шукає роздільну лінію, яка максимізує інформаційний виграш. Уміння змішувати дискретні та числові дані корисно для багатьох класів задач, а традиційні статистичні методи, наприклад регресійний аналіз, мають в цій частині труднощі. З іншого боку, дерева рішення не такі хороші для прогнозування числових результатів. Дані можна розділити за критерієм мінімальної дисперсії, але якщо вони досить складні, то дерево виявиться дуже великим і нездатним давати точні прогнози. Основна перевага дерев рішень в порівнянні з байєсовським класифікатором — здатність легко справлятися з взаємозв'язаними змінними [8].

Штучні нейронні мережі — математичні моделі, а також їхня програмна та апаратна реалізація, побудовані за принципом функціонування біологічних нейронних мереж — мереж нервових клітин живого організму. Системи, архітектура і принцип дії базується на аналогії з мозком живих істот. Ключовим елементом цих систем виступає штучний нейрон як імітаційна модель нервової клітини мозку — біологічного нейрона. Цей термін виник при вивченні процесів, які відбуваються в мозку, та при спробі змоделювати ці процеси. ШНМ представляють собою систему з'єднаних і взаємодіючих між собою простих процесорів(штучних нейронів). З точки зору машинного навчання, нейронна мережа являє собою окремий випадок методів розпізнавання образів, дискримінантного аналізу, методів кластеризації тощо. А з точки зору штучного інтелекту, ШНМ є основою філософської течії коннективізму і основним напрямком

в структурному підході з вивчення можливості побудови (моделювання) природного інтелекту за допомогою комп'ютерних алгоритмів. Нейронні мережі не програмуються в звичайному розумінні цього слова, вони навчаються.

Головна перевага нейронних мереж полягає в тому, що вони здатні розв'язувати складні нелінійні функції і розкривати залежності між різними вхідними даними. Нейронні мережі також допускають інкрементне навчання і зазвичай не потребують багато місця для зберігання навчених моделей, оскільки моделюється списком чисел, що представляють ваги синапсів. Зберігати вихідні дані для подальшого навчання не потрібно, отже нейронні мережі здатні працювати з безперервним потоком навчальних даних. Основний недолік нейронних мереж в тому, що вони є «чорним ящиком». Неможливість ознайомитися з процесом міркування для деяких додатків неприйнятна. Ще один недолік — відсутність твердих правил щодо вибору швидкості навчання та розміру мережі для вирішення конкретного завдання. У цьому разі необхідно експериментувати. Якщо швидкість навчання занадто висока, то мережа робитиме зайві загальні висновки на основі зачумлених даних, а якщо вона занадто мала, то, мережа може не навчиться робити висновки, на основі запропонованих даних [7].

Вище було розглянуто одні з найпоширеніших та найбільш універсальних математичних методів. Порівняємо їх та виберемо найбільш доцільні в нашому випадку за допомогою наведених нижче критеріїв.

1) Не потребує постійних великих обрахунків.
2) Не потребує перенавчання при додаванні даних.
3) Здатний враховувати взаємозв'язані критерії
4) Здатен класифікувати складні випадки

Порівняємо методи відповідно до вищезазначених критеріїв (табл. 1.2).

Таблиця 1.2 – Порівняння математичних методівкластеризації

Метод/критерії	Великі обрахунки	Не потребує перенавчання	Враховує взаємозв'язані критерії	Здатність класифікувати складні випадки
Метод опорних векторів	–	–	+	+
K-найближчих сусідів	–	+	+	–
Байєсовський класифікатор	+	+	–	–
Дерево прийняття рішень	+	+	+	–
Нейронні мережі	+	+	+	+

З порівняльної таблиці видно, що штучні нейронні мережі – це найбільш оптимальний метод для класифікації речень. Зазвичай обирають метод дерев, але для української мови, синтаксичні конструкції якої є складнішими, ніж у англійської чи німецької, краще обрати нейронні мережі.

1.3. Висновки

Розглянуши основні класи методів для пошуку ключової інформації в тексті, а саме статистичні методи, позиційні, дескрипторні, синтаксичні

та семантичні і методи що засновані на дослідженні структури зв'язного тексту, було проведено їх порівняльний аналіз. Аналіз показав що для задачі пошуку ключової інформації в системі автоматизованої генерації запитань до наукових текстів найкращим варіантом буде поєднання статистичного методу та методу що заснованій на дослідженні структури зв'язного тексту. Для поєднання було обрано метод ключових слів та метод міжфразових зв'язків.

Також порівнявши основні методи класифікації даних, а саме метод опорних векторів, K-найближчих сусідів, Байєсовський класифікатор, дерево прийняття рішень та нейронні мережі, було обрано метод для класифікації типу речень. Для вирішення цієї задачі було обрано метод штучних нейронних мереж.

РОЗДІЛ 2. МОДЕЛЬ ПОШУКУ КЛЮЧОВОЇ ІНФОРМАЦІЇ

Із розширенням світової мережі Інтернет швидкими темпами зростає обсяг науково-технічної інформації, яку має переглянути користувач під час пошуку необхідних даних. Тому автоматизована обробка цієї інформації, реферування й анотування джерел, що несуть інформацію для конкретного користувача, стає все більш актуальним питанням.

Головним завданням автоматизованого реферування текстів є виокремлення найбільш важливих понять, фраз та речень із тексту. Виявлення найголовнішого в тексті є основою для низки напрямів роботи з текстовими документами – від створення невеликих рефератів і книг до задач пошуку інформації в Інтернеті.

На сьогодні поняття автоматизованого аналізу тексту перебуває в колі інтересів багатьох дослідників галузі. Зокрема, питання автоматизованого реферування знайшло відображення у працях В. Берзона, В. Горькової, О. Тріщук та ін. Створення моделей автоматизованого реферування викладено, зокрема, у працях О. Лазаренко та А. Яковенко.

Проблема адекватного автоматизованого реферування документів на сьогодні особливо актуальна для тих досліджень, в рамках яких розробляються відповідні алгоритми і програмні засоби. На нашу думку, застосування реферованих за допомогою ПЗ текстів є перспективним

напрямом дослідження. Особливо корисним він буде для сфери освіти: для вчителів загальноосвітніх та спеціалізованих середніх шкіл та викладачів ВНЗ, а також для осіб, котрі займаються самоосвітою

2.1. Моделювання знань у тексті

Вивчаючи питання автоматизованого аналізу та синтезу текстів з точки зору їхнього смислового навантаження, дослідники наголошують на складності та імпліцитності процесу узагальнення змісту під час реферування, що породило виникнення методу моделювання знань. За визначенням дослідника галузі Р. Підтровського, під час застосування цього методу «прямому спостереженню і дослідженню піддається не сам об'єкт, а його аналог чи, як звичайно говорять, модель» [9]. На його думку, метод моделювання знань у тексті є єдиним способом вивчити процес узагальнення смислу. Цей метод використовується під час опису лінгвістичних явищ у межах досліджень з прикладної лінгвістики, оскільки «…спілкування людини з ЕОМ може здійснюватися лише за умови, що в пам'ять комп'ютера буде введена певна модель, яка становить скорочений опис природної мови» [9].

Такі моделі, як стверджують науковці, використовують під час вивчення та узагальнення структури оригіналу (структурні моделі), його поведінки (функціональні моделі), а також для вивчення одночасно структури та поведінки оригіналу (структурно-функціональні моделі).

Зокрема, аналогами лінгвістичного оригіналу дослідник В. Рубашкін уважає такі об'єкти: реально існуючі; штучно створені матеріальні об'єкти; ідеальні математичні чи логіко-математичні конструкції [10].

Важливе значення для нашого дослідження мають гносеологічні аспекти моделювання знань. Візьмемо до уваги ті з них, котрі виокремив згадуваний вище учений В. Рубашкін. На його думку, до них належать такі положення:

– «модель має виступати особливим засобом відображення дійсності, здійснюваного на основі аналогії;

– модель не повинна бути складніша за сам оригінал;

– побудова моделі повинна бути вільною від протиріч (логічно коректною), вичерпною і максимально простою;

– модель повинна мати загальний характер, що дозволяє застосовувати її для опису різних натуральних об'єктів;

– модель повинна володіти пояснювальною силою (експланаторністю), суть якої полягає в здатності моделі розкривати і пояснювати невідомі дотепер властивості натурального об'єкта;

– модель повинна містити в собі евристичні можливості, тобто давати такі знання, які самі стають джерелами нових ідей і теорій» [10].

За В. Рубашкіним, кожна така модель має відображати зміст справжнього об'єкта, для чого, на його думку необхідне виконання певних вимог. По-перше, модель має відображати й відтворювати основні та найхарактерніші риси початкового об'єкта. Саме вони є найважливішими для конкретного експерименту, у той час, як побічні, не важливі та не релевантні з погляду досвіду ознаки, на думку дослідника, модель може взагалі не виокремлюватися і не враховуватися. По-друге, дослідник має добре розумітися на особливостях внутрішньої будови такої моделі, в той

час як про оригінал він повинен мати мінімум знань, що, на думку В. Рубашкіна, «дозволяють припускати, що елементи і їхні відносини в оригіналі, з одного боку, та елементи і їхні зв'язки в моделі, з іншого боку, перебувають у відносинах повної (ізоморфної) або часткової (гомоморфної) подібності. На основі цих даних будується процедура перенесення на оригінал тієї інформації, що була отримана в результаті спостережень над структурою і роботою моделі» [10].

Однак, як уважає дослідник галузі Е. Попов, неможливо задати алгоритм, котрий може бути застосовуваний як універсальний для моделювання процесу розуміння текстів. Адже, на сьогодні розроблені лише такі алгоритми, котрі здатні висвітлити певні конкретні аспекти знань і під час застосування можуть бути ефективними лише для чітко окресленого кола завдань [11].

Питання моделювання знань у тексті, як спосіб вивчити процес узагальнення смислу, нерозривно пов'язане з новою науково-технічною дисципліною – *інженерією знань*, котра виникла на потреби науки та суспільства формалізувати знання на сучасному етапі розвитку систем обробки текстової інформації. Ми погоджуємося з дослідницею І. Замаруєвою, котра стверджує, що на сьогодні проблемне поле цієї науки вже майже окреслене і об'єднує здобутки фахівців суміжних галузей, на основі яких ця дисципліна формувалася, а саме: прикладна інформатика, математика, лінгвістика та психосемантика. Серед основних завдань нової науки інженерії знань є пошук мови логічного систематизування знань, що, на думку дослідників, є важливою ланкою між мовами опису знань, котрі застосовуються під час розроблення сучасних систем штучного інтелекту, і знаннями спеціаліста, котрі вербалізуються на професійному діалекті ПМ [12].

