УДК 004.912

О.В. БАРМАК, О.В. МАЗУРЕЦЬ, А.В. ЖИВІЛІК
Хмельницький національний університет

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ АВТОМАТИЗОВАНОГО АНОТУВАННЯ ТА РЕФЕРУВАННЯ ЦИФРОВИХ ТЕКСТІВ

В статті було досліджено проблему семантичного аналізу текстів, зокрема автоматизованого згортання текстових документів. Визначено актуальність досліджень в напрямку автоматизованого анотування та реферування цифрових текстів. Запропоновано інформаційну технологію використання моделі Sub-Verb-Sub для формування анотацій та рефератів текстів. Інформаційна технологія автоматизованого анотування та реферування цифрових текстів призначена для знаходження ключових термінів у введеному тексті, створення на базі ключових термінів конструкції Sub-Verb-Sub у тексті, автоматизованого формування анотації та реферату тексту. Розглянуто тестовий програмний продукт, розроблений на засадах запропонованої інформаційної технології. Досліджено практичну ефективність розробленої інформаційної технології. Встановлено, що інформаційна технологія автоматизованого анотування та реферування може бути використана для ефективного автоматизованого реферування текстів.

Ключові слова: семантичний аналіз текстів, згортання текстів, дисперсійна оцінка, ключові терміни, анотація, реферат.

O.BARMAK, O.MAZURETS, A.ZHYVILIK
Khmelnytsky National University

INFORMATION TECHNOLOGY FOR AUTOMATIC CREATION OF ANNOTATIONS AND ABSTRACTS FROM DIGITAL TEXTS

The problem of semantic analysis of texts, in particular automation of compression of text documents, was researched in the article. The relevance of research in the direction of the automatic creation of annotations and abstracts from digital texts is determined. The information technology of using Sub-Verb-Sub model for creation of annotations and abstracts was suggested. Information technology for automatic creation of annotations and abstracts from digital texts is intended to find key terms in the text entered, to create of Sub-Verb-Sub constructs on the basis of key terms in the text, to create annotations and abstracts of the text atomically. The test program, which was developed based on suggested information technology, was estimated. The practical efficiency of the developed information technology was investigated. It has been established that information technology of automated annotation and abstract can be used for effective automated text referencing, but for

Keywords: semantic analysis of texts, curtailment of texts, dispersion evaluation, key terms, annotation, abstract.

Постановка проблеми в загальному вигляді

Використання комп'ютерів в людській діяльності, у тому числі і науковій, не тільки прискорює процеси створення та обробки документів, а й надзвичайно збільшує їх кількість і об'єм. Багато користувачів регулярно стикаються з необхідністю швидкого перегляду великого обсягу документів і вибору з них найбільш релевантних і дійсно потрібних документів. Виходом із ситуації є перегляд не всього документа, а його стислого опису – анотації або реферату [1]. Це зумовило необхідність проведення досліджень у вирішенні проблеми автоматичного реферування повнотекстових документів.

Одним із найважливіших напрямів у даних дослідженнях, є пошук шляхів і методів автоматичного стиснення (обсягового згортання) тексту. Під стисненням мається на увазі сукупність операцій аналітико-синтетичної переробки інформації, що переслідують мету створення вторинних документів чи вираження змісту вихідного тексту в більш економічній формі при максимальному збереженні його інформативності в похідному тексті. Реферування й анотування займають центральне місце у згортанні інформації, і всі проблеми, пов’язані з іншими різновидами згортання, так чи інакше відбиті в цих процесах.

При цьому, різниця між рефератом та анотацією полягає в наступному: реферат передає фактографічну інформацію і відповідає на питання, яку інформацію закладено в первинному документі; анотація ж являє собою стислу описову характеристику першоджерела і відповідає на запитання, про що говориться в первинному документі [2]. Крім того, в анотації основний зміст передається «своїми словами», які припускають високий ступінь абстрагування та узагальнення матеріалу. У рефераті ж використовуються ключові фрагменти, тобто формулюються узагальнення, запозичені з тексту оригіналу.

Реферування та анотування документів відносяться до числа основних видів інформаційної діяльності людини в ряду традиційних пошукових технологій [3]. Отриманий в результаті аналітичний огляд являє собою унікальний інформаційний продукт, здатний надати користувачеві повну і концентровану інформацію. Формування рефератів і анотацій вручну вимагає колосальних людських ресурсів, тому завдання створення ефективних методів автоматичного реферування та анотування набуває все більшої важливості.

Аналіз останніх досліджень

Автоматизоване вилучення знань з тексту є однією з основних задач штучного інтелекту і безпосередньо пов’язане з розумінням текстів на природній мові [4]. Задачу автоматизованої аналітичної обробки текстової інформації намагаються вирішити багато іноземних та вітчизняних вчених, серед яких можна виділити роботи В.І. Горькової, Є.А. Борохова, Х.П. Луна, В.Є. Берзона, І.П. Севбо, В.П. Леонова,

С.І. Гінді та інших.

Більшість існуючих програмних продуктів в розглядуваній області призначені переважно для статистичної обробки текстів. Наприклад, аналізатор від біржі контенту “Адвего” [6] вираховує кількість слів, кількість граматичних помилок і т.п. Сервіс також дозволяє побачити перелік ключових слів, які вираховуються за методом частотної оцінки (рис. 1). Текстовий аналізатор від компанії “Seozor” [7] допомагає визначити вагу слів в тексті (рис. 2) для с складання анкор-листа. За допомогою даного аналізатора тексту можна проаналізувати тексти ТОП-ової десятки сайтів, і визначити середню кількість входження до текстів ключових слів і символів. На основі частотного SEO аналізу тексту, сервіс самостійно визначає ключові слова в тексті і вагу кожного ключового слова.

Рис. 1. Обробка тексту в системі “Адвего”

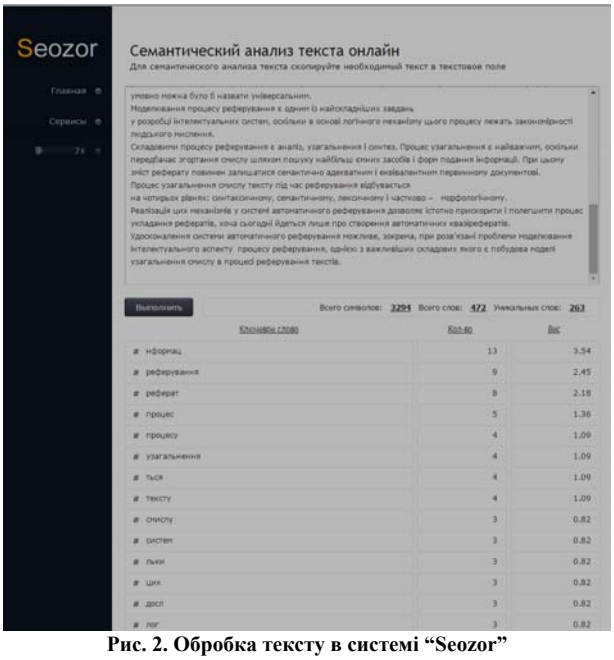

Рис. 2. Обробка тексту в системі “Seozor”

В Європі та США протягом останніх десятиліть проводяться активні теоретичні й практичні дослідження, спрямовані на пошуки ефективних методів автоматичного реферування. Незважаючи на початок активного вивчення альтернативних щодо екстрагування методів реферування, більшість алгоритмів сьогодні ґрунтується на екстрагуванні речень з оригінального тексту для побудови тексту реферату. Зараз активно проводяться машинні експерименти з оцінки існуючих систем автоматичного реферування, зі створення нових систем, що використовують Інтернет для одержання масивів оригінальних текстів і текстів рефератів для дослідження. В окрему проблему виділяються питання індикативного реферування і теоретичні питання ролі реферату в документальному пошуку.

