Якість – гарантія успіху та конкурентоспроможності / І.О. Сидоренко // Економічні науки: науковий вісник Чернівецького торговельно-економічного інституту. – Чернівці: АНТ Лтд, 2005. – Вип. ІІІ. – С. 458 – 461. 35. Системи управління якістю. Вимоги : ISO 9001–2001.– [Чинний від 2001–104–01]. – К.: Держстандарт України, 2001. – 23 с. 36. Системи управління якістю. Настанови щодо поліпшення діяльності. Вимоги :ISO 9004–2001. – [Чинний від 2001–01–01]. – К.: Держстандарт України, 2001. – 44 с. 37. Советов Б.Я. Информационные технологии / Б.Я. Советов: учеб. для студ. вузов. – 2-е изд. – М.: Высш. шк., 2005. – 264 с. 38. Ткаченко Т.І. Управління якістю готельних послуг: монографія / Т.І. Ткаченко, Мельниченко С.В., Новак М.В. – К.: Київ. нац. торг.-екон. ун-т, 2006. – 234 с. 39. Ткаченко Т.І., Гаврилюк С.П. Аналіз конкуренції на туристичному ринку України // Вісник Донецького інституту туристичного бізнесу. – № 2, 2000. – С.113–121. 40. Шаповал М.І. Менеджмент якості : підручник / М.І. Шаповал. – К.: КОО «Знання», 2007. – 457 с. 41.Якість і конкурентоспроможність продукції [Електронний ресурс]. – Режим доступу:http://referaty.pp.ua/abstracts/ua/economicapidpriemstva/economica-pidpriemstva_ 5647.php 42. Чуковенков А.Ю. Правила оформления документов / А.Ю.Чуковенков, В.Ф. Янковая. – М.: Проспект, 2004. – 210 с. 43. Crosby P.B. Quality Is Free: The Art of Making Quality Certain / P.B. Crosby. – M.: Mentor Books, 1992. – 272 p. 44. Healy S. ISO 15489 Records management – its development and significance/S. Healy // Records Management Journal. – December 2001. – Vol. 11. – P. 133–142. 45. Juran J.M. Quality Control Handbook. /J.M. Juran J.M., F. Gryna. – M.: Mcgraw-Hill, 1998. – 1774 p. 46. Larry Shillito M. Acquiring, processing, and deploying voice of the customer / M. Larry Shillito. – St. Lucie Press, 2001. – 279 p. 47. Prideaux B. Managing Tourism and hospitality services: theory and international applications/B. Prideaux.,G. Moscardo, E. Laws. – London, 2006. – 337 p.

УДК 519.7

О.В. Каніщева
Національний технічний університет "Харківський політехнічний інститут"

ВИКОРИСТАННЯ КАРТ ВІДНОШЕНЬ (TRM) ДЛЯ АВТОМАТИЧНОГО РЕФЕРУВАННЯ

© Каніщева О.В., 2013

Запропоновано використання статистичних методів TFIDF, TLTF та Text Relationship Map (TRM) для автоматичної побудови реферату для українсько- та російськомовних текстів. Ці методи програмно реалізовано за допомогою мови програмування C++ у середовищі Borland Builder 6.0 та бази даних, створеної в Microsoft Access.

Ключові слова: автоматична обробка природної мови, автоматичне реферування, TFIDF, TLTF, Text Relationship Map (TRM).

This paper is devoted to the use of statistical TFIDF, TLTF and Text Relationship Map (TRM) methods for automatic construction of abstracts for the Russian and Ukrainian languages. These methods are implemented using C++ software programming language in Borland Builder 6.0 and databases created in Microsoft Access.

Key words: automatic processing of natural language, automatic abstracting, TFIDF, TLTF, Text Relationship Map (TRM).

Вступ

Розвиток цивілізації є причиною неухильного зростання обсягу накопичених людством знань. Мільйони паперових книг і рукописів містять інформацію різної тематики, різних галузей науки і культури, але їх все більше замінюють електронні носії. Вже існують електронні версії багатьох книг, популярні друковані видання виходять як в паперовому, так і в електронному вигляді,

кількість документів у мережі Інтернет зростає експоненціально. У зв'язку з цим виникає багато проблем, таких як класифікація, аналіз, пошук інформації, вирішення яких пов'язане з інтелектуальною обробкою великих масивів природномовних текстів.