Провівши детальний аналіз стану дослідження наукових праць з ШІ, Е. Скороходько зазначає, що на сьогодні серед інтерес науковців довкола засобів представлення знань є значно вищим, ніж довкола змісту, котрий підлягає формалізації [13].

Однак дослідження формального представлення інформації для систем ШІ бере початок іще з кінця 60-х рр. XX ст. Тоді було розроблено семантичну сітку як один із найефективніших методів опису знань. Нині семантичні сітки з успіхом застосовують в процесах моделювання мовних зв'язків під час створення інтелектуальних інформаційних систем.

Детальний аналіз наукової літератури з теми дав змогу окреслити загальне розуміння сучасних науковців поняття семантичної сітки – це граф, вершини якого відповідають семантичним елементам мови чи тексту, а ребра – семантичним зв'язкам між ними [14]. Дослідники галузі виділяють мовну, мовленнєву та лексичну семантичні сітки. Мовна семантична сітка має на меті фіксацію парадигматичних відношень, мовленнєва – синтагматичні відношення. Вершини лексичної сітки співставляються зі значенням слів [15].

Досліджуючи семантику тексту і його формалізацію, А. Новиков виокремлює такі основні операції під час сіткового моделювання тексту:

1. «Виявлення семантичних зв'язків між елементами тексту.
2. Їх подання у зручній для подальших досліджень формі.
3. Встановлення (з використанням цього подання) закономірностей, що характеризують досліджуваний об'єкт» [15].

У своїй праці «Семиотические основы информатики» Ю. Шрайдер стверджує, що для семантичного аналізу тексту слід враховувати основні його структурні елементи, такі як слово, речення та абзац. Важливим для смислового аналізу текстової інформації, на думку дослідника, є розуміння

того, що семантично пов'язаними в тексті слова є в тому разі, «якщо в описуваній ситуації їх денотати з'єднані деяким відношенням» [16]. Однак автор зауважує, що синтаксичний аналіз не дає повного розуміння змісту тексту, що є необхідною умовою для адекватного змістового аналізу текстової інформації. Адже синтаксичний аналіз покликаний лише встановлювати синтаксичну структуру речення, тоді як для визначення його змістового наповнення необхідний докладний семантичний аналіз.

На думку Ю. Шрайдера, речення є семантичного пов'язаними у тому разі, «якщо їхні денотати, тобто ситуації, описані цими реченнями, зв'язані на предметному рівні, що дозволяє розглядати їх як компоненти єдиної більш великої ситуації» [16]. Дослідник також зауважує, що «семантичний, чи міжфразовий, зв'язок між реченнями, що описують ситуації зі спільним семантичним елементом, іменується *номінативним*, а зв'язок між реченнями, що описують ситуації, зв'язані будь-яким відношенням, – *релятивним*. Семантичну сітку, вершини якої відповідають лексичним одиницям, а ребра – релятивним лексичним і граматичним одиницям, називають *послівною*. Семантична сітка, вершини якої відповідають реченням, а ребра – семантичним відношенням між ними, називається *пофразовою*». [16]. Автор також зазначає, що семантичні сітки з успіхом були застосовані під час опису МФЗ для розробки систем АР.

Як бачимо, питанню семантичних сіток як засобу моделювання знань у тексті приділено досить багато уваги в науковій літературі, однак вони не є єдиним способом рішення питання кодування знань з метою їхньої обробки за допомогою комп'ютерних програм.

Досліджуючи математичні методи розпізнавання змісту тесту, В. Гончаренко звертає увагу на один із способів організації комп'ютерної моделі реального світу, котрий був запропонований ученим М. Мінським. Його суть полягає в тому, що останній вбачав можливість представити

знання у вигляді значної сукупності даних, котрі впорядковані певним чином і є стереотипними ситуаціями. За М. Мінським, таким чином ми отримуємо структури, котрі були названі *фреймами*.

Такі фрейми уявляються як сітка, яка об'єднує вузли, а також відповідні зв'язки між цими вузлами. Виокремлюють *верхні рівні фрейму*, а також *осередки*. Верхні рівні утворюються такими поняттями, які є справедливими весь час відповідно до деякої передбачуваної ситуації, а тому вони є чітко визначеними. Характерними ознаками для осередків, або слотів, є те , що вони мають бути заповнені певними відомостями, тобто містити конкретну інформацію про ситуацію і розміщуватися нижче за рангом, ніж верхні рівні фрейму.

Важливим, на нашу думку, є те, що кожен із терміналів може створювати умови, котрі мають задовольняти його завдання. Існують прості умови та складні. До простих умов дослідник Р. Шенк відносить питання обробки концептуальної інформації відносить ті, котрі «визначаються маркерами, наприклад, у вигляді вимоги, щоб завданням терміналу був який-небудь суб'єкт або предмет придатних розмірів, або вказівка на субфрейм певного типу. Субфрейми, фрейми і суперфрейми – ієрархічно упорядковані елементи, що утворюють системи фреймів» [17]. Щодо більш складних умов, то тут автор зазначає, що ними задаються відношення між поняттями, які були внесені до різних термінальних вершин.

Близькі за семантичним навантаженням фрейми об'єднують у групи, групи – в системи фреймів, а останні пов'язані мережею пошуку інформації [17].

Л. Кокорєва та інші дослідники виділяють такі два види фреймів:

– фрейми – візуальні образи;

– фрейми – сценарії (скрипти) [18].

В. Котов уважає, що теорія машинного представлення текстової інформації за допомогою фреймів де в чому схожа на спосіб людського мислення, і тому може пояснити деякі характерні його риси. Адже ця теорія представлення знань об'єднала в собі чимало досягнень із таких наук, як психологія, лінгвістика та ШІ.

Використання фреймового підходу знайшло відображення в діалогових системах. Адже, як уважають дослідники, кожен пункт діалогу є певним фреймом конкретної діалогової ситуації, а окремий сценарій діалогу має формуватися як система фреймів діалогових ситуацій [19].

Фреймовий спосіб подачі інформації активно застосовують під час вирішення задач опису смислової (семантичної) і синтаксичної структур тексту, а також під час розроблення онтологічних систем.

На думку дослідників, вирішуючи питання створення змістового компонента, який має забезпечувати безперебійну роботу системи машинного реферування тексту, слід опиратися на вивченні глибинних структур зв'язного тексту, звертаючи особливу увагу на його систематизацію. При цьому, як уважає Е. Добрускіна, слід враховувати те, що текст є єдиною системою, що поєднує в собі три різних за своєю природою структур, а саме:

– семіотична (або знакова) система;
– мовна система;
– система знань про світ.

Лише за умови наявності усіх трьох компонентів стає можливим відтворювати справжній зміст тексту, коли після автоматизованого реферування тексту можна буде отримати адекватний реферат, зрозумілий людині і наповнений сенсом початкового тексту.

Під час вивчення природніх текстів значної уваги слід приділити дослідженню та відтворенню міжфразових зв'язків, без яких неможливо створити повноцінний текст [20].

Важливим поняттям у сфері автоматизованого реферування є поняття *синтаксичного міжфразового зв'язку (МФЗ)*. Досить вдале, на нашу думку, визначення МФЗ подають С. Приходько та Е. Скороходько у своїй праці «Автоматическое реферирование на основе анализа межфразовых связей». За ними, МФЗ – «це такий зв'язок між реченнями, коли одне з них є насиченням іншого в плані вираження (експліцитні МФЗ) або в плані змісту (імпліцитні МФЗ)» [21].

Згадуваний нами вище Ю. Шрейдер зазначає, що «кількість міжфразових зв'язків може слугувати критерієм відбору речень у реферат, оскільки вимірює функціональну вагу речення в оригінальному тексті, яка є показником його значущості» [22].

Вирішуючи завдання пошуку основної інформації в тексті, слід починати з визначення основного засобу вираження смислового навантаження тексту. Беручи це до уваги, А. Анісімов розглядає текст як знак, котрий має певну внутрішню будову і подає його схематичне зображення таким чином:

$$T = < \text{И}, \text{М}, \varphi_1 \dots \varphi_m, \Theta > \qquad (2.1)$$

, де Т – текст; И – словник; М – множинність місць; φ – набір відношень на цій множинності; Θ – відображення множинності місць у словнику [23].

Вивчаючи системи автоматизованого реферування, У. Хан звертає увагу на певну комунікативну завершеність тексту, котра визначається конкретною темою, прагматикою, а також творчим задумом його автора. На думку автора, мова не містить формально-структурного інваріанта тексту, тому що структурні властивості кожного тесту можуть значно відрізнятися одна від одної. Мова не містить достатньо стійких

формально-структурних характеристик тексту, що дає змогу говорити про те, що текст є мовленнєвою, а не мовною одиницею. На підтвердження своєї теорії автор наводить аргумент, що за наявності великої кількості типологій текстів, власне лінгвістична типологія відсутня [24].

Однак, на думку дослідника, з цього не випливає, що такий текст не можна аналізувати з лінгвістичної точки зору. Незважаючи на те, що текст не можна розглядати як одиницю опису смислу, його складовими частинами є речення, котрі можуть виступати як «мінімальний текст». Це і спонукає мовознавців розглядати саме речення як одиницю для аналізу. Таким чином речення є не лише мовленнєвою, а й мовною одиницею. Автор зазначає, що «Речення виступає як одиниця тексту, що безпосередньо співвідноситься з мовним актом і разом з тим є мовним засобом вираження думок, ідей. Речення є найбільш значимою структурною текстовою та комунікативною одиницею» [25].

Думки сучасних мовознавців сходяться в тому, що вищенаведені особливості речення роблять його головним структурним компонентом мови. Усі інші структурні одиниці є допоміжними.

Для автоматизованого аналізу тексту те, що кожне речення є не лише мовленнєвою, а й мовною одиницею, є важливим через те, що його структура, на відміну від структури зв'язного тексту, завжди конкретна і має певну відповідність з однією з формально-змістових моделей синтаксичної структури мови.