Дослідження і розробки в галузі автоматичної обробки тексту (АОТ) в Європі і США привертають увагу найбільших приватних фірм і державних організацій найвищого рівня. Європейський Союз вже декілька років координує ряд програм у галузі автоматичної обробки тексту. Наприклад, Human Language Technology Sector of the Information Society Technologies (IST) Programme [8]. Основні розробки присвячено автоматизації процесу синтаксичного аналізу для різних систем АОІ, в тому числі й АР.

На відміну від лексико-граматичного аналізу тексту, синтаксичний аналіз – галузь прикладної лінгвістики, що перебуває в стані розвитку. Мета синтаксичного аналізу – автоматична побудова функціонального дерева фрази, тобто пошук взаємозалежності між різнорівневими елементами речення.

Так, синтаксичний аналізатор Ergo Linguistic Technologies Parser [9], розроблений Дереком Бікертоном і Філіпом Браліком з Університету Гонолулу, використовує широковідому схему аналізу і має наочне вираження. ERGO орієнтує свій парсер на використання інтерфейсів у вигляді питань і відповідей. ERGO поки що є єдиною компанією, яка має парсер, здатний визначати тип запитання (питання до підмета, суб’єкта, прямого або непрямого додатка чи обставини) і «миттєво» конструювати відповідь.

Один із найбільш вдалих синтаксичних аналізаторів Functional Dependency Grammar [10] створений дослідниками з Гельсінського університету, котрі пізніше заснували дві фірми: Lingsoft і Conexor. Рання версія під назвою ENGCG (English Constraint Grammar) була використана для анотації найбільшого у світі корпусу – Bank of English, що належить видавництву Collins/Harper Publishers. Особливістю даного синтаксичного аналізатора є те, що у випадках, коли неможливо зняти багатозначність, синтаксичний аналізатор або видає декілька варіантів аналізу, або не добудовує дерево для даної частини фрагменту.

Один із найбільш оригінальних підходів до синтаксичного аналізу тексту – Link Parser – розроблено в Carnegie-Melon University. Цей синтаксичний аналізатор – єдиний, чиї початкові коди були опубліковані он-лайн [11]. Тоді як більшість систем синтаксичного аналізу використовують структури рівня іменних і дієслівних груп у побудові дерева фрази, Link Grammar, яка покладена в основу Link Parser, використовує

інформацію про типи зв'язків, які кожне слово може мати зі словами, що знаходяться праворуч або ліворуч, а також декілька загальних граматичних правил.

На ринку існує зовсім невелика кількість традиційних програм реферування, тобто таких, які виділяють найбільш вагомі фрагменти з тексту, використовуючи статистичні алгоритми або слова-підказки. Inxight Summarizer – одна з найбільш відомих комерційно поширюваних систем реферування. Inxight Summarizer був створений у Дослідницькому центрі Ксерокса в Пасло Альто [12].

Серед комерційних систем також можна відзначити Prosum [13] – систему реферування, розроблену British Telecommunications Laboratories у межах експериментальної комерційної онлайн-платформи TranSend, що являє собою cgi-скрипт, вбудований до веб-сторінки.

Оскільки інтерес до традиційних систем автоматичного реферування неухильно знижується, багато компаній пропонують інші підходи. Одним із нетрадиційних рішень є використання іменних груп, виділених за допомогою часткових синтаксичних аналізаторів. Алгоритми такого типу використовуються в програмному продукті Extractor, що створений в Інституті інформаційних технологій Національної дослідницької ради Канади. Він являє собою модуль, що виділяє з наданого йому на вхід тексту найбільш інформативні іменні групи. За замовчанням кількість таких груп – сім, незалежно від довжини тексту. Extractor використовується в програмних продуктах фірм ThinkTank Technologies і Tetranet, а також у пошуковій системі Журналу досліджень в галузі штучного інтелекту.

Система автоматичного реферування, інтегрована в текстовий редактор Microsoft Word, працює на основі методу екстрагування. Ця система далека від досконалості, однак виробляє більш-менш вдалі квазірефераты.

Отже, останнім часом над завданнями синтаксичного аналізу речення та розробки методів автоматичного згортання текстів працює багато дослідницьких груп, і на даний момент цей напрям залишається актуальним. Якщо семантичний вміст тексту може бути переданий множиною ключових термінів [14], то для побудови реферату та анотації необхідне зв'язування термінів в рамках закладеної у текст семантичної моделі, що може бути здійснено, зокрема, шляхом використання відомої моделі «суб'єкт-об'єкт-дія» (СОД, англ.: Sub-Verb-Sub) [15].

Постановка задачі

Метою даної роботи є розробка інформаційної технології використання моделі «Sub-Verb-Sub» для формування згорнутих зразків текстових документів.

Викладення основних матеріалів дослідження

Вирішення задачі автоматизованого формування згорнутих зразків текстових документів з використанням моделі СОД складається з ряду етапів перетворення інформації. Вхідними даними є цифровий текст або його визначена частина; вихідними даними є текст анотації чи реферату, релевантний вхідному тексту.

На рисунку 3 надано функціональну діаграму (за стандартом IDEF0 [16]), яка ілюструє послідовність дій при формуванні згорнутих зразків текстових документів шляхом використання моделі Sub-Verb-Sub.

Рис. 3. Діаграма етапів у формуванні згорнутих зразків текстових документів

Кожному етапу з наведених відповідає ряд послідовностей перетворення даних, які загалом формують інформаційну технологію використання моделі Sub-Verb-Sub для формування згорнутих зразків текстових документів, наведену на рисунку 4.

Попередня технічна обробка тексту (Блок 1) полягає в обробці розділових знаків й усуненні неоднозначного іменування термінів. Неоднозначне іменування термінів свідчить про недотримання норм формування та ведення наукової літератури. Хоча є допустимим використання абревіатур, напівскорочень та повних назв (наприклад, «СКРБД», «Система керування РБД», «Система керування розподіленими базами даних»), зокрема кількома мовами (наприклад, українською та англійською), для використання в рамках окремого матеріалу обирається лише один варіант, оскільки як для машинної, так і для антропосемантичної ідентифікації термінів бажаною (а в науковій літературі – обов'язковою) є символьна уніфікація ідентифікатора. Навіть у випадку введення нового скорочення, в цій же позиції у тексті присутня також і повна назва. Отже, усунення неоднозначної іменованості термінів переслідує мету уникнути випадків, коли при автоматичному аналізі контенту навчальних матеріалів одне поняття буде розглядатись як кілька окремих термінів.

При обробці розділових знаків проводиться їх видалення й стандартизація. Власне розділові знаки видаляються, а якщо вони є частиною слів – уніфікуються. Цей етап є суто технічним і покликаний зменшити кількість «сміття» у контенті, яке заважає автоматизованому аналізу.

Формування множини термінів (Блок 2) проводиться за кілька етапів. Спершу виконується пошук важливих слів з використанням методу дисперсійного оцінювання, який показав свою достатню ефективність у рамках попередніх досліджень [17].

Рис. 4. Схема інформаційної технології автоматизованого формування згорнутих зразків текстових документів

Дисперсійна оцінка є оцінкою дискримінантної сили слів й дозволяє відкинути із загальної множини широковживаних у тексті слів слова, що розташовані рівномірно, із достатньою ефективністю [18]. Для формування вихідної множини слів слова у множині, сформованої у результаті дисперсійного

оцінювання, сортуються за зменшенням значення дисперсійної оцінки. Для якісного пошуку ключових слів використовується не вся множина, а визначена частина її елементів із дисперсійною оцінкою вище порогової величини.