Розвиток штучного інтелекту як наукового напряму уможливився тільки після створення електронно-обчислювальної машини (ЕОМ) (50-ті, 60-ті роки XX ст.), коли в його межах відбулося об'єднання математиків (теоретиків і прикладників), психологів і фахівців в області робототехніки, електроніки, кібернетики, для того щоб навчити ЕОМ у певному значенні думати і поводитися як людина (природний інтелект). Одночасно виникла ще одна важлива галузь, яка отримала назву обчислювальної (комп'ютерної) лінгвістики або відповідно до оригінальної назви Computational Linguistics (CL). Її завдання як наукового напряму полягали у тому, щоб навчити ЕОМ розуміти та обробляти природну мову (природномовні тексти).

Незважаючи на велику кількість теоретичних розробок, виключно важливою дотепер залишається проблема створення ефективних промислових систем у межах кожного з напрямів, що визначають сучасний рівень розвитку такої найважливішої науково-технічної галузі, як інформатика.

До кола проблем автоматичного опрацювання інформації входить автоматичне реферування й анотування науково-технічних текстів. Адже зберігати на електронних носіях усе, що зроблено людиною, немає сенсу, адже технічні описи застарівають, про них достатньо лише залишити зовнішню інформацію: автор, тема, що зроблено. Ще більше це стосується потоків інформації – їх треба сортувати у різні масиви за спільними темами, джерелами, з яких вони отримані, треба стискати змістову інформацію, формалізувати записи, розміщуючи їх у базах знань, звідки їх видобувати і видавати відповіді на запити.

Цей клас проблем є одним з найскладніших серед інших задач автоматичної обробки текстів, оскільки потребує глибокого лінгвістичного аналізу, який має виявити найінформативніші, найважливіші частини змісту тексту. А це компетенція складних інтелектуальних систем.

Найпростішим прийомом стиснення тексту є автоматичне видобування з тексту тих речень, які містять одне або більше ключових слів чи словосполучень, що є свого роду «піками» у розподілі смислової інформації тексту. Ці речення, виведені комп'ютером на екран у послідовності їх слідування, утворюють машинний реферат або анотацію документа.

Задача реферування як процес автоматизованої обробки мови

Анотація і реферат ефективно забезпечують швидкий обмін новою науково-технічною інформацією, саме вони істотно скорочують час фахівців на обробку інформації. Суть анотування і реферування полягає в максимальному зменшенні обсягу джерела інформації за істотного збереження його основного змісту.

Принциповою основою для такої компресії інформації є надмірність мови і відсутність однозначної відповідності між змістом і формою мовного твору. Під час реферування повідомлення звільняється від всього другорядного, ілюстративного, такого, що пояснює, зберігається лише основна ідея.

Анотація і реферат покликані давати лише найістотнішу інформацію про нові досягнення науки і техніки. Якщо реферат та анотація зацікавлять читача, а інформації, що міститься в них, йому виявиться недостатньо, то за вказаними у них вихідними даними можна завжди знайти першоджерело і отримати необхідну інформацію в повному обсязі. Отже, анотація і реферат виконують важливу функцію: вони здійснюють систематизацію необхідної користувачеві інформації.

Реферат і анотація належать до вторинних документальних джерел наукової інформації. Це такі документи, в яких повідомляється інформація про первинні документи. Обробка інформації передбачає процес вивчення кожного первинного документа або певної їх сукупності, наприклад, збірки статей, й підготовки інформації, що відображає найважливіші елементи цих документів. На основі використання вторинних документів комплектуються інформативні видання, такі як реферативні журнали, довідкова література, наукові переклади тощо. Здійснюючи компресію першоджерел, анотація і реферат роблять це принципово різними способами. Якщо анотація лише

перераховує ті питання, які висвітлені в першоджерелі, не розкриваючи самого змісту цих питань, то реферат не тільки перераховує всі ці питання, але і зосереджує увагу на змісті кожного з них. Можна сказати, що анотація лише повідомляє, про що йдеться у першоджерелі, а реферат інформує про те, що написано стосовно кожного з висвітлених у першоджерелі питань. Звідси випливає, що анотація є лише покажчиком для відбору першоджерел і не може їх замінити, тоді як реферат цілком може замінити саме першоджерело, оскільки відображає зміст матеріалу. Як зазначено вище, і для анотації, і для реферату характерний певний ступінь згортання інформації на основі її попереднього аналізу.