Наприклад, найпростіша структура розповідного предикативного речення виглядає так: підмет + присудок. Дещо ускладнивши нашу структуру додамо означення перед підметом або обставину перед присудком. Складне речення складається з декількох підметів і декількох присудків: (означення + підмет + присудок) + (підмет + обставина + присудок) тощо. Знання про те, які частини мови якими виступають членами речення дасть певне розуміння структури речення для його

подальшої математичної обробки. Порядок слів (хоч в українській мові він чітко не визначений) також дає певну інформацію про місце певного слова в структурі речення, а отже допоможе описати семантичну його структуру. Однак подібна модель може бути застосована для визначення семантичної структури ряду типових речень, у яких наявний один підмет та один присудок або декілька підметів і декілька присудків (поєднання декількох простих речень). У мові наявна величезна кількість речень зі складними синтаксичними зв'язками, багатовалентністю дієслів тощо, котрі описати дещо складніше.

2.2. Модель виділення ключової інформації

Оскільки речення виступають не тільки як мовленнєва, але й як мовна одиниця і являють собою центральну структурну одиницю мови, створенню якої служать у кінцевому підсумку всі інші компоненти мовної системи, можна припустити, що виділення ключової інформації з тексту – це те саме, що й виділення ключових речень з тексту. Таким чином ми отримуємо задачу пошуку найважливіших речень в тексті.

Цю задачу досліджують науковці, які займаються питаннями і проблемами автоматичного реферування текстів. Більшість з нині наявних методів придатні для створення лише «квазірефератів». Але якщо використовувати модель пошуку, засновану на використанні зв'язності тексту, а також на вираховуванні теми і тематики тексту, можна отримувати замість квазірефератів спавжні реферати.

Для цього можна поєднати метод, що заснований на використанні зв'язності тексту, та один з методів частотного автоматизованого реферування, наприклад метод ключових слів.

Частотне авто реферування – це спосіб створення авторефератів, що використовує тільки кількісні характеристики текстів (наприклад, частоту входження слів). Частотний автореферат складається з деякої сукупності пропозицій, витягнутих з тексту, причому порядок речень автореферату може бути змінений щодо початкового тексту. В результаті, прочитавши автореферат, не завжди вдається правильно зрозуміти його зміст [25].

Зазвичай частотні автореферати складаються на основі частоти входження ключових слів. Це терміни, які частіше за інших зустрічаються в тексті і використовуються для передачі його змісту. Тому в автореферат потрапляють речення, які включають в себе найбільшу кількості ключових слів [25].

Ключовими словами можна вважати не тільки слова, які найчастіше зустрічаються в тексті, а й слова, які є термінами чи основними поняттями, що пов'язані з темою тексту. Слова, які є назвами чи скороченнями, також можна вважати важливими. Звісно, якщо скорочення зустрічається лише один раз його неможна віднести до ключових слів.

Слід також звернути увагу сигнальні фрази. Це слова або словосполучення, які апріорі містять ключову інформацію, мають максимальну вагу і речення, з такими фразами, повинні бути включені в список найважливіших. Наприклад, «висновок», «необхідно підкреслити». Їх також називають маркерами. Для того, щоб виявити таку фразу в тексті, використовують спеціальні словники маркерів [26].

Таким чином поєднавши все вище зазначене можна отримати набір ключових слів та висловів в тексті. Присвоїти цим словам певну вагу залежно від частоти їх застосування в тексті та наближеності до тематики

тексту, можна вибрати ключові слова, за якими в подальшому і буде відбуватись відбір речень.

Для пошуку ключових речень важливо враховувати смислове навантаження кожного речення. Смислове навантаження – набір ключової інформації, яке передає слово, речення тощо. Враховувати смислове навантаження кожного речення можна за допомогою міжфразових зв'язків.

Якщо речення має декілька речень, що доповнюють його зміст або розкривають певну суть, подану в початковому реченні, то кількість таких речень можна вважати числовою характеристикою МФЗ.

Як зазначалось раніше, для опису міжфразових зв'язків можна використовувати семантичні сітки.

Семантична сітка – інформаційна модель предметної області, що має вигляд орієнтованого графа, вершини якого відповідають об'єктам предметної області, а дуги (ребра) задають відношення між ними. Об'єктами можуть бути поняття, події, властивості чи процеси [27].

Таким чином, семантична сітка є одним із способів подання знань. У назві поєднано терміни з двох наук: семантика в мовознавстві вивчає сенс одиниць мови, а сітка в математиці являє собою різновид графа – набір вершин, що з'єднані дугами (ребрами), яким привласнено деяке число. В семантичній сітці роль вершин виконують поняття бази знань, а дуги (зазвичай, спрямовані) задають відношення між ними. Таким чином, семантична мережа відображає семантику предметної області у вигляді понять і відношень [27].

Зазвичай у вузлах в подібних сітках знаходяться слова, означення, в нашому же випадку в вузлах знаходяться речення, а дуги відповідають за зв'язок між цими реченнями. Таким чином вузли з найбільшою кількістю дуг будуть відповідати реченням з найбільшим смисловим навантаженням, тобто найбільш інформативним реченням.

Поєднавши метод синтаксичного МФЗ та метод ключових слів, отримаємо модель аналізу тексту, що дозволяє найбільш чітко визначити ключові фрази та речення в тексті. На основі цієї моделі можна розробити алгоритм вибору ключових речень із тексту.

В основу алгоритму покладемо метод лінійних зважених коефіцієнтів. Визначимо два критерії відбору найбільш інформативних речень: перший – кількість синтаксичних міжфразових зв'язків у речення, другий – наявність у ньому ключових термінів чи понять. Для подальшого виявлення значущих речень будемо використовувати модель лінійних зважених коефіцієнтів. Залежно від кількості ключових термінів і їх частоти згадування в тексті, а також середньої кількості зв'язків у реченнях, цим двом критеріям присвоюється певна оцінка рівня значущості. Далі для кожного речення отримуємо його характеристику інформативності W як суму ключових слів A_i, помножених на коефіцієнт важливості кожного слова k_i, які залежать від кількості згадок терміну в тексті і його наближеності до теми, та кількості зв'язків S, помножених на коефіцієнт зв'язків k:

$$W = \sum_i k_i * A_i + k * S \tag{2.2}$$

, де коефіцієнти задовольняють рівнянню 3:

$$\frac{k}{1{,}25 * i} + \sum_i k_i = 1 \tag{2.3}$$

Вибравши речення з найбільшим значенням характеристики, отримаємо набір найважливіших речень у тексті, розуміння яких дасть найбільш чітку картину для розуміння всього тексту в цілому. Ці речення можна використовувати для створення рефератів, анотацій, для індексування пошуку початкового тексту в Інтернеті та генерації тестів для виявлення якості засвоєння матеріалу тощо.

2.3. Висновки

Удосконалена методика дає можливість ефективно знаходити найбільш інформативні і важливі речення в науково-технічному тексті. Семантико-синтаксичний аналіз дає змогу виявити частини тексту, які є найбільш важливими для автоматичного реферування. Під час застосування вдосконаленої методики виокремлення ключових речень із цієї статті було показано кращий результат, аніж під час застосування окремо інших прийомів.

Використання лінійних зважених коефіцієнтів для обраної моделі дає досить простий, проте ефективний алгоритм відбору речень.

Ця методика підходить для поєднання з методами генерації запитань до речень для створення автоматичної системи генерації відкритих запитань до природо-мовних текстів, що, на нашу думку, є перспективним напрямом дослідження в галузі автоматизованого аналізу текстів.

РОЗДІЛ 3. МЕТОД КЛАСИФІКАЦІЇ РЕЧЕНЬ ЗА ДОПОМОГОЮ НЕЙРОННИХ МЕРЕЖ

Методи автоматизованого аналізу текстової інформації почали досліджувати ще в 60-х рр. XX ст., що було пов'язано з появою великої кількості текстових документів і необхідності їх швидкого аналізу та обробки. На сьогодні розроблено багато методів обробки текстової інформації починаючи з індексування документів в Інтернеті і закінчуючи виділенням ключових елементів з тексту. Питаннями автоматизованої обробки інформації займалися такі вчені, як: Є. Борохова, Х. Лун, М. Жинкіна А. Новикова, В. Рубашкіна, Г. Жданов, H. Kunichika, A. Takeuchi, S. Otsuki та ін.

Серед завдань генерацій текстів розрізняють задачі з генерації тексту «риби» (що не має сенсу), або генерації більш осмисленого тексту. Одним із напрямів в області генерації осмисленого тексту є задача генерації запитань. Існує декілька способів автоматизованого створення запитань: 1) з ключових слів, 2) за заданими критеріями, 3) генерація запитання до наявного речення. У цій роботі будуть представлені методи генерації запитань до речень методом класифікації типу речення.

Найбільш поширеним методом генерації запитань до речення є метод через класифікацію самого речення. Якщо стає зрозумілим тип

речення, то запитання генерується методом заміни декількох слів та адаптацією закінчень у змінюваних частинах мови.

3.1. Метод штучних нейронних мереж

На сьогодні розроблено велику кількість математичних методів, які здатні розв'язувати широкий спектр задач. Одним із таких методів, на рівні з методом «дерево прийняття рішень», методом «генетичні алгоритми», являється метод «штучних нейронних мереж». Нейронні мережі здатні розв'язувати різні задачі: розпізнавання образів, дискримінантного аналізу, а також задачі класифікації і кластеризації.

Сучасні дослідники зазначають, що на створення штучних нейронних мереж учених надихнула природа організації діяльності вищої нервової системи живих істот. Мережеві конфігурації та спосіб роботи машинних логічних систем має аналогію з роботою мозку біологічних видів. Однак, не зважаючи на високі досягнення науки взагалі, включаючи розробку високошвидкісних та високопродуктивних електронних машин та освоєння космосу, людство ще не накопило достатньо знань про особливості функціонування нашого мозку, а отже не має можливості сповна наслідувати його способи організації, сприйняття та обробки інформації. У зв'язку з цим розробники нейронних мереж лише опираються на сучасні знання з біології мозку, шукаючи нові способи створення структур, котрі зможуть виконувати відповідні функції. Застосування такого підходу призводить до того, що біологічна правдоподібність порушується і створюються такі нейронні мережі, котрі

неможливі в живій природі або ж засновані на вірогідних, проте ще не підтверджених припущеннях про роботу мозку [28].

Аналіз наукової літератури з теми дав змогу представити таке визначення штучних нейронних мереж – це «математичні моделі, а також їхня програмна та апаратна реалізація, побудовані за принципом функціонування біологічних нейронних мереж — мереж нервових клітин живого організму. Системи, архітектура і принцип дії базується на аналогії з мозком живих істот. Ключовим елементом цих систем є штучний нейрон як імітаційна модель нервової клітини мозку — біологічного нейрона» [29].