Словосполучення ж є стійкими сукупностями важливих слів, що згруповані у визначеній послідовності та у такій комбінації неодноразово присутні в розглядуваному контенті. Для формування множини словосполучень проводиться пошук неперервних скупчень ключових слів протягом тексту й знайдені зразки додаються у масив словосполучень. Отриманий масив словосполучень сортується за частотою вживання, після чого з нього видаляються неключові словосполучення – такі, що зустрічаються в тексті один раз, або не містять двох і більше іменників чи прикметників.

На основі отриманих даних проводиться формування вихідної множини термінів [19]. Інтеграція множини ключових слів та множини ключових словосполучень відбувається шляхом заміщення словосполученнями слів, які є переважно елементами відповідних словосполучень. Результатом етапу є об’єднана вихідна множина термінів. Об’єднана множина термінів компактифікується шляхом видалення повторів серед її елементів і сортується за значеннями їх дисперсійної оцінки.

Пошук речень з термінами у моделі Sub-Verb-Sub (Блок 3) ставить за мету знайти важливі речення в тексті й визначити їх семантичний вміст шляхом їх співставлення з множиною термінів через ідентифікацію моделей Sub-Verb-Sub. Конструкція моделі Sub-Verb-Sub (рис. 5) включає складові:

- іменник, що виступає суб’єктом – підметом зі значенням виробника дії або носія стану;
- дієслово, що виражає дію – діяльнісне співвідношення між суб’єктом і об’єктом;
- іменник, що виступає об’єктом – додатком, на який спрямовано практичну або пізнавальну діяльність суб'єкта.

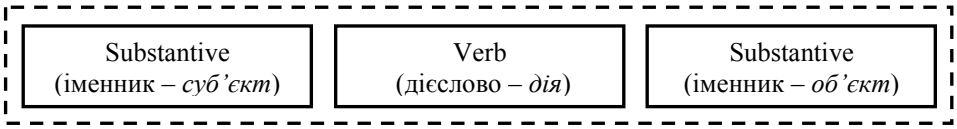

Рис. 5. Схема конструкції моделі Sub-Verb-Sub

Модель Sub-Verb-Sub дозволяє фіксувати природні елементи формування зв’язних текстів, коли введення чи переміщення фокусу до нових термінів передбачає наступне їх використання в подальших семантичних побудовах. При цьому в межах речення відбувається семантичне об'єднання суб'єкта, про який що-небудь стверджується чи заперечується, та об'єкта, на який спрямовано практичну або пізнавальну діяльність суб'єкта в рамках єдиної семантичної моделі. При цьому об'єктом може бути і власне суб'єкт. Ці одиниці утворюють основу семантичної структури речення. Центральними категоріями семантичної структури речення є:

1) суб'єкт – виробник дії або носій стану;

2) дія – предикативна ознака, що виявляється в об'єктивно-модальному плані: у часі й у певному відношенні до дійсності, й ця категорія в подальшому реалізується як дія;

3) об'єкт – те, на що спрямована дія або до якого звернений стан.

Категорія предикативної ознаки (дії або стану), тобто ознаки, віднесеної в певний часовий план, завжди присутня в компоненті структурної схеми речення: в одному з головних членів двукомпонентного речення або в головному члені однокомпонентного речення. Всередині предикативної ознаки основне протиставлення здійснюється на підставі його віднесеності або невіднесеності до суб’єкта. Категорія суб’єкта є категорією носія предикативної ознаки: виробника дії або носія стану. Категорія об’єкта присутня у тих реченнях, в яких ознака з’являється як конкретна дія, діяльність, розумова, емоційна або вольова активність чи як чиєсь відношення до чогось. А об’єкт – це те, на що безпосередньо спрямовано дію, діяльність, до чого безпосередньо звернене чиєсь відношення.

Початковим етапом пошуку речень з термінами у моделі Sub-Verb-Sub є формування множини речень з дієслівним роздільником. У отриманій множині вилучаються речення, в яких відсутні дві ключові іменникові групи – суб’єкта та об’єкта. Співставлення речень з результуючої множини множині термінів дозволяє визначити оцінку речень на ступінь відповідності моделі Sub-Verb-Sub шляхом врахування значень дисперсійної оцінки використаних у реченнях термінів.

Формування реферату (Блок 4) є процесом екстрагування з тексту послідовно семантично пов’язаних речень. В рамках розглядуваної моделі цей процес передбачає визначення вузлових елементів онтології та його обсягове згортання (рис. 6). Вузлові елементи онтології представлені множиною ключових термінів, множина конструкцій Sub-Verb-Sub виражає семантичний базис тексту, а множина актуальних речень з термінами у моделі Sub-Verb-Sub передає семантичний вміст тексту.

Шляхом запропонованого екстрагування з тексту послідовно семантично повязаних речень проводиться побудова всіх можливих варіантів семантичних ланцюгів. Етап оцінки варіантів семантичних ланцюгів враховує: сумарну оцінку речень на ступінь відповідності моделі Sub-Verb-Sub; довжину семантичного ланцюга; ступінь охоплення онтології – що ставить за мету зменшити імовірність включення до результуючого семантичного ланцюга менш важливих термінів при відсутності в ньому більш важливих термінів. У результаті порівняння оцінок варіантів семантичних ланцюгів забезпечується вибір

результуючого набору речень, що буде екстрагований з вхідного тексту в якості реферату.

Формування анотації (Блок 5) відбувається шляхом побудови послідовності семантичних конструкцій через відтворення їх сенсу на основі моделі Sub-Verb-Sub. При цьому в якості каркасу використовується один з семантичних ланцюгів пов'язаних речень, які сформовані на попередньому етапі.

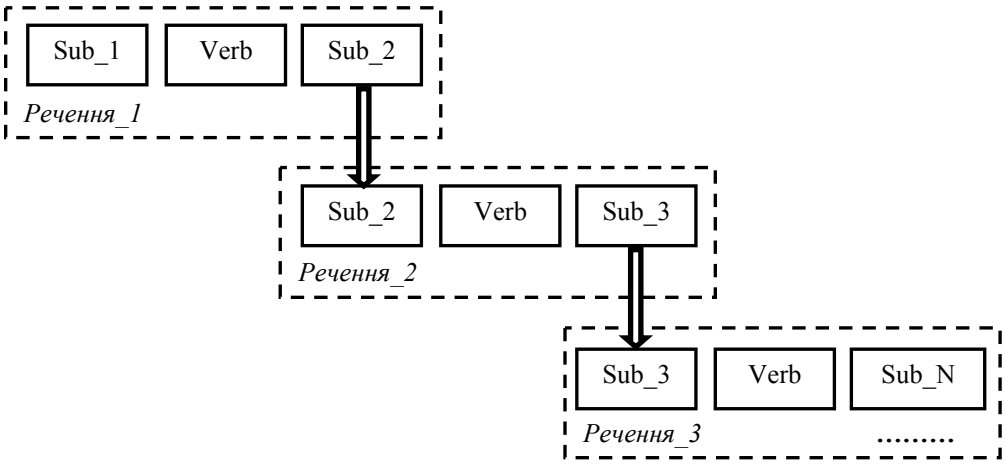

Рис. 6. Схема методу послідовного зв'язування речень за моделлю Sub-Verb-Sub

З точки зору цифрового вмісту, поверхнева структура речення ототожнюється з його формальною синтаксичною організацією, а глибинна структура – із семантичною. При цьому синтаксична структура є формальним виразником семантичної структури, що відбиває ідентичні або тотожні ситуації, які розглядаються в реченні. Тому поверхнева структура речення є основою для аналізу структури речення, адже саме вона доступна для цифрової обробки.