Анотування текстових документів

Анотація (от лат. *annotatio* – зауваження) – коротка характеристика змісту друкованого твору або рукопису. Вона є гранично стислою описовою характеристикою першоджерела. У ній в узагальненому вигляді вказано тематику публікації без повного розкриття її змісту. Анотація дає відповідь на питання, про що йдеться в первинному джерелі інформації.

Анотації за змістом і цільовим призначенням можуть бути довідкові й рекомендаційні. Довідкові анотації розкривають тематику документа і повідомляють певні відомості про нього, але не дають критичної оцінки. Рекомендаційні анотації містять оцінку документа з погляду його придатності для певної категорії читачів.

За охопленням змісту анотованого документа і читацького призначення розрізняють загальні та спеціалізовані анотації. Загальні анотації характеризують документ загалом і розраховані на широкий круг читачів. Спеціалізовані анотації розкривають документ лише в певних аспектах, що цікавлять вузького фахівця. Вони можуть бути зовсім короткими, такими, що складаються з декількох слів або невеликих фраз, і розгорнутими, до 20–30 рядків, але і в цьому випадку, на відміну від реферату, дають в стислій формі тільки найосновніші положення і висновки документа. У анотації вказують лише істотні ознаки змісту документа, тобто ті, які дають змогу виявити його наукове і практичне значення і новизну, відрізнити його від інших, близьких до нього за тематикою і цільовим призначенням.

Складаючи анотації, не слід переказувати зміст документів (висновки, рекомендації, фактичний матеріал). Потрібно звести до мінімуму використання складних зворотів, вживання особистих і вказівних займенників.

Загальні вимоги, що висуваються до написання анотацій, такі:

1. Сфера призначення анотації. Від цього залежить повнота охоплення і зміст завершальної частини.

2. Обсяг анотації має коливатися у межах від 500 до 2000 друкованих знаків.

3. Дотримання логічності структури, яка може відрізнятися від послідовності викладу в оригіналі.

4. Дотримання мовних особливостей анотації, зокрема:

- виклад основних положень оригіналу просто, ясно, коротко;
- уникнення повторень, зокрема й заголовка статті;
- дотримання єдності термінів і скорочень;
- використання загальноприйнятих скорочень;
- вживання безособових конструкцій типу «розглядається..., аналізується..., повідомляється…» і пасивного стану;
- уникнення використання прикметників, прислівників, вставних слів, що не впливають на зміст;
- використання деяких узагальнюючих слів і словосполучень, що забезпечують логічні зв’язки між окремими частинами висловів, наприклад, «як показано...», «…, проте», «отже…» і так далі.

Склад анотації:

Вступна частина – бібліографічний опис.

Основна частина – перелік основних проблем, які згадуються в публікації.

Завершальна частина – коротка характеристика і оцінка, призначення роботи, що анотується (кому адресується публікація).

Отже, анотація – це короткий, узагальнений опис (характеристика) тексту книги, статті. Перед текстом анотації подають вихідні дані (автор, назва, місце і час видання) в номінативній формі. Ці дані можна вводити і в першу частину анотації. Анотація зазвичай складається з трьох частин.

Зразок анотації:

Башмаков А. И., Башмаков И. А. Интеллектуальные информационные технологии / А. И. Башмаков, И. А. Башмаков. – М. : Изд-во МГТУ им. Н.Э. Баумана, 2005. – 304 с.