Засновниками теорії штучних нейронних систем можна вважати У. Мак-Каллока та В. Піттса, котрі у своїй публікації припустили, що модель нервової діяльності живих істот можна застосувати під час створення мережі елементів з двома стійкими станами, рисунок 3.1. Кожен такий елемент називається формальним нейроном [30].

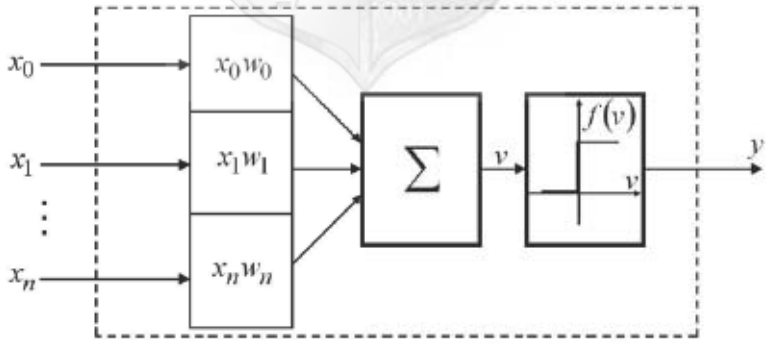

Рисунок. 3.1 – Представлення ШНМ

Основою для кожної ШНМ виступають зазвичай прості та однотипні елементи, або осередки. Їх функцією є імітувати функціонування нейронів мозку вищої нервової системи живих істот. Надалі під нейроном ми

розумітимемо штучний нейрон – елемент ШНМ. На рисунку 3.2 подано структуру такого нейрона у загальному вигляді. Характеристикою кожного нейрона є його поточний стан, який може бути, як і мозку живих істот, збуджений чи загальмований.

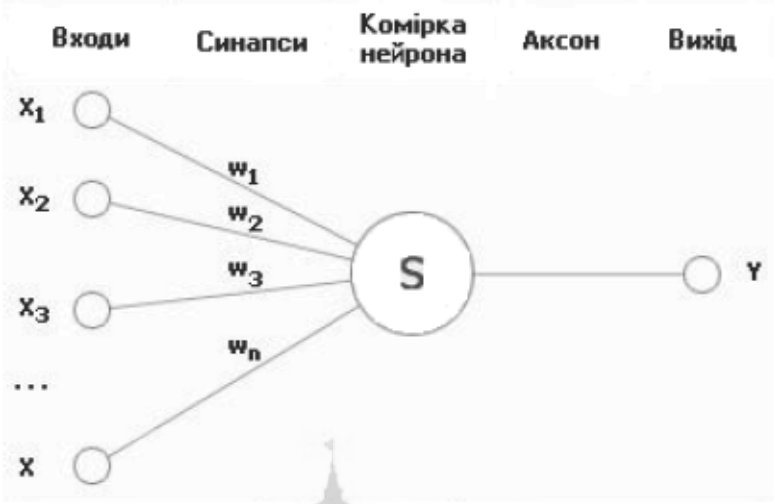

Рисунок 3.2 – Штучна нейронна мережа

Рисунок 3.2 демонструє, що так само, як і біологічний нейрон, штучний нейрон поєднує синапси, котрі мають на меті зв'язувати входи нейрона з ядром, тоді як ядра нейрона оброблюють вхідні сигнали та аксон, що пов'язує цей нейрон з іншими, що знаходяться на наступному шарі.

Кожен синапс має вагу, яка визначає наскільки відповідний вхід нейрона впливає на його стан. Стан нейрона визначається за формулою:

$$S = \sum_{i=1}^{n} x_i w_i \qquad (3.1)$$

де n - число входів нейрону, x_i – значення i-го входу нейрона, w_i – вага i-го синапсу

Значення виходу визначається за формулою:

$$Y = f(S) \tag{3.2}$$

де f – певна функція, що називається передатною.

Найчастіше в якості передатної функції використовують, так звану, сигмоїду, що має наступний вид:

$$f(x) = \frac{1}{1 - e^{ax}} \tag{3.3}$$

Основною перевагою сигмоїди є те, що вона диференційована по всій осі абсцис і має просту похідну:

$$f'(x) = f(x)(1 - f(x)) \tag{3.4}$$

При зменшенні параметра a сигмоїда стає більш пологою, вироджуючись у горизонтальну лінію на рівні 0,5 при a=0. При збільшенні a сигмоїда наближається до функції одиничного стрибка [31].

Кожна мережа призначена для виконання певної функції, тому існує досить багато їх різновидів. Подамо таку класифікацію нейронних мереж:

Нейронні мережі можна класифікувати за видом навчання:

- Мережі, що навчаються з учителем.
- Мережі, що навчаються без учителя.

Також нейронні мережі розрізняються за типом налаштування вагів:

- Мережі з фіксованими зв’язками.

- Мережі з динамічними зв’язками.

Розділяють нейронні мережі за типом вхідних даних:

- Аналогова.
- Двійкова.

Нейронні системи поділяють за кількість прошарків нейронів:

- Одношарові мережі.
- Багатошарові мережі.

Найбільш важливою є класифікація за моделлю нейронної мережі:

- Мережі прямого розповсюдження.
- Рекурентні нейронні системи.
- Радіально базисні функції.
- Карти що самоорганізуються або мережі Кохонена [32].

Для максимальної ефективності нейронної мережі слід встановлювати оптимальну кількість нейронів та застосовувати відповідні зв’язки між ними.

Під час описування нейронних мереж дослідники використовують деякі усталені терміни, значення яких може по-різному трактуватися у різних джерелах. У нашому дослідженні ми візьмемо до уваги таке розуміння найважливіших термінів:

– *Структура нейромережі* – спосіб зв'язків нейронів у нейромережі.
– *Архітектура нейромережі* – структура нейромережі та типи нейронів.
– *Парадигма нейромережі* – спосіб навчання та використання, іноді містить поняття архітектури.

На базі однієї архітектури може бути реалізовано різні парадигми нейромережі і навпаки.

За способом архітектурного рішення нейронні мережі поділяють на дві групи. До першої групи – слабо зв’язані нейронні мережі – належать нейромережі, кожен нейрон кожної пов'язаний лише з сусіднім нейроном (рис. 3.3), а до другої – повнозвязані нейронні мережі – відносять ті з них, у яких входи кожного нейрона пов’язані з виходами усіх інших нейронів (рис 3.4).

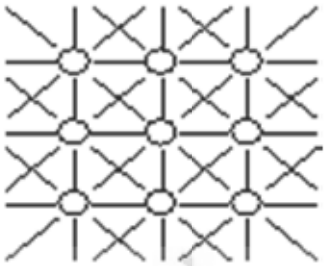

Рисунок 3.3 – Слабозв'язані нейронні мережі

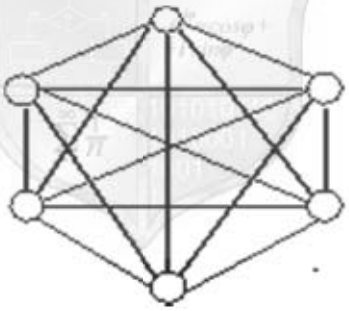

Рисунок 3.4 – Повнозв'язані нейронні мережі

Найбільш поширеним способом внутрішньої будови нейромережі є їх організація у багатошарові мережі. У такому разі усі нейрони об’єднуються у шари, котрих у свою чергу об’єднує спільний вектор вхідних сигналів. На вхідний шар нейронної мережі (або рецептори) подається зовнішній вхідний вектор, а виходами нейромережі служать

вихідні сигналу, котрі знаходяться на останньому прошарку (або ефектори). Слід зазначити, що не всі прошарки нейромережі контактують із зовнішнім середовищем. Кожна мережа повинна мати від одного до декількох прихованих прошарків нейронів.

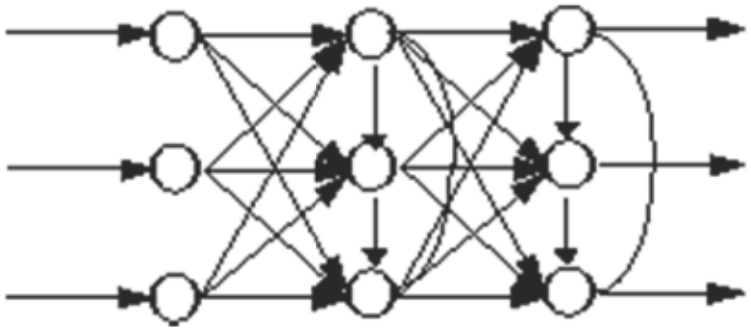

Рисунок 3.5 – Багатошаровий тип з'єднання нейронів

Виділяють такі види зв’язків між нейронами в штучних нейронних мережах:

- Проективні – зв’язки між нейронами різних прошарків;
- Бічні, або літеральні, - зв’язки між нейронами одного прошарку.

На рисунку 5 у схематичному вигляді зображено типову внутрішню будову штучних нейронних мереж. Під час застосування нейромереж зазвичай необхідно, щоб кожна з них мала хоча б три типи прошарків, а саме: вхідний, вихідний та прихований. Однак є такі мережі, котрі містять у собі лише один прошарок, а іноді й один елемент.

«Прошарок вхідних нейронів отримує дані або з вхідних файлів, або безпосередньо з електронних давачів. Вихідний прошарок пересилає інформацію безпосередньо до зовнішнього середовища, до вторинного комп'ютерного процесу, або до інших пристроїв. Між цими двома прошарками може бути багато прихованих прошарків, які містять багато

нейронів в різноманітних зв'язаних структурах. Входи та виходи кожного з прихованих нейронів сполучені з іншими нейронами» .

Розглянемо типові структури нейронних мереж. Виділяють мережі прямого поширення та рекурентні нейромережі.

Мережі прямого поширення відносять до статичних, тут на входи нейронів надходять вхідні сигнали, які не залежать від попереднього стану мережі.

Рекурентні мережі вважаються динамічними, оскільки за рахунок зворотних зв'язків (петель) входи нейронів модифікуються в часі, що призводить до зміни станів мережі.

До мереж прямого поширення належать:

- Перцептрони
- Мережа Back Propagation
- Мережа зустрічного поширення
- Карта Кохонена.

Рекурентні мережі охоплюють такі:

- Мережа Хопфілда
- Мережа Хемінга
- Мережа адаптивної резонансної теорії
- Двоскерована асоціативна пам'ять.