Типології речень з розширеними структурними схемами наразі практично не розроблено. Втім основою для розширеної структури слугує мінімальна структурна схема, що залежно від її граматичного оформлення, лексичного наповнення й ступеня адекватності описуваної ситуації визначає лексичні й граматичні параметри розширеної моделі речення, що є похідною від простої моделі. Тому можна припустити, що шляхом використання простої моделі речень також можливо достатньо повно виразити певний зміст поза контекстом.

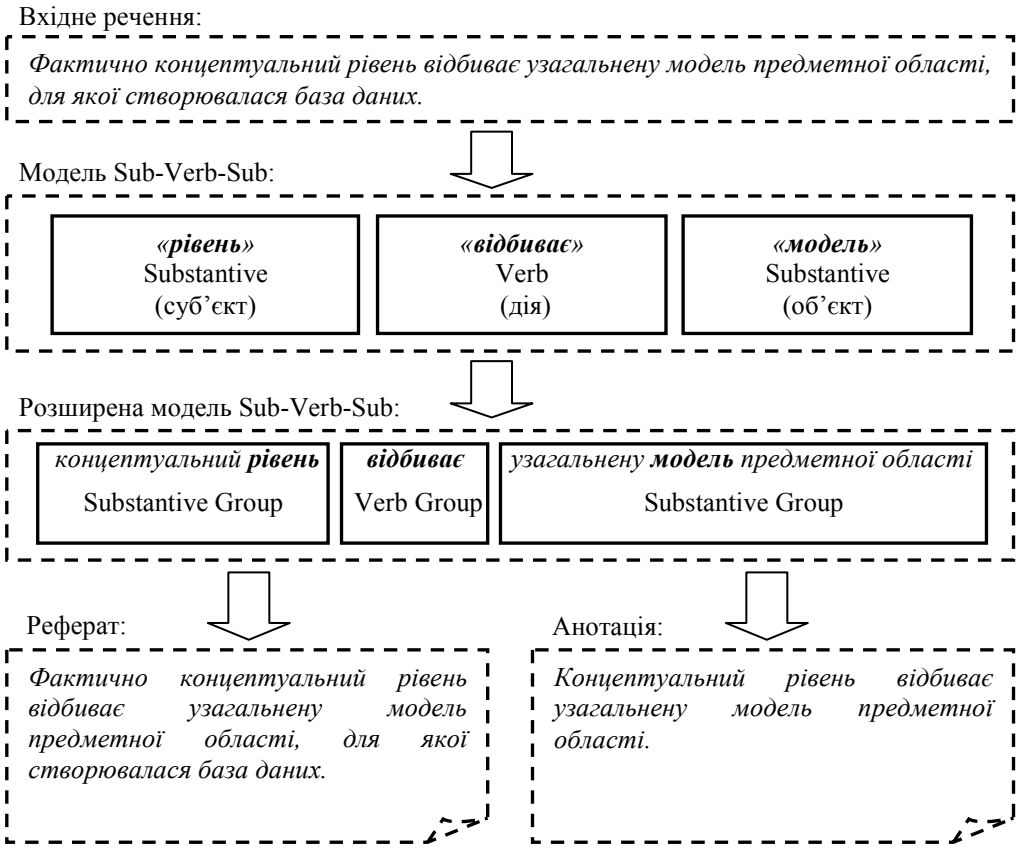

Рис. 7. Еволюція конструкції речення за моделлю Sub-Verb-Sub

Використовуючи множину ключових слів як перелік прототипів Sub, беруться до уваги речення, що містять конструкції Sub-Verb-Sub, причому в рамках речень обов'язкова наявність двох елементів з множини Sub, розділених дієсловом з множини Verb. При цьому наявність всередині фрагменту речення, що охоплюється конструкцією Sub-Verb-Sub, інших слів, які не входять до зазначених множин, ігнорується як другорядна інформація. При цьому для семантичної характеристики речення актуальним є поняття позиції. Під позицією розуміється та ланка в структурі речення, яку займає певний термін з множини Sub. Таким чином, для повноти передавання сенсу необхідне використання не тільки власне конструкції Sub-Verb-Sub, що складається з різних словесних форм, а й її розширеної версії, у якій кожен з елементів являє собою структуру, до якої входять терміни із зазначених переліків в поєднанні з уточнюючими додатковими членами речення (рис. 7). Зокрема, до іменникової групи можуть входити інші іменники (не в називному відмінку) та прикметники (без врахування малих частин речення на кшталт сполучників), а дієслівна група може включати декілька дієслів.

Таким чином, шляхом прив'язки до кожного з елементів моделі множини доповнюючих другорядних слів відбувається ідентифікація іменникових і дієслівних груп, й для формування анотації проводиться генерація речень шляхом відтворення семантичного вмісту речень наборами іменникових і дієслівних груп. У якості текстового каркасу використовується один з семантичних ланцюгів пов'язаних речень, які сформовані на етапі формування й оцінки варіантів семантичних ланцюгів.

Програмна реалізація

З метою перевірки ефективності запропонованої інформаційної технології, на її засадах було створено тестовий програмний продукт, який дозволяє за введеним текстом виконати наступні функції:

- попередня обробка введеного тексту
- автоматизоване визначення ключових термінів у введеному тексті;
- формування на базі ключових термінів конструкцій Sub-Verb-Sub;
- витяг статистичних відомостей по простих семантичних конструкціях Sub-Verb-Sub у тексті;
- автоматизоване формування анотації тексту;
- автоматизоване формування реферату тексту.

Автоматизована система формування згорнутих зразків текстових документів за моделлю Sub-Verb-Sub реалізована з використанням мови програмування C# на платформі .NET Framework. Система використовує базу даних (рис. 8) на СКБД Microsoft SQL Server, призначену для зберігання даних про слова української мови, їх леми та назви словоформ.