Интеллектуальные информационные технологии — одна из наиболее перспективных и быстро развивающихся научных и прикладных областей информатики. В учебном пособии рассматриваются ее основные направления: обработка текстов на естественном языке, моделирование знаний и базы знаний, управление знаниями, распознавание образов, нейротехнологии, интеллектуализация Internet, концептуальное программирование и др. Основное внимание уделяется математическим моделям, методам и инструментальным средствам разработки программного обеспечения интеллектуальных автоматизированных систем.

Задача реферування текстових документів

Реферат (від лат. *refero*, що означає «повідомляю») є коротким викладом змісту наукової праці, літератури з теми, зробленим письмово або у формі публічної доповіді, який розкриває його основний зміст щодо усіх питань, що також супроводжується оцінкою і висновками референта. Він має дати читачеві об'єктивне уявлення про характер роботи, показати найсуттєвіші моменти її змісту.

На відміну від анотації, реферат не тільки дає відповідь на питання, про що йдеться в первинному друкованому документі, але і що мовиться, тобто яка основна інформація міститься в реферованому першоджерелі. Реферат дає опис первинного документа, оповіщає про вихід в світ і про наявність відповідних первинних документів, також він є джерелом для отримання довідкових даних і самостійним засобом наукової інформації. Реферат може бути виконаний письмово і у формі усної доповіді.

Мета реферату – дати читачеві порівняно повне уявлення щодо висвітлених у першоджерелі питань і тим самим звільнити користувача від необхідності повного перегляду першоджерела.

Розрізняють два основні види рефератів:

- інформативний реферат (реферат-конспект);
- індикативний реферат (реферат-резюме).

Інформативний реферат містить в узагальненому вигляді всі основні положення оригіналу, відомості про методику дослідження, використання устаткування і сферу застосування. Найпоширенішою формою є інформативний реферат. У індикативному рефераті наводять не всі положення, а тільки ті, які тісно пов'язані з темою реферованого документа.

Реферати, складені за одним джерелом, називаються монографічними. Реферати, складені за декількома джерелами на одну тему, є оглядовими.

Серед численних видів рефератів виділимо спеціалізовані реферати, в яких виклад орієнтований на фахівців певної області або певного роду діяльності (наприклад, викладачів мовознавства) і враховує їх запити.

Попри різноманіття, реферати мають деякі загальні риси. У рефераті не наводять міркування та історичні екскурси. Матеріал подається у формі консультації або опису фактів. Інформація висловлюється точно, коротко, без спотворень змісту і суб'єктивних оцінок. Стислість досягається за рахунок використання термінологічної лексики, а також застосування таблиць, формул, ілюстрацій. Текст реферату не повинен бути скороченим перекладом або механічним переказом реферованого матеріалу. У ньому повинно бути виділено все те, що заслуговує на особливу увагу з погляду новизни й можливості використання в майбутній виробничій або науково-дослідній роботі. У тексті реферату не повинно бути повторень і загальних фраз, недопустиме використання прямої мови і діалогів. Доцільно включити в текст реферату основні висновки автора першоджерела.

Вони допомагають з максимальною точністю передати зміст первинних документів. У рефератах розумно використовувати скорочення термінів. Система скорочень дозволяє досягти значної економії місця без втрати змісту. Такі скорочення можуть бути і загальноприйнятими в мові (наприклад, *adj*. – прим.), і типовими для джерела.

Для мови реферату характерним є використання певних граматико-стилістичних засобів. До них належать передусім прості закінчені речення, які сприяють швидкому сприйняттю реферату. Для характеристики різних процесів можуть бути використані дієприкметникові звороти, що забезпечують економію обсягу. Вживання невизначено-особистих речень дозволяє зосередити увагу читача тільки на істотному, наприклад, «аналізують, застосовують, розглядають тощо».

Обсяг реферату коливається залежно від обсягу первинного документа і характеру реферату і може становити 1/8 або 10–15 % від обсягу першоджерела.