На сьогодні найпоширенішою моделлю для пошуку закономірностей, прогнозування та якісного аналізу є нейронні мережі зворотного поширення (back propagation). Модель була так названа через те, що в ній застосовується специфічний алгоритм навчання, у якому похибка поширюється в напрямку, котрий є протилежним до напрямку поширення сигналу в процесі правильної роботи мережі, тобто від вихідного прошарку до вхідного.

Структура нейронної мережі зворотного поширення така: кожна

мережа містить декілька прошарків нейронів, у якому кожен нейрон певного прошарку *i* пов'язаний з кожним нейроном прошарку *i+1*, тобто ідеться про повнозв'язану мережу [32].

Нейрони впорядковані в пошарову структуру з прямою передачею сигналу. Кожен такий нейрон продукує зважену суму своїх входів, пропускає цю величину через передатну функцію і видає вихідне значення. Мережа може моделювати функцію практично будь-якої складності, причому кількість прошарків і число нейронів у кожному прошарку визначають складність функції. Під час проектування архітектури для визначеної проблеми застосовують такі правила [33]:

1. Кількість входів та виходів мережі визначаються кількістю вхідних та вихідних параметрів досліджуваного об'єкту, явища, процесу, тощо. На відміну від зовнішніх прошарків, число нейронів прихованого прошарку $n_{\text{прих}}$ обирається емпіричним шляхом. В більшості випадків достатня кількість нейронів становить $n_{\text{прих}} = n_{\text{вх}} + n_{\text{вих}}$, де $n_{\text{вх}}$ та $n_{\text{вих}}$ - кількість нейронів у вхідному і у вихідному прошарках.
2. Якщо складність у відношенні між отриманими та бажаними даними на виході збільшується, кількість нейронів прихованого прошарку повинна також збільшитись.
3. Якщо процес, що моделюється, може розділятись на багато етапів, потрібен додатковий прихований прошарок (прошарки). Якщо процес не розділяється на етапи, тоді додаткові прошарки можуть допустити перезапам'ятовування і, відповідно, негативно вплинути на загальне рішення [33].

Після визначення кількості прошарків і кількості нейронів у кожному з них, знаходять значення для синаптичних ваг і порогів мережі, котрі зводять до мінімуму похибку кінцевого результату.

Алгоритми навчання, котрі були розроблені для зменшення похибки, працюють за принципом підгонки мережі до наявних навчальних даних. У цьому разі похибку для певної моделі мережі визначають проходженням через мережу всіх навчальних прикладів, після чого проводиться порівняння спродукованих вихідних даних із бажаними значеннями. Множина похибок створює функцію похибок, значення якої розглядають, як похибку мережі. Як функцію похибок найчастіше використовують суму квадратів похибок .

На нашу думку, щоб краще зрозуміти алгоритм навчання мережі прямого поширення *Back Propagation,* перш за все слід розібратися в понятті поверхні станів. У науковій літературі так пояснюється принцип роботи цього алгоритму: «Кожному значенню синаптичних ваг і порогів мережі (вільних параметрів моделі кількістю *N*) відповідає один вимір в багатовимірному просторі. *N*+1-ий вимір відповідає похибці мережі. Для різноманітних сполучень ваг відповідну похибку мережі можна зобразити точкою в *N*+1-вимірному просторі, всі ці точки утворюють деяку поверхню - поверхню станів. Мета навчання нейромережі полягає в знаходженні на багатовимірній поверхні найнижчої точки» [34].

Поверхня станів має складну будову і досить неприємні властивості, зокрема, наявність локальних мінімумів (точки, найнижчі в своєму певному околі, але вищі від глобального мінімуму), пласкі ділянки, сідлові точки і довгі вузькі яри. Аналітичними засобами неможливо визначити розташування глобального мінімуму на поверхні станів, тому навчання нейромережі по суті полягає в дослідженні цієї поверхні. Відштовхуючись від початкової конфігурації ваг і порогів (від випадково обраної точки на поверхні), алгоритм навчання поступово відшукує глобальний мінімум. Обчислюється вектор градієнту поверхні похибок, який вказує напрямок найкоротшого спуску по поверхні з заданої точки. Якщо трошки просунутись по ньому, похибка зменшиться. Зрештою алгоритм

зупиняється в нижній точці, що може виявитись лише локальним мінімумом (в ідеальному випадку - глобальним мінімумом). Складність тут полягає у виборі довжини кроків. При великій довжині кроку збіжність буде швидшою, але є небезпека перестрибнути рішення, або піти в неправильному напрямку. При маленькому кроці, правильний напрямок буде виявлений, але зростає кількість ітерацій. На практиці розмір кроку береться пропорційним крутизні схилу з деякою константою - швидкістю навчання. Правильний вибір швидкості навчання залежить від конкретної задачі і здійснюється дослідним шляхом. Ця константа може також залежати від часу, зменшуючись по мірі просування алгоритму .

Алгоритм діє ітеративно, його кроки називаються епохами. На кожній епосі на вхід мережі по черзі подаються всі навчальні приклади, вихідні значення мережі порівнюються з бажаними значеннями і обчислюється похибка. Значення похибки, а також градієнту поверхні станів використовують для корекції ваг, і дії повторюються. Процес навчання припиняється або коли пройдена визначена кількість епох, або коли похибка досягає визначеного рівня малості, або коли похибка перестає зменшуватись (користувач переважно сам вибирає потрібний критерій зупинки) [34].

Алгоритм навчання мережі такий [33]:

1. Ініціалізація мережі: вагові коефіцієнти і зсуви мережі приймають малі випадкові значення.
2. Визначення елемента навчальної множини: (вхід - вихід). Входи $(x_1, x_2 \dots x_N)$, повинні розрізнятися для всіх прикладів навчальної множини.
3. Обчислення вихідного сигналу:

$$S_{i_m} = \sum_{i_{m-1}}^{N_{m-1}} W_{i_m j_{m-1}} y_{i_{m-1}} - b_{i_m} \quad (3.5)$$

$$y_{i_m} = f(S_{j_m}) \quad (3.6)$$

$$i_m = 1,2,\dots,N_m, m = 1,2,\dots,L \quad (3.7)$$

Де S - вихід суматора, w - вага зв'язку, y - вихід нейрона, b - зсув, i - номер нейрона, N - число нейронів у прошарку, m - номер прошарку, L - число прошарків, f- передатна функція.

4. Налаштування синаптичних ваг:

$$w_{ij}(t+1) = w_{ij}(t) + r g_j x_i' \quad (3.8)$$

де w_{ij} - вага від нейрона i або від елемента вхідного сигналу i до нейрона j у момент часу t, x_i' - вихід нейрона i, r - швидкість навчання, g_j - значення похибки для нейрона j.

Якщо нейрон з номером j належить останньому прошарку, тоді

$$g_j = y_j(1 - y_j)(d_j - y_j) \quad (3.9)$$

де d_j - бажаний вихід нейрона j, y_j - поточний вихід нейрона j.

Якщо нейрон з номером j належить одному з прошарків з першого по передостанній, тоді

$$g_j = (1 - x'_j) \sum_k g_k w_{jk} \tag{3.10}$$

де k пробігає всі нейрони прошарку з номером на одиницю більше, ніж у того, котрому належить нейрон j.

Зовнішні зсуви нейронів b налаштовуються аналогічним образом.

Тип вхідних сигналів: цілі чи дійсні.

Тип вихідних сигналів: дійсні з інтервалу, заданого передатною функцією нейронів.

Тип передатної функції: сигмоїдальна. В нейронних мережах застосовуються кілька варіантів сигмоїдальних передатних функцій.

Функція Ферми (експонентна сигмоїда):

$$\mathrm{f(S)} = \frac{1}{1 + \mathrm{e}^{-2\mathrm{aS}}} \tag{3.11}$$

де s - вихід суматора нейрона, α - деякий параметр.

Раціональна сигмоїда:

$$f(S) = \frac{S}{|S| + a} \tag{3.12}$$

Гіперболічний тангенс:

$$f(S) = \text{th}\frac{S}{a} = \frac{e^{\frac{S}{a}} - e^{-\frac{S}{a}}}{e^{\frac{S}{a}} + e^{-\frac{S}{a}}} \tag{3.13}$$

Наведені функції належать до одно параметричних функцій. В кожній із них значення залежить від аргументу та одного параметра. Нині також можуть застосовуватися багато параметричні передатні функції, типу [04]:

$$f(S) = p_1 \frac{S}{|S| + p_2} + p_3 . \tag{3.14}$$

3.2. Класифікація речень для генерації запитань

В наш час більшість дослідників, які займаються генерацією запитань до речень використовують системи, які засновані на методі дерев прийняття рішень. Це пов'язано з тим, що представлення речення в вигляді графа є достатньо простою і зрозумілою задачею, також більшість таких систем та досліджень проводяться для мов з чітким синтаксичним порядком слів у реченні. На жаль даний підхід погано працює для української або російської мови.

При класифікації речень за допомогою нейронних мереж можна натрапити на ряд проблем. По-перше нейронні мережі не можуть приймати на вхід слова, вони здатні працювати або з векторами, або з числами. По-друге речення представляють собою структури не визначеної довжини, в той час, як нейронні мережі прямого поширення мають чіткий розмір вхідного вектора.

Для вирішення першої проблеми існує декілька підходів. Можна пронумерувати кожне слово в реченні, відповідно до його порядкового номера в словнику. Цей підхід є поганим через те, що слова, які є різними можуть опинитись близько, з точки зору метрики простору, в той час як схожі за значенням слова опиняться далеко. Це ускладнить класифікацію і не призведе до жодних позитивних результатів. Другим підходом до представлення слів в числовому вигляді є метод, коли кожному слову присвоюється двійковий код, в якому одиниця стоїть в розряді який відповідає порядковому номеру слова в словнику. Цьому підході всі слова є рівно віддаленими, це набагато краще ніж перший підхід, але залишається проблема «не близькості».

Існує і третій варіант розв'язання проблеми з форматом даних – це кодування слів за допомогою алгоритмів, які дозволяють зберегти максимальну наближеність між спорідненими словами. Зараз існую декілька систем, які реалізують такі алгоритми. Ці системи використовують в собі методи кластеризації слів. Однією з таких систем є система кодування слів Word2Vec, розроблена компанією Google. Ця система дозволяє представити слово в вигляді десяткового числа з 0 в цілій частині і 6 знаками після крапки.

Друга проблема в використанні нейронних мереж для класифікації речень, що пов'язана з розмірністю вхідної інформації, має два варіанта вирішення.

Перший пов'язаний з представленням слів в реченні в вигляді вектора фіксованої довжини (довжина визначається розміром найдовшого з можливих речень, в науково-технічній літературі, зазвичай не більше 50, в той час як в художній – до декількох сторінок) а елементами вектора є числова характеристика слова.