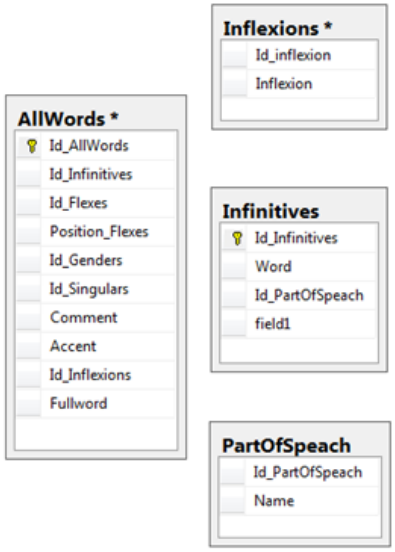

Рис. 8. Схема БД системи

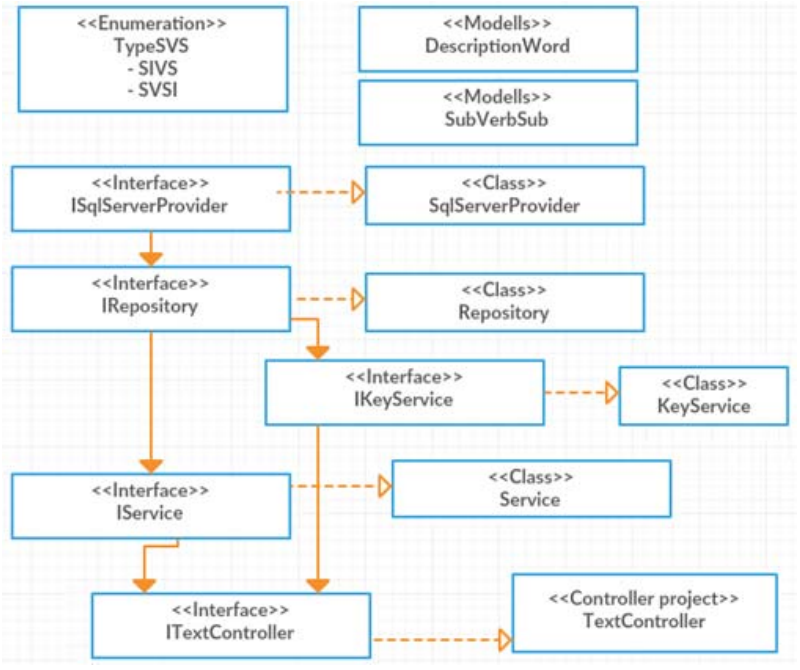

Рис. 9. Діаграма класів системи

Таблиця «AllWords» містить в собі перелік слів української мови в різних словоформах, використовується для визначення правильності написання слова і для подальшого визначення його властивостей. Таблиця «Infinitives» зберігає леми – інфінітивні слова до тих, які зберігаються в таблиці «AllWords». Таблиця «PartOfSpeach» зберігає в собі назви частин мови і використовується, зокрема, для подальшого формування іменникових та дієслівних груп слів. Таблиця «Inflexion» містить в собі дані про відмінок слова чи інший параметр його модифікації.

Діаграму класів автоматизованої системи формування згорнутих зразків текстових документів подано на рисунку 9. Система побудована на основі 4 проектів і одного проекту для юзер-інтерфейсу (UI).

Так, проект “Models” описує моделі даних. Проект “Repository” працює з базою даних і займається вибіркою даних з БД, також вибірка даних відбувається в асинхронному режимі. Проект “Service” займається бізнес-логікою системи, використовує “Repository” для вибірки даних; на основі цих даних сервіс займається формуванням семантичних конструкцій, ключових слів, лематизацією, нормалізацією. Проект “Controller” виступає обгорткою і попереднім шаром перед UI, а також зберігає проміжні дані для роботи з сервісом; він працює із сервісом і не має доступу до репозитарію. Така архітектура дозволяє системі працювати максимально незалежно.

Клас “TextHelper” виконує нормалізацію тексту, а також методом дисперсійної оцінки формує множину ключових слів з їхніми вагами, які в подальшому використовуються в системі. Клас “Repository” відповідає за роботу з базою даних і дає запит до БД, яка в свою чергу повертає дані в програму, й використовуючи ключові терміни, сформовані класом “KeyService”, цей клас обробляє їх. Клас “Repository” також використовується для обробки вхідного тексту і формування семантичних конструкцій.

В рамках дослідження ефективності запропонованої інформаційної технології, у відповідний елемент тестового програмного продукту вводився текст (рис. 10), який після приведення до стилістичної та термінологічної однорідності, видалення й стандартизації розділових знаків зчитувався й оброблявся системою. В результаті в UI автоматично формуються дві таблиці: перша містить в собі ключові терміни, а друга – їх відповідні лематизовані аналоги (рис. 11).

На наступному етапі проводиться аналіз тексту з метою сформувати множину наявних елементів моделі Sub-Verb-Sub. Шляхом співставлення наявних елементів моделі Sub-Verb-Sub з термінами у моделі Sub-Verb-Sub формується множина речень з дієслівним роздільником. Для дослідницьких цілей система дозволяє проводити аналіз як речень на предмет наявності в них елементів моделі Sub-Verb-Sub (рис. 12), так і елементів моделі Sub-Verb-Sub на предмет пошуку пов’язаних з ними речень тексту й формування статистики (рис. 13).