Аналіз існуючих методів анотування і реферування для обробки повнотекстових документів

Майже одночасно з роботами стосовно машинного перекладу почалися дослідження щодо використання ЕОМ для цілей автоматичного реферування науково-технічних текстів. Перший такий машинний експеримент проведено у 1957 р. в США. На відміну від машинного перекладу, де увага дослідників, принаймні на початковому етапі, була зосереджена на окремих реченнях, оскільки машинний переклад розуміли як переклад "фраза за фразою", в області автоматизованого реферування увага була зосереджена на більших ділянках тексту (найчастіше на абзацах), в яких концентрувалися судження стосовно однієї теми. Інакше кажучи, увага дослідників у цій області з самого початку була орієнтована на виявлення закономірностей, що визначають смислову єдність тексту.

На першому етапі цих робіт найпопулярнішими були підходи, основані на виявленні тих або інших статистичних закономірностей розподілу термінів в тексті або їх взаємного розташування в ньому [1, 2]. Надалі дослідження в області автоматизованого реферування змістилися у бік використання внутрішніх структур тексту, виявлення тієї інформаційної основи, яка організовує весь текст [3, 4]. Роботи в цьому напрямі суттєво вплинули на використання ЕОМ для створення штучних текстів.

Процес реферування здійснюється у три етапи: аналіз початкового тексту, визначення його характерних фрагментів, формування відповідного виводу. Більшість сучасних робіт концентруються навколо розробки технології реферування одного документа.

Метод складання виписок зосереджує увагу на виділенні характерних фрагментів (як правило, речень). Для цього методом зіставлення фразових шаблонів виділяються блоки з найбільшою лексичною і статистичною релевантністю. Створення підсумкового документа в цьому випадку – просто з'єднання вибраних фрагментів.

У більшості методів застосовується модель лінійних вагових коефіцієнтів, представлена у формулі (1). Основою аналітичного етапу в цій моделі є процедура призначення вагових коефіцієнтів для кожного блока тексту відповідно до таких характеристик, як розташування цього блока в оригіналі, частота появи в тексті, частота використання в ключових реченнях, а також показники статистичної значущості. Сума індивідуальних ваг, як правило, визначена після додаткової модифікації відповідно до спеціальних параметрів настроювання, пов'язаних з кожною вагою, дає загальну вагу всього блока тексту U:

$$Weight(U) := Location(U) + KeyPhrase(U) + StatTerm(U) + AddTerm(U). \quad (1)$$

Ваговий коефіцієнт розташування (Location) в цій моделі залежить від того, де у всьому тексті або в окремо узятому параграфі з'являється фрагмент – на початку, в середині або в кінці, а також чи використовується він у ключових розділах, наприклад, вступній частині або у висновках.

Ключовими фразами є лексичні або фразові резюмувальні конструкції, такі як «на закінчення», «в цій статті», «згідно з результатами аналізу» тощо. Ваговий коефіцієнт ключової фрази може залежати також і від значення терміна, яке він має в цій предметній області.

Крім того, під час призначення вагових коефіцієнтів у цій моделі враховується показник статистичної важливості (Statterm). Статистична важливість обчислюється на підставі даних, отриманих в результаті аналізу автоматичної індексації, при якому дослідники виявляють і оцінюють цілий ряд метрик, що визначають вагові коефіцієнти терміна. Ці метрики дають змогу виділити документ з-поміж інших в певному наборі документів.

Одна група метрик, наприклад, метрика TFIDF, характеризує баланс між частотою появи терміна в документі і частотою його появи в наборі документів (як правило, використовується з іншими метриками частоти і засобами нормалізації довжини).

Основою TFIDF методики є визначення вагового коефіцієнта для кожного елемента структури (див. формулу (2)). Ступінь впливу елемента на відстань між об'єктами пропорційний до значення коефіцієнта, визначеного за формулою:

$$K_{fr}(i,j) = 1 - \frac{ITF(i,j) \cdot IDF(i,j)}{N \cdot |D|}, \qquad (2)$$

де N – кількість елементів у ланцюжку; $|D|$ – кількість ланцюжків у наборі даних; $ITF(i,j)$ – величина, зворотна до частоти, що складається з i-го та j-го елементів у ланцюжку; $IDF(i,j)$ – величина, зворотна до частоти використання ланцюжків (що містять цю пару елементів) у наборі даних.