Другий варіант був запропонований дослідником Yoon Kim [35]. В цій роботі дослідник запропонував подавати на вхід загортальної нейронної мережі слова послідовно попарно з'єднані, приєднуючи наступне слово в хвіст попередньому (спочатку перше і друге, потів друге і третє і т.д.). Цей принцип показано на рис. 3.6.

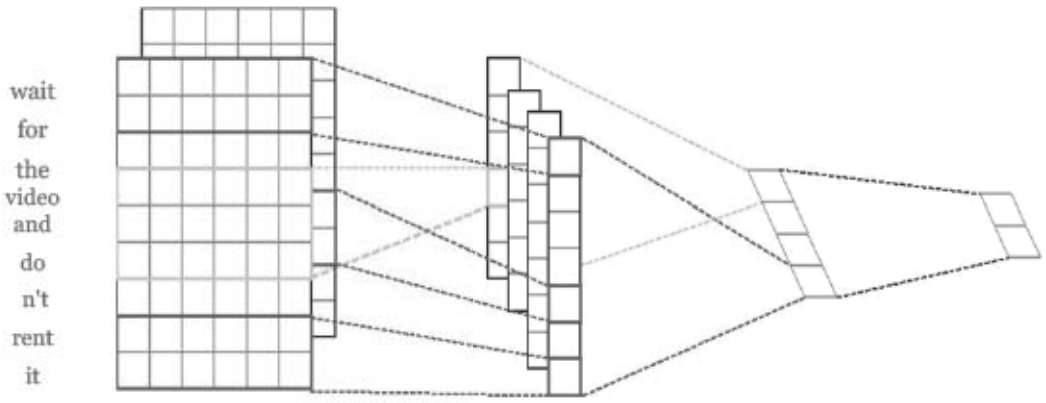

Рисунок 3.6 – Принцип згортання слів

Звісно, для української мови, як і для більшості інших, неможливо передбачити найбільшу довжину речення. Але якщо враховувати, що ця робота присвячена системі, яка більше підходить для науково-технічних текстів (метод пошуку розроблювався саме для такої літератури) то тут можна зробити припущення про найбільшу довжину речення. Оскільки в наукових текстах важливо передавати інформацію так, щоб читач її зрозумів, речення не можуть бути дуже довгими. З легкістю можна обмежитись 50 словами. Це також випливає з методу відбору речень, адже чим більше речення тим менше у нього доповнень і тим менша кількість МФЗ.

Отже, в якості класифікатора будемо використовувати нейронну мережу зворотного поширення похибки з двома схованими прошарками нейронів. В якості передавальної функції пропонуємо обрати функцію Ферми. Для представлення слів в вигляді речень будемо використовувати систему Word2Vec, та у випадку, коли слово є ключовим, для нашого тексту (див. розділ 2), будемо додавати 0.0000005.

Тепер розглянемо класи, на які будуть поділятись вхідні речення. Ці

класи повинні відображати типи запитань, які найдоцільніше буде поставити до речення, щоб перевірити степінь засвоєння інформації. Оскільки метод пошуку ключової інформації та метод класифікації орієнтовані на науково-технічні тексти, класи запитань будуть також орієнтовані на них.

На першому рівні класифікації будемо виділяти такі класи:

- Речення з визначенням чи означенням
- Речення, що містять класифікації
- Просто речення з ключовими словами
- Всі інші речення

Такі класи дають можливість виділяти в першу чергу речення, які легко трансформувати в питальні. Лише останній клас може викликати труднощі в генерації запитання до нього, але тут нам допоможе другий рівень класифікації і другий прошарок прихованих нейронів.

Речення з визначеннями чи означеннями пропонуємо поділяюті на:

- Речення в яких явно присутній означуваний термін
- Речення в яких явно не присутнє означуване слово

Речення що містять класифікацію поділяються на:

- З означеною кількістю на яку поділяють
- З неозначеною кількістю на яку поділяють

Речення з ключовими словами пропонуємо розділити на:

- Речення з одним ключовим словом
- Речення з декількома ключовими словами

У всі інших реченнях будемо виділяти наступні підкласи:

- З ключовим підметом
- З ключовим дієсловом
- З ключовими другорядними членами

Таким чином ми маємо набір з 4 класів і 9 підкласів, які дають можливість достатньо просто перетворити вхідне речення в запитання.

Навчання мережі проводилось на вибірці з 500 речень, які відносились до різних класів. Після навчання мережа була здатна класифікувати речення з науково технічних текстів з вірогідністю 80% для класів і приблизно 62% для підкласів. Що є достатньо хорошим результатом.

3.3. Висновки

Вивчивши існуючи методи класифікацій речень було розроблено варіант набору класів для генерації запитань з речень, та запропонована модель класифікації речень. Отримана модель є найбільш простим варіантом для реалізації. Треба враховувати, що всі міркування, які призвели до вибору саме такого підходу до класифікації, ґрунтувались на припущенні що будуть оброблюватись виключно наукові тексти.

Також потрібно звернути увагу на те, що вхідні речення, які будуть класифікуватись на основі запропонованої моделі повинні містити не більше 50 слів. Тим не менш, запропонований варіант класифікації дає прийнятні результати при тестуванні.

РОЗДІЛ 4. ПРОГРАМНА РЕАЛІЗАЦІЯ ТА ТЕСТУВАННЯ

За умови появи великої кількості нової наукової та навчальної літератури виникає необхідність в контролі засвоєння вивченої літератури. Це стосується як вчителів та викладачів, так і людей що займаються самонавчанням. Саме тому в наш час з'явилась необхідність в автоматизованій системі, які б дозволяла генерувати запитання до тексту. Нажаль готових програмних продуктів які повністю вирішували б цю задачу не існує, адже подібна задача включає в себе не тільки генерацію запитань, а й пошуку ключових речень, до яких будуть ставитись запитання.

У попередніх розділах була розроблена методика відбору ключової інформації в тексті, та модель класифікації речень по типу запитання, яке найдоцільніше поставити до вхідного речення. На базі вище описаних методів було розроблено програмні модулі для автоматизованої системи генерації відкритих запитань до природо-мовних текстів.

4.1. Опис програмного продукту

Автоматизована система генерації запитань до природо-мовних текстів очевидно складається з трьох основних модулів:

1) Модуль синтаксичного та семантичного розбору речень.
2) Модуль вибору ключової інформації з тексту.
3) Модуль генерації запитань до обраних речень.

Перший модуль (синтаксичного та семантичного розбору) повинен відповідати за повний аналіз кожного речення та слова. В даній роботі цей модуль не розглядається і не реалізується. Оскільки цей модуль досить складний і потребує великої бази даних слів, а при реалізації цього модуля для кількох мов кількість необхідних даних збільшується в багато раз. Подібні системи є реалізовані для англійської та німецької мови наприклад Functional Dependency Grammar, Link Parser та інші. Схожі системи розроблюються для російської та української мови. Примітивний варіант такої системи вбудований в Microsoft Word.

Другий модуль (вибору ключової інформації з тексту) реалізовує запропоновану вище модель пошуку ключової інформації. Цей модуль відповідає за вибір найбільш інформативної інформації з тексту.

Третій модуль (генерації запитань до обраних речень) реалізовує систему класифікації речень, за допомогою нейронних мереж, та генерації запитання за допомогою заміни частини речення.

Програмні модулі написані на мові програмування Java. Для роботи з документами Microsoft Word використовується Apache POI бібліотека,

розроблена компанією Apache. В якості нейронних мереж використовувалась бібліотека Fast Artificial Neural Network Library.

Принцип роботи автоматизованої системи та зв'язків між модулями можна побачити на рис 3.

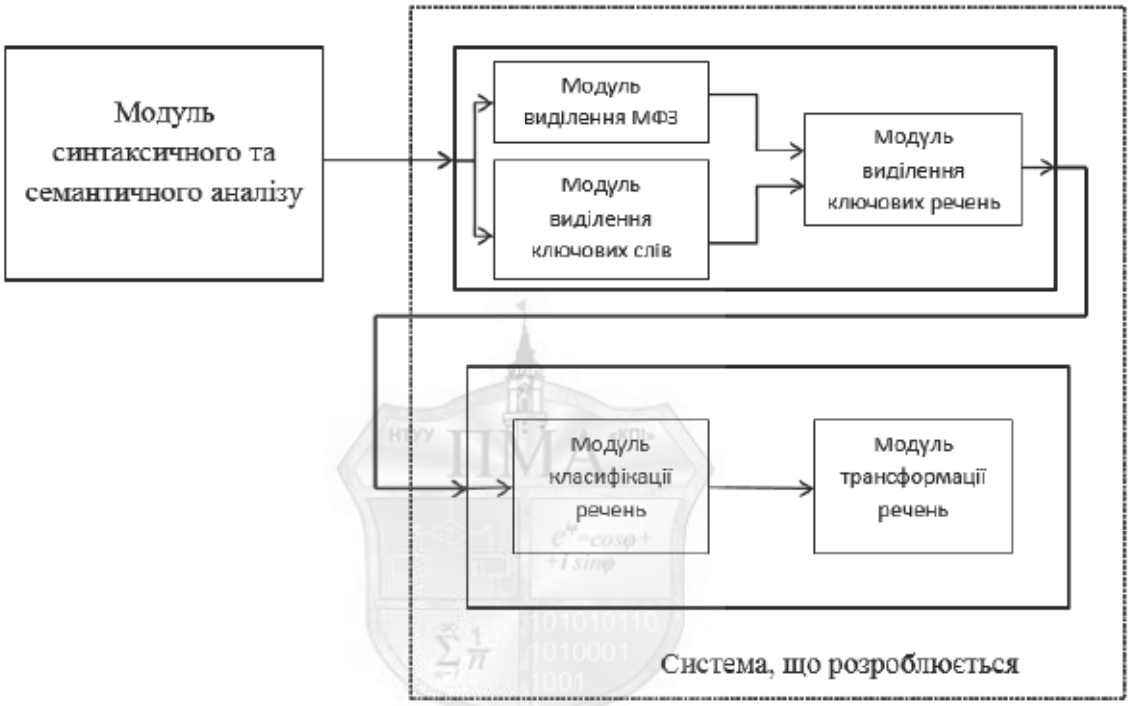

Рисунок 4.1 – Зв'язки між модулями

В результаті роботи модуля синтаксичного та семантичного розбору буде отримано два файли:

1) Перший, Microsoft Word файл з розширенням doc, відповідає вхідному файлу, де перед кожним реченням знаходяться два числа в наступних скобках - []. Перше число відповідає за кількість міжфразових зв'язків у речення, а друге число відповідає за кількість ключових слів у реченні. Також всі знаки, які демонструють

закінчення речення повинні бути замінені на {@}, а всі «–»в означеннях на «це» як це зроблено в прикладі.