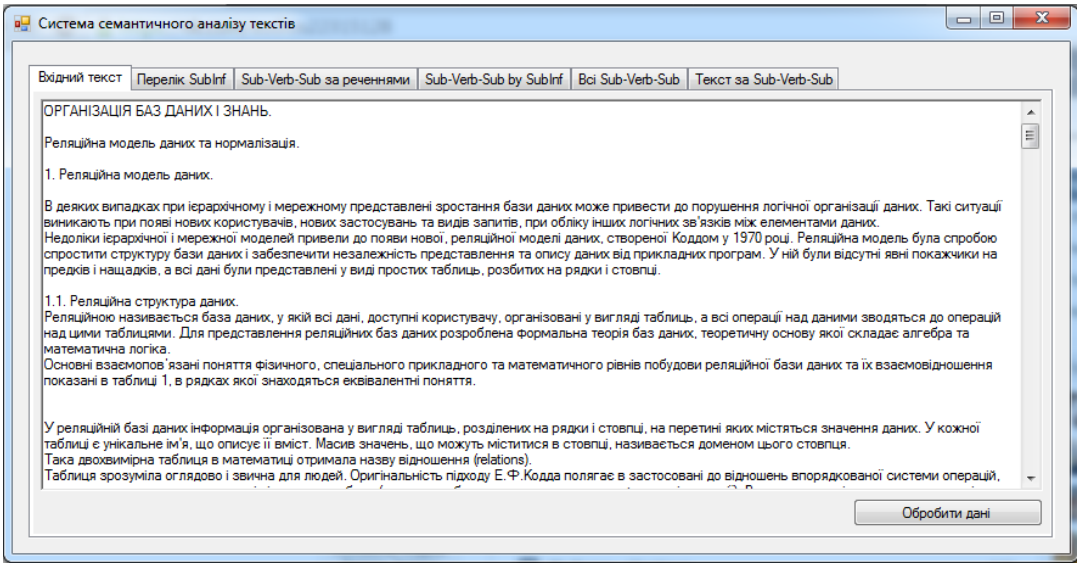

Рис. 10. Введення тестового тексту в систему

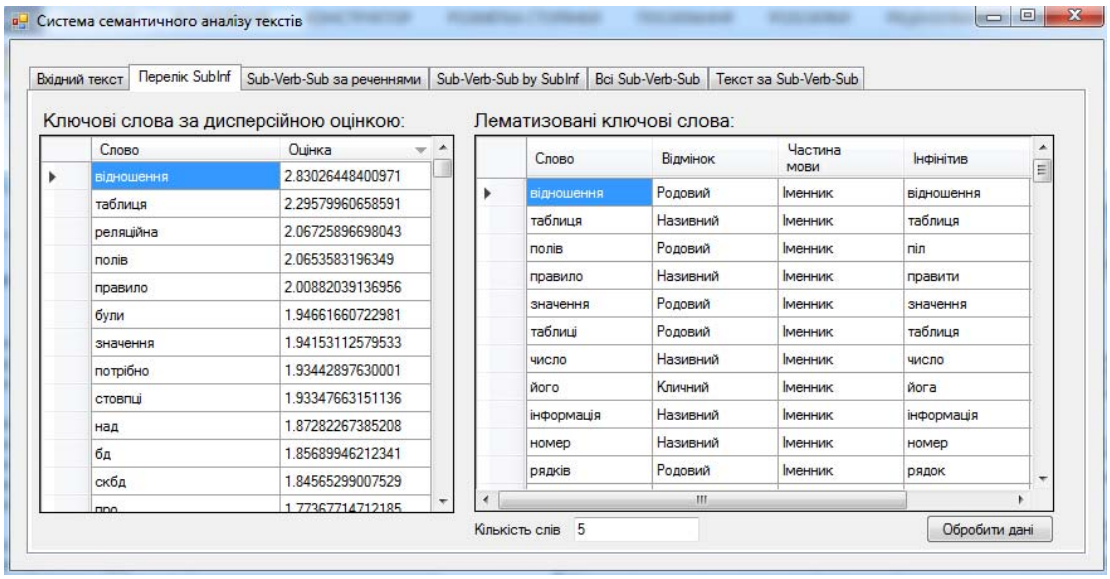

Рис. 11. Формування системою множини ключових термінів та їх аналіз

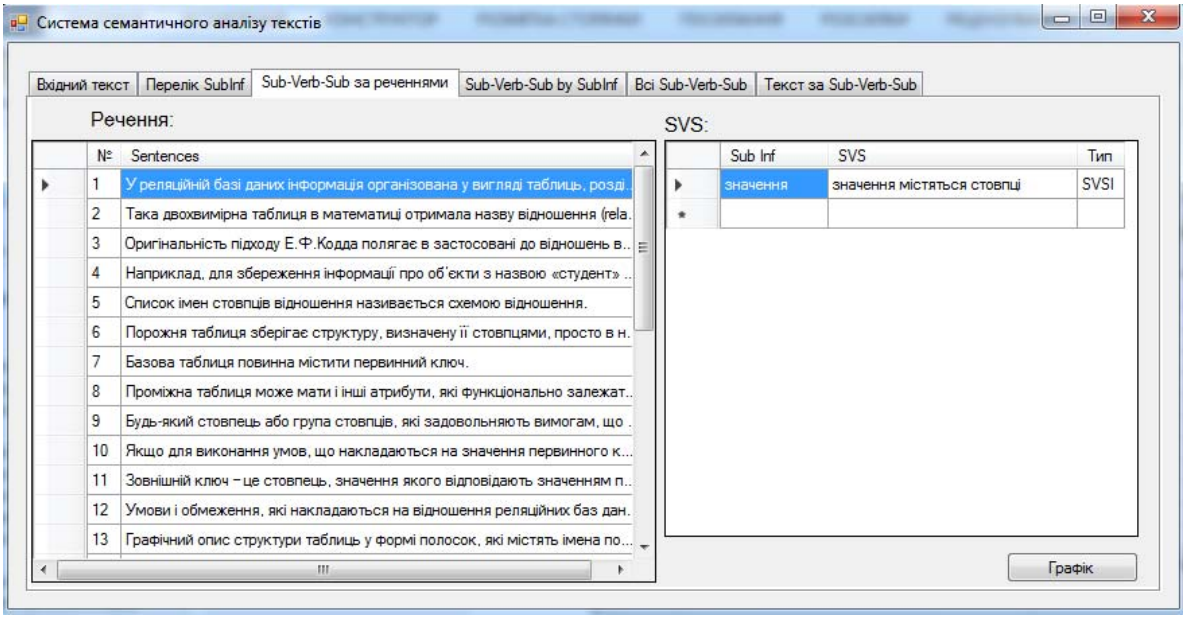

Рис. 12. Аналіз тексту на відповідність елементів моделі Sub-Verb-Sub реченням

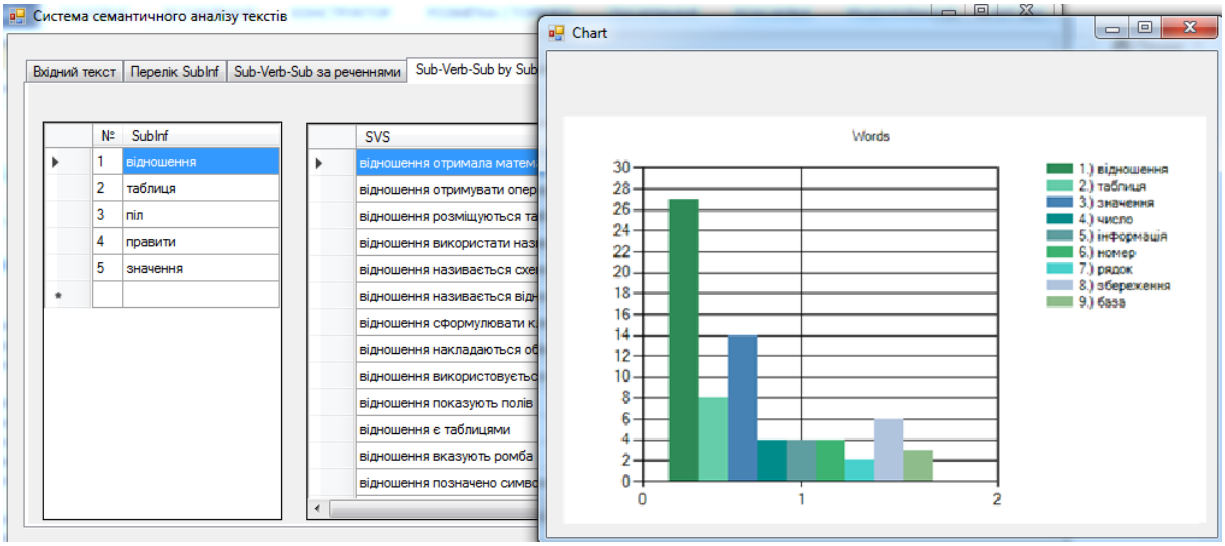

Рис. 13. Аналіз тексту на відповідність речень елементам моделі Sub-Verb-Sub

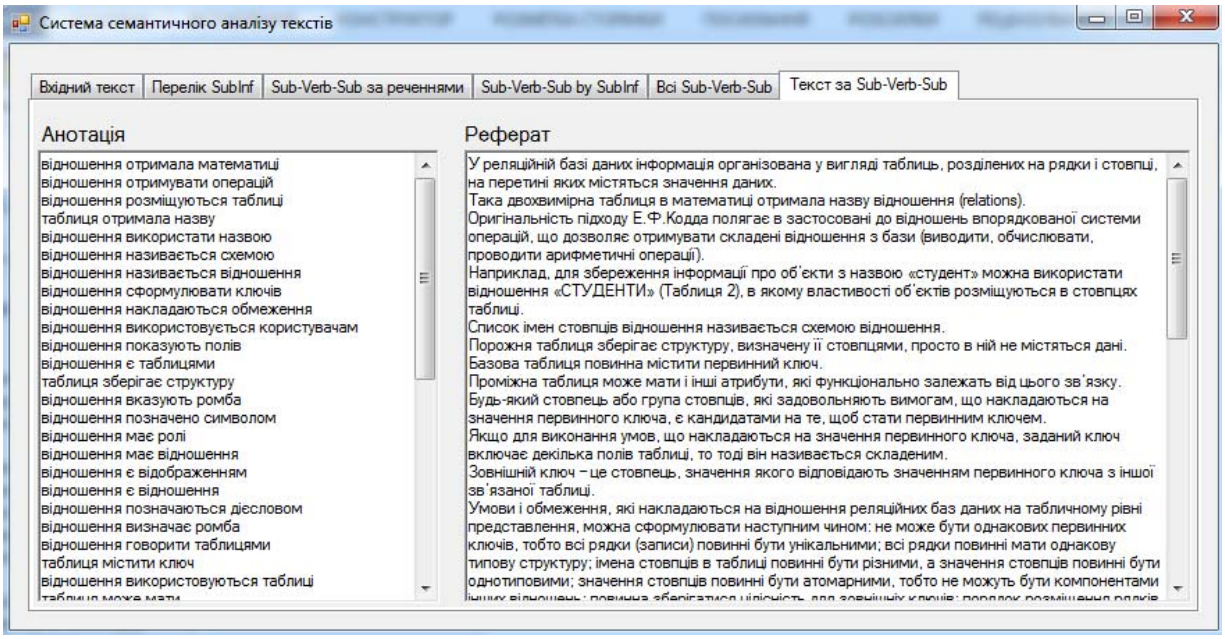

Рис. 14. Формування згорнутих текстів шляхом послідовного зв'язування речень моделі Sub-Verb-Sub

В результаті будуються всі можливі варіанти семантичних ланцюгів з моделями Sub-Verb-Sub, оцінюються, й вибирається найбільш прийнятний варіант. Саме на його основі шляхом екстрагування з тексту послідовно семантично пов'язаних речень проводиться побудова реферату. На основі цього ж семантичного ланцюгу формується анотація. Побудова анотації відбувається шляхом побудови послідовності семантичних конструкцій шляхом відтворення їх сенсу на основі моделі Sub-Verb-Sub (рис. 14).