2) Другий файл, файл характеристики слів, повинен мати інформацію про всі слова в реченнях, а саме – їх приналежність до певної частини мови, чи є вони ключовими для даного тексту та синтаксичну роль в реченні.

Вхідною інформацією для модуля вибору ключової інформації з тексту є отриманий в процесі роботи першого модуля Microsoft Word документ, що відповідає вхідному файлу.

Використовуючи запропонований метод пошуку ключової інформації другий модуль виділить певну кількість речень. Кількість цих речень залежать від вхідного параметру. Кількість обраних речень буде не менше ніж задано вхідним параметром. Причому речення будуть скомпоновані в list-ому java об'єкту в порядку важливості цього речення, найбільш важливі йдуть раніше.

Третій модуль отримує на вхід набір виділених речень в вигляді list-ого, та другий файл, отриманий після роботи модуля синтаксичного та семантичного розбору. Третій модуль за рахунок нейронних мереж класифікує вхідні речення. Вхідною інформацією для нейронних мереж і є данні з файлу характеристики слів, а також результат роботи системи Word2Vec, безкоштовна система для представлення слів в вигляді вектора, розроблена компанією google. Таким чином, визначивши клас речення, модуль трансформує його в питальне речення певного типу. Результатом роботи модуля є документ Micrtosoft Word, що містить сформовані запитання.

4.2. Тестування програмного продукту

Для тестування розроблених модулів було використано текст тез доповіді, що публікувалась на конференції ПМК-2015 [37]. За для зручності тестування було створено простий інтерфейс рис 4.2. В полі InputFile вводиться повний путь до розібраного файлу. В полі Number Of Question вводиться мінімальна кількість запитань, що буде створена. Поле Characteristic File заповнюється інформацією про путь до файлу характеристики.

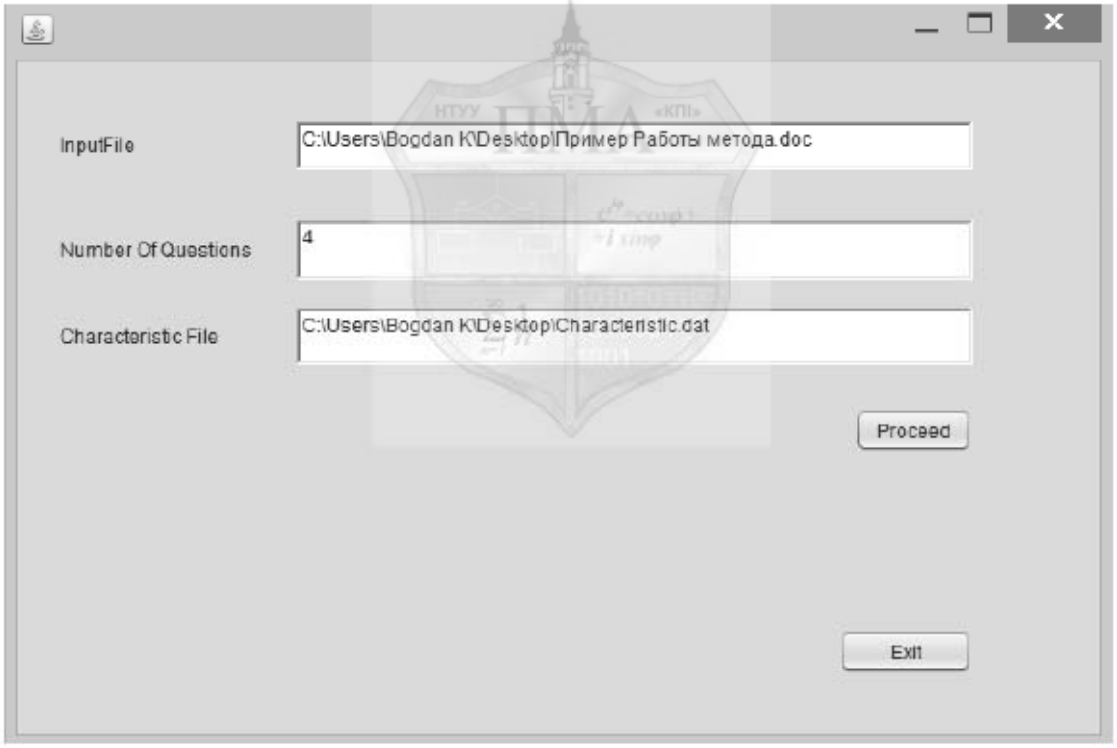

Рисунок 4.2 – Екранна форма для тестування

Приклад змісту одного з розібраних вхідних файлів, на яких проводилось тестування:

«[2][0] Детальний аналіз наукових джерел дає змогу констатувати, що на сьогодні розроблено велику кількість методів автоматизованого реферування текстів, класифікувати які, на думку дослідників, досить складно{@} [1][0] Це пов'язано, зокрема, з тим, що вчені все частіше використовують комбіновані класифікації, намагаючись поєднати переваги різних підходів{@}

[1][0] Досить докладну, на нашу думку, класифікацію методів реферування представив у своїй роботі В. Є. Берзон{@} [2][6] Дослідник виділяє такі методи: статистичні, позиційні, дескрипторні, анкетні, засновані на статистичних зв'язках елементів тексту, семантичні, синтаксичні та засновані на дослідженні структури зв'язного тексту{@}

[1][0] Важливим, на наш погляд, теоретичним результатом пошуку у сфері автоматизованого реферування текстів є створення алгоритму аналізу початкового тексту, який запропонувала Г. С. Жданова{@} [3][0] В основу методу покладено імітацію інтелектуального процесу реферування, під час якого з тексту виокремлюється інформація відповідно до ступеня її важливості{@} [1][0] Щоб визначити важливість інформації необхідно розв'язати задачу семантичного та синтаксичного розбору тексту{@}

[2][1] Під час дослідження синтаксичного розбору тексту особливу увагу слід приділити вивченню механізму міжфразових зв'язків як текстоутворювального чинника{@} [4][1] Синтаксичний міжфразовий зв'язок (МФЗ) це це такий зв'язок між реченнями, в якому одне з них є насиченням іншого в плані вираження (експліцитні МФЗ) або в плані змісту (імпліцитні МФЗ){@} [2][0] Кількість таких міжфразових зв'язків може слугувати критерієм відбору речень, які несуть найбільшу смислову значущість, оскільки вимірює функціональну вагу речення в оригінальному тексті, яка є показником його значущості{@}

[3][1] Іншим способом виявлення важливих за семантичним навантаженням речень є наявність у них термінів або ключових слів, котрі співпадають з темою тексту{@} [2][0] Особлива увага приділяється першій згадці наукового терміна чи скорочення в тексті, адже вона може вміщувати визначення цих термінів чи їх характеристики є важливою інформацією{@} [1][0] Це дає можливість не втратити відомості про ключові поняття{@}

[3][2] Поєднавши метод синтаксичного МФЗ та метод ключових термінів, отримаємо модель аналізу тексту, що дозволяє найбільш чітко визначити ключові фрази та речення в тексті{@} [2][0] На основі цієї моделі розробимо алгоритм вибору ключових речень із тексту{@}

[1][1] В основу алгоритму покладемо метод лінійно зважених коефіцієнтів{@} [4][2] Визначимо два критерії відбору найбільш інформативних речень: перший це кількість синтаксичних міжфразових зв'язків у речення, другий це наявність у ньому ключових термінів чи понять{@} [3][1] Для подальшого виявлення значущих речень будемо використовувати модель лінійних зважених коефіцієнтів{@} [2][0] Залежно від кількості ключових термінів і їх частоти згадування в тексті, а також середньої кількості зв'язків у реченнях, цим двом критеріям присвоюється певна оцінка рівня значущості{@} [4][1] Далі для кожного речення отримуємо його характеристику інформативності W як суму ключових слів A_i, помножених на коефіцієнт важливості кожного слова k_i, і кількості зв'язків S помножених на коефіцієнт зв'язків k{@}

[2][0] Вибравши речення з найбільшим значенням характеристики отримаємо набір найважливіших речень у тексті, розуміння яких дасть найбільш чітку картину для розуміння всього тексту в цілому{@} [1][0] Ці речення можна використовувати для створення рефератів, анотацій, для

індексування пошуку початкового тексту в Інтернеті та генерації тестів для виявлення якості засвоєння матеріалу тощо{@} 10»

В результаті обробки цього тексту програмними модулями по вибору ключової інформації з тексту та генерації запитань до обраних речень було отримано файл з наступним набором запитань:

Які методи виділяють дослідники?

Які два критерії визначені для відбору речень?

Синтаксичний міжфразовий зв'язок (МФЗ) - це?

Поєднавши які два методи, отримаємо модель аналізу тексту, що дозволяє найбільш чітко визначити ключові фрази та речення в тексті?

Як отримати для кожного речення його характеристику інформативності?

Який інший спосіб виявлення важливих за семантичним навантаженням речень?

Для подальшого виявлення значущих речень будемо використовувати яку модель?

За рахунок отриманих запитань досить легко виявити степінь засвоєння вхідного тексту користувачем. Саме це і є ключовим завданням для системи, що розроблювалась.

4.3. Висновки

Програмні модулі, розроблені в процесі виконання магістерського дослідження, показали адекватність і коректність запропонованих методик та моделей для створення автоматизованої системи генерації запитань.

Як і очікувалось, виходячи з теоретичних міркувань, програмний модуль, створений на основі запропонованої методики виявлення ключової інформації, дав кращі результати, ніж схожі системи автоматизованого реферування для наукових текстів. Для художніх та публіцистичних текстів розроблена методика і програмний модуль не показали жодних покращень, порівняно з більшістю раніше розроблених.

Класифікатор на основі нейронних мереж дав 80 % правильних класифікацій в класи та 62 % під час класифікації на підкласи. Ці результати є досить гарними, адже найкращі результати, які показували інші класифікатори становить 83 % для бінарної класифікації.