Розглянутим чином, на основі введеного у систему тексту відбувається формування відповідних реферату та анотації як зразків згорнутого тексту.

Результати дослідження

Дослідження практичної ефективності розробленої інформаційної технології використання моделі Sub-Verb-Sub для формування згорнутих зразків текстових документів переслідує перевірку ефективності формування анотацій та рефератів як кінцевої мети створення інформаційної технології. Для перевірки якості формування анотацій було досліджено ефект від використання розробленої системи за критерієм оцінки за введеним текстом (37 речень) анотацій, сформованих за опціями: 3 речення, 5 речень, 10 речень. Кожне речення анотацій містить одну конструкцію моделі Sub-Verb-Sub. Оцінка проводилась дванадцятьма піддослідними, що оцінювали якість анотації по десятибальній шкалі. В результаті було отримано статистичні дані, середні значення яких показано на рисунку 15.

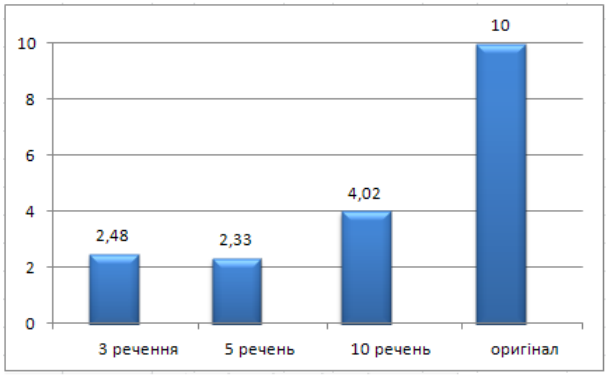

Рис. 15. Середня оцінка якості автоматично сформованої анотації (шкала – 10 балів)

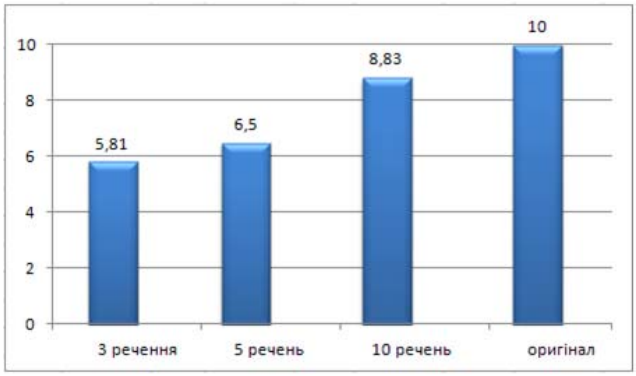

Рис. 16. Середня оцінка якості автоматично сформованого реферату (шкала – 10 балів)

Для перевірки якості формування реферату було досліджено ефект від використання розробленої системи за критерієм оцінки за введеним текстом (37 речень, текст той самий) рефератів, сформованих за аналогічними опціями: 3 речення, 5 речень, 10 речень. Кожне речення реферату може містити кілька конструкцій моделі Sub-Verb-Sub, хоча використовується внаслідок наявності лише однієї з них. Оцінка проводилась також дванадцятьма тими ж піддослідними, що оцінювали якість реферату по десятибальній шкалі. На рисунку 16 відображено середні значення отриманих статистичних даних.

Дискусія

Дослідження практичної ефективності розробленої інформаційної технології використання моделі Sub-Verb-Sub для формування згорнутих зразків текстових документів показало низьку ефективність застосування моделі Sub-Verb-Sub для автоматизованого анотування текстів. При цьому, дослідження показало достатньо високу ефективність застосування моделі Sub-Verb-Sub для автоматизованого реферування текстів. У зв'язку з чим розроблена інформаційна технологія може бути використана для автоматизованого реферування текстів, хоча й існують перспективи її подальшого удосконалення. Однак розроблена інформаційна технологія без додаткових елементів удосконалення не може бути використана для якісного автоматизованого анотування текстів внаслідок виявленої низької фактичної ефективності.

Висновки

В статті було досліджено проблему семантичного аналізу текстів, а саме автоматизованого згортання текстових документів шляхом автоматизації формування анотацій та рефератів. Зокрема, було проведено аналіз ефективності застосування моделі Sub-Verb-Sub для автоматизованого реферування та анотування текстів.

Запропоновано інформаційну технологію використання моделі Sub-Verb-Sub для формування згорнутих зразків текстових документів, яка дозволяє за введеним текстом автоматизовано визначати ключові терміни у введеному тексті, формувати на базі ключових термінів конструкції Sub-Verb-Sub у тексті, автоматизовано формувати анотації та реферати тексту.

Розглянуто тестовий програмний продукт, розроблений на засадах запропонованої інформаційної технології й призначений для дослідження ефективності розробленої інформаційної технології. Дослідження практичної ефективності розробленої інформаційної технології визначення семантичного вмісту тексту показало низьку ефективність застосування моделі Sub-Verb-Sub для автоматизованого анотування текстів, проте достатньо високу ефективність застосування моделі Sub-Verb-Sub для автоматизованого реферування

текстів. У зв'язку з чим розроблена інформаційна технологія може бути використана для автоматизованого реферування текстів.

Подальші дослідження спрямовані на вдосконалення методів та алгоритмів, що забезпечують коректне автоматизоване визначення й локалізацію іменникових груп у реченнях, а також на поширення можливостей технології на ефективну роботу із складними реченнями та семантичними конструкціями. В результаті планується досягти суттєвого підвищення якості автоматизованого анотування текстів з використанням розробленої інформаційної технології використання моделі Sub-Verb-Sub для формування згорнутих зразків текстових документів.

Література

1. Кліменко В.І. Аналіз сучасних методів автоматизації анотування та реферування текстів / В. І. Кліменко, А. В. Живілік, О. В. Мазурець // Збірник наукових праць за матеріалами дев'ятої міжнародної науково-технічної конференції «Актуальні проблеми комп'ютерних технологій 2015». – Хмельницький, 2015. – С. 116–123.

2. Коханова І. О. Проблеми та похибки методів автоматизованого реферування документів / І. О. Коханова // Вісник Книжкової палати. – 2014. – № 9. – С. 31–32.

3. Снитюк В. Е. Интеллектуальное управление оцениванием знаний / В. Е. Снитюк, К. Н. Юрченко. − Черкассы, 2013. − 262 с.

4. Даревич Р. Р. Підвищення ефективності інтелектуального аналізу тесту шляхом зважування понять моделі онтології / Р. Р. Даревич // Штучний інтелект. – 2005. – № 3. – С. 571–577.