ВИСНОВКИ

Питання автоматизованої обробки тексту стає все більш популярним серед науковців. Задачами пов'язаними з автоматизацією аналізу текстів займаються переважно закордонні науковці, але це питання увійшло до кола інтересів вітчизняних дослідників. Здебільшого розробки в цій сфері стосуються задач автоматизованого реферування, задач генерації тексту «риби» чи систем питання–відповідь. Також досить багато уваги приділено розробці питання класифікації текстів.

Аналіз наукових джерел з автоматизованого аналізу текстів засвідчив, що більшість досліджень, які зараз проводяться, направлені на роботу з текстами з чітким порядком слів. Для розв'язання завдань дослідження ми виходили з того, що подібні задачі для іншого виду мов, наприклад таких як українська, часто вимагають інших підходів і методик. Це призводить до необхідності розробки або універсальних методів, або адаптації раніше розроблених до специфіки української та подібних мов.

Ґрунтовний аналіз наявних методів автоматизованого реферування показав наявність досить цікавих новітніх методик, які показують гарні результати, але мають певні вади. Ми виявили, що їх можна усунути за рахунок поєднання декількох методів, з різними підходами. Багато дослідників пропонують саме методики засновані на поєднанні декількох методів. Виходячи з потреб системи генерації запитань ми запропонували

поєднати методи міжфразових синтаксичних зв'язків та метод ключових слів.

Дослідження зв'язного тексту та виокремлення знань з нього підтвердило припущення про можливість використання речень як ключового носія інформації і знань у тексті. Це дало змогу використовувати пошук ключових речень в тексті для виявлення важливих знань. Саме на виділенні певних речень для передачі інформації початкового тексту під час генерування реферату ґрунтуються методи автоматизованого реферування.

На основі дослідження було розроблено нову методику пошуку ключової інформації в наукових текстах. Її суть полягає в поєднанні методу МФЗ та методу ключових слів для виявлення найважливіших речень з тексту. Поєднання методів виконувалось з урахуванням того, що виявлення ключових речень буде проводитись в наукових текстах і що далі з них генеруватимуться запитання. Запропонована методика дозволяє враховувати інформативність кожного речення, а також тематичне спрямування самого тексту.

Вивчення питання генерації запитань показало наявність декількох напрямів дослідження: генерації запитань на певну тему та генерації запитання до наявного речення. Кожен з напрямів має свої методики вирішення поставленої задачі. Генерація запитання до наявного речення зазвичай розв'язується представленням речення в вигляді зв'язного графа. Такий метод більше підходить для мов з чітким порядком слів у реченні. Також існує підхід, коли питання генерується за рахунок кластеризації речень. Цей підхід є більш універсальним, але складнішим у розробці та застосуванні.

Дослідивши наявні методи кластеризації даних ми отримали ґрунтовне розуміння переваг і недоліків кожного з методів. Завдяки збільшенню обчислювальних можливостей комп'ютерів з'являється

можливість використовувати складніші з точки зору реалізації і роботи методи з метою отримання кращих результатів. Тому зараз вчені все більше намагаються використовувати нейронні мережі для задач класифікації даних, в тому числі і текстових.

В роботі подано результати розробки методу класифікації речень за допомогою нейронних мереж з метою подальшої генерації запитань. Частка правильних класифікації під час застосування запропонованого методу становить 62 %, що є гарним результатом для такої задачі. Також ми виявили, що під час його застосування в поєднанні з розробленою методикою виділення ключової інформації, отриманий результат значно покращився.

На основі розроблених методик було створено програмні модулі, здатні виявляти ключові речення в тексті та генерувати запитання до них. Саме за допомогою цих модулів було проведено тестування результатів досліджень цієї дисертації.

Дослідження та розробки, представлені в цій роботі, є актуальними через відсутність розробок подібних систем та перспективними, адже вони є універсальними і можуть використовуватись для різних типів мов. Наразі є можливість подальшого удосконалення методу класифікації за рахунок збільшення кількості прошарків нейронів або зміни нейронної системи. У найближчій перспективі нам бачиться розширення розробленої системи для створення повноцінної системи автоматизованої генерації тестів.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Husam Ali Automatic Question Generation from Sentences [Електроний ресурс] — Режим доступу: http://www.iro.umontreal.ca/~felipe/TALN2010/Xml/Papers/all/taln2010_submission_172.pdf
2. Anthony Khoo Experiments with Sentence Classification [Електроний ресурс] — Режим доступу: http://www.aclweb.org/anthology/U06-1005
3. Yoon Kim Convolutional Neural Networks for Sentence Classification [Електроний ресурс] — Режим доступу: http://emnlp2014.org/papers/pdf/EMNLP2014181.pdf
4. Берзон В. Е. Классификация коннекторов и диалоговые системы автоматического реферирования [Текст] / В. Е. Берзон, А. Б. Брайловский // НТИ. Сер. II. – 1979. – № 11. – С. 19–23.
5. Леонов В. П. О методах автоматического реферирования [Текст] / В. П. Леонов // НТИ. Сер. II. – 1975. – № 6. – С. 16 – 20.
6. Лазаренко О.В. Моделювання процесу узагальнення в системі автоматичного реферування : [Монографія] / О.В. Лазаренко, А.А. Яковенко. - Х. : Вид-во НУА, 2007. - 123 с.
7. Програмуємо колективний розум [Текст] / Тоби Сегаран: переклад з англійської А. Слинкина // Символ-Плюс, 2008. – 368 с.
8. Технический справочник по алгоритму дерева принятия решений [Електронний ресурс] — Режим доступу: http://msdn.microsoft.com/ru-ru/library/cc645868(v=sql.105).aspx.
9. Пиотровский Р. Г. Методы автоматического анализа и синтеза текста: [Учеб. пособие для ин-тов и фак. иностр. яз.] / Р.Г. Пиотровский, В.Н. Билан, М.Н. Боркун, А.К. Бобков. – Мн.: Выш. шк., 1985. – 222 с

10. Рубашкин В. Ш. Представление и анализ смысла в интеллектуальных информационных системах [Текст] / М.: Наука, 1989. – 192 с.
11. Попов Э. В. Общение с ЭВМ на естественном языке [Текст] / М.: Наука, 1982. – 360с.
12. Замаруєва І. Мова – інформація – знанням [Текст] / Інформація та нові технології. – К. – 1996. – № 1. – С. 2 – 4.
13. Скороходько Е. Ф. Сіткове моделювання лексики: лінгвістична інтерпретація парметрів семантичної складності [Текст] / Мовознавство. – 1995, № 6. – С. 19 – 28.
14. Минский М. Фреймы для представления знаний [Текст] / Энергия, 1975. – 152 с.
15. Новиков А. И. Семантика текста и ее формализация [Текст] / А. И. Новиков // – М.: 1983. – 215 с.
16. Шрейдер Ю. А. Семиотические основы информатики [Текст] / Ю. А. Шрейдер // ИПКРИР. – М., 1974. – 81 с.
17. Шенк Р. Обработка концептуальной информации [Текст] / Р. Шенк // – М.: Энергия, 1980. – 361с.
18. Кокорєва Л. В. Діалогові системи та представлення знань [Текст] / Л. В. Кокорєва, О. Л. Перевозчикова, К. Л. Ющенко: АН України. Ін-т кібернетики. – К.: Наук. думка, 1992. – 448 с.
19. Котов В. Е. Сети Петри [Текст] / В. Е. Котов // – М.: Наука, 1984. – 160 с.
20. Добрускина Э. М. Синтаксические сверхфразовые связи и их инженерно-лингвистическое моделирование [Текст] / Э. М. Добрускина, В. Е. Берзон // Кишинев: Штиинца, 1986. – 167 с.
21. Приходько С. М. Автоматическое реферирование на основе анализа межфразовых свіязей [Текст] / С. М. Приходько, Э. Ф. Скороходько // НТИ. – Сер. 2, № 1, 1982 – С. 27 – 31.

22. Шрейдер Ю. А. Семиотические основы інформатики [Текст] / Ю. А. Шрейдер // ИПКРИР. – М., 1974. – 81 с.
23. Анисимов А. В. компьютерная лингвистика для всех: Мифы. Алгоритмы. Язык. [Текст] / А. В. Анисимов // – К.: Наук. думка, 1991. – 208 с.
24. Хан У.. Системы автоматического реферирования [Електроний ресурс] — Режим доступу: http://www.osp.ru/os
25. Третьяков Ф.И. Методы автоматического построения рефератов на основе частотного анализа текстов [Електроний ресурс] /Доклады БГУИР // Минск 2003 – С. 40-43
26. Серебряная Л.В., Чебаков С.В. // Информатизация образования. 2011. № 2. С. 52–61.
27. Продукційна модель представлення знань [Електроний ресурс] — Режим доступу: http://victoria.lviv.ua/html/ai/knowledge.html
28. Основи штучних нейронних мереж [Електроний ресурс] — Режим доступу: http://www.victoria.lviv.ua/html/wosserman/rozdil1.htm#r1_1
29. Штучні нейронні мережі. [Електронний ресурс]. – Режим доступу: http://victoria.lviv.ua/html/neural_nets/.
30. М.А. Новотарський Штучнінейронні мережі : обчислення. [Текст]./ Б.Б. Нестеренко // Інститут математики НАН України 2004 – 407 с.
31. Штучний нейрон. [Електронний ресурс]. – Режим доступу: http://victoria.lviv.ua/html/oio-l/help/neuron.pdf
32. Классификация нейронных сетей. [Електронний ресурс]. – Режим доступу: http://www.aiportal.ru/articles/neural-networks/classification.html
33. Класифікація відомих нейронних мереж по основних категоріях застосування [Електронний ресурс]. – Режим доступу: http://www.victoria.lviv.ua/html/oio/html/theme7.htm

34. Нейромережа зворотного поширення похибки [Електронний ресурс]. – Режим доступу: http://www.victoria.lviv.ua/html/neural_nets/Lecture3.htm
35. Yoon Kim Convolutional Neural Networks for Sentence Classification [Електронний ресурс]. – Режим доступу: http://emnlp2014.org/papers/pdf/EMNLP2014181.pdf
36. Гончаренко В. В. Фреймы для распознавания смысла текста / В. В. Гончаренко, Е. А. Шингарева // – Кишинев: Штиинца, 1984. – 198 с.
37. Сирота С.В., Кривоніс Б.Ю. Методи автоматизованого реферування для пошуку ключової інформації в тексті // Прикладна математика та комп'ютинг. ПМК, 2015 : сьома наук. конф. магістрантів та аспірантів, Київ, 15—17 квіт. 2015 р. : зб. пез доп. / [редкол.: Дичка І. А. та ін.]. — К. : Просвіта, 2015.