5. Автоматизоване реферування – Вільна енциклопедія [Електронний ресурс]. – Режим доступу : http://uk.wikipedia.org/wiki/Автоматизоване реферування.

6. Семантичний аналіз тексту – seo-аналізатор «Адвего» [Електронний ресурс]. – Режим доступу : http://advego.ru/text/seo/.

7. Семантичний онлайн аналізатор тексту «Seozor» [Електронний ресурс]. – Режим доступу : http://seozor.ru/tools/analyzer.php.

8. Human Language Technology Sector of the Information Society Technologies (IST) Programme [Електронний ресурс]. – Режим доступу : http://www.linglink.lu.

9. ERGO Linguistic Tecnologies [Електронний ресурс]. – Режим доступу : http://www.ergo-ling.com.

10. Functional Dependency Grammar [Електронний ресурс]. – Режим доступу : http://www.conexor.fi.

11. Link Grammar Homepage [Електронний ресурс]. – Режим доступу : http://bobo.link.cs.cmu.edu/link.

12. Inxight Summarizer [Електронний ресурс]. – Режим доступу : http://www.inxight.com.

13. Prosum Summarizer [Електронний ресурс]. – Режим доступу : http://transend.labs.bt.com/cgi-bin/prosum/prosum.

14. Бармак О. В. Інформаційна технологія автоматизованого визначення термінів у навчальних матеріалах / О. В. Бармак, О. В. Мазурець // Міжнародний науково-технічний журнал «Вимірювальна та обчислювальна техніка в технологічних процесах» – Хмельницький, 2015. – № 2. – С. 94–102.

15. Про поняття моделі простого речення [Електронний ресурс]. – Режим доступу : https://www.jstor.org/stable/43659932.

16. IDEF5 – Ontology Description Capture Method [Електронний ресурс]. – 2015. – Режим доступу : http://www.idef.com/IDEF5.htm.

17. Бармак О. В. Методи автоматизації визначення семантичних термінів у навчальних матеріалах / О. В. Бармак, О. В. Мазурець // Вісник Хмельницького національного університету. Серія: Технічні науки. – Хмельницький, 2015. – № 2(223). – С. 209–213.

18. Ландэ Д. В. Компактифицированный горизонтальный граф видимости для сети слов / Д. В. Ландэ, А. А. Снарский // Труды Международной научной конференции «Интеллектуальный анализ информации ИАИ-2013. Знания и рассуждения». – Киев : КПИ,2013. – С. 158–164.

19. Крак Ю. В. Практичне дослідження ефективності інформаційної технології автоматизованого визначення семантичних термінів в контенті навчальних матеріалів / Ю. В. Крак, О. В. Бармак, О. В. Мазурець // Прикладне програмне забезпечення : збірник наукових праць за матеріалами десятої міжнародної науково-практичної конференції по програмуванню «УкрПРОГ'2016». – Київ, 2016. – С. 237–245.

References

1. Klimenko V.I. Analiz suchasnykh metodiv avtomatyzatsii anotuvannia ta referuvannia tekstiv / V. I. Klimenko, A. V. Zhyvilik, O. V. Mazurets // Zbirnyk naukovykh prats za materialamy deviatoi mizhnarodnoi naukovo-tekhnichnoi konferentsii «Aktualni problemy kompiuternykh tekhnolohii 2015». – Khmelnytskyi, 2015. – S. 116–123.

2. Kokhanova I. O. Problemy ta pokhybky metodiv avtomatyzovanoho referuvannia dokumentiv / I. O. Kokhanova // Visnyk Knyzhkovoi palaty. – 2014. – # 9. – S. 31–32.

3. Snytiuk V. E. Yntellektualnoe upravlenye otsenyvanyem znanyi / V. E. Snytiuk, K. N. Yurchenko. − Cherkassы, 2013. − 262 s.

4. Darevych R. R. Pidvyshchennia efektyvnosti intelektualnoho analizu testu shliakhom zvazhuvannia poniat modeli ontolohii / R. R. Darevych // Shtuchnyi intelekt. – 2005. – # 3. – S. 571–577.

5. Avtomatyzovane referuvannia – Vilna entsyklopediia [Elektronnyi resurs]. – Rezhym dostupu : http://uk.wikipedia.org/wiki/Avtomatyzovane_referuvannia.

6. Semantychnyi analiz tekstu – seo-analizator «Adveho» [Elektronnyi resurs]. – Rezhym dostupu : http://advego.ru/text/seo/.

7. Semantychnyi onlain analizator tekstu «Seozor» [Elektronnyi resurs]. – Rezhym dostupu : http://seozor.ru/tools/analyzer.php.

8. Human Language Technology Sector of the Information Society Technologies (IST) Programme [Elektronnyi resurs]. – Rezhym dostupu : http://www.linglink.lu.

9. ERGO Linguistic Tecnologies [Elektronnyi resurs]. – Rezhym dostupu : http://www.ergo-ling.com.

10. Functional Dependency Grammar [Elektronnyi resurs]. – Rezhym dostupu : http://www.conexor.fi.

11. Link Grammar Homepage [Elektronnyi resurs]. – Rezhym dostupu : http://bobo.link.cs.cmu.edu/link.

12. Inxight Summarizer [Elektronnyi resurs]. – Rezhym dostupu : http://www.inxight.com.

13. Prosum Summarizer [Elektronnyi resurs]. – Rezhym dostupu : http://transend.labs.bt.com/cgi-bin/prosum/prosum.

14. Barmak O. V. Informatsiina tekhnolohiia avtomatyzovanoho vyznachennia terminiv u navchalnykh materialakh / O. V. Barmak, O. V. Mazurets // Mizhnarodnyi naukovo-tekhnichnyi zhurnal «Vymiriuvalna ta obchysliuvalna tekhnika v tekhnolohichnykh protsesakh» – Khmelnytskyi, 2015. – # 2. – C. 94–102.

15. Pro poniattia modeli prostoho rechennia [Elektronnyi resurs]. – Rezhym dostupu : https://www.jstor.org/stable/43659932.

16. IDEF5 – Ontology Description Capture Method [Elektronnyi resurs]. – 2015. – Rezhym dostupu : http://www.idef.com/IDEF5.htm.

17. Barmak O. V. Metody avtomatyzatsii vyznachennia semantychnykh terminiv u navchalnykh materialakh / O. V. Barmak, O. V. Mazurets // Herald of Khmelnytsky National University. Seriia: Tekhnichni nauky. – Khmelnytskyi, 2015. – # 2(223). – S. 209–213.

18. Lande D. V. Kompaktyfytsyrovannyi horyzontalnyi hraf vydymosty dlia sety slov / D. V. Lande, A. A. Snarskyi // Trudy Mezhdunarodnoi nauchnoi konferentsyy «Yntellektualnyi analyz ynformatsyy YAY-2013. Znanyia y rassuzhdenyia». – Kyev : KPY,2013. – S. 158–164.

19. Krak Yu. V. Praktychne doslidzhennia efektyvnosti informatsiinoi tekhnolohii avtomatyzovanoho vyznachennia semantychnykh terminiv v kontenti navchalnykh materialiv / Yu. V. Krak, O. V. Barmak, O. V. Mazurets // Prykladne prohramne zabezpechennia : zbirnyk naukovykh prats za materialamy desiatoi mizhnarodnoi naukovo-praktychnoi konferentsii po prohramuvanniu «UkrPROH2016». – Kyiv, 2016. – S. 237–245.

Рецензія/Peer review : 27.06.2017 р. Надрукована/Printed :09.09.2017 р.
Рецензент: д.т.н., проф. Сорокатий Р.В.