Ця модель надає можливість перегляду термінів у блоці тексту і визначення вагового коефіцієнта цього блока відповідно до додаткової наявності термінів (Addterm) – чи з'являються вони також в заголовку, в колонтитулі, першому параграфі і в призначеному для користувача профілі запиту. Виділення пріоритетних термінів, які найточніше відображають інтереси користувача, – це один зі шляхів побудови реферату або анотації для конкретної людини або групи. На рис. 1 наведено узагальнену архітектуру реферування без опори на знання.

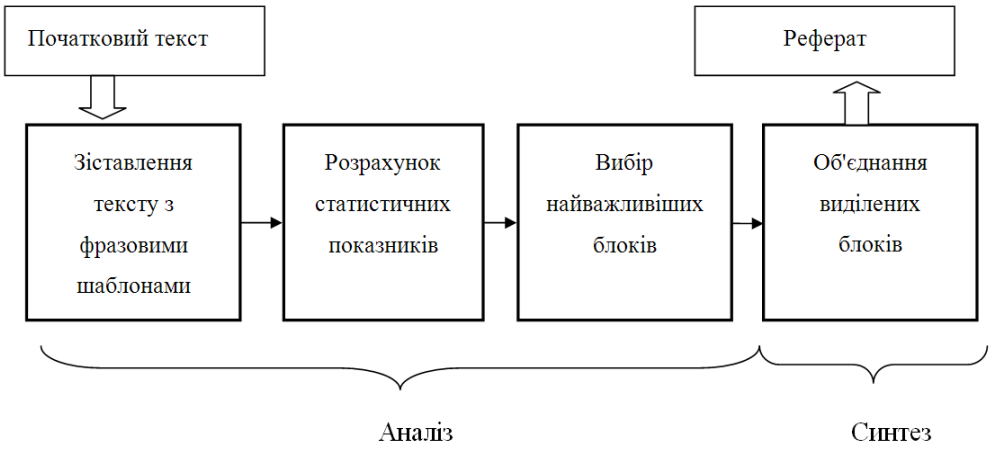

Рис. 1. Узагальнена архітектура реферування без опори на знання

На аналітичному етапі застосовується модель лінійних вагових коефіцієнтів (Naive-bayes Method), що припускає виконання послідовності обчислень частоти і операцій зіставлення рядків або шаблонів, які для кожного блока початкового тексту видають вагові коефіцієнти чотирьох типів (Location, Cuephrase, Statterm, Addterm), за формулою [5]:

$$P(s \in S | F_1, F_2, .. F_k) = \frac{\prod_{i=1}^{k} P(Fi | s \in S) \cdot P(s \in S)}{\prod_{i=1}^{k} P(Fi)}, \qquad (3)$$

де s – речення; S – блок початкового тексту; $F_1, ..., F_k$ – коефіцієнти.

Потім ці коефіцієнти підсумовують для кожного блока, після чого вибираються n блоків, що мають найвищу суму коефіцієнтів (значення n можна визначити за ступенем стиснення), для введення у реферат.

Цей метод був створений ще в 60–70-ті роки XX ст., але більшість систем, що готують такі конспекти на основі виписок, до сьогодні використовують підхід, проілюстрований на рис. 1. Аналіз порівняльних характеристик різних моделей, проведений з метою визначення продуктивності кожної з них, показав, що локалізацію блоків тексту можна вважати однією з найкорисніших функцій, особливо у поєднанні з функцією виявлення ключових фраз.

У багатьох системах користувач самостійно задає параметри, які залежать від його потреб в цей момент, тому що характеристики можуть сильно розрізнятися для текстів різного стилю. Намагаючись автоматизувати цей процес і, можливо, підвищити продуктивність, дослідники з Xerox PARC, такі як Джуліан Купьєч (1995) та його колеги, розробили класифікатор, здатний навчатися правилам виділення фрагментів. На рис. 2 показано, як цей класифікатор використовує набір визначених користувачем рефератів і відповідні початкові тексти для автоматичного визначення критеріїв адекватного вибору фрагментів [6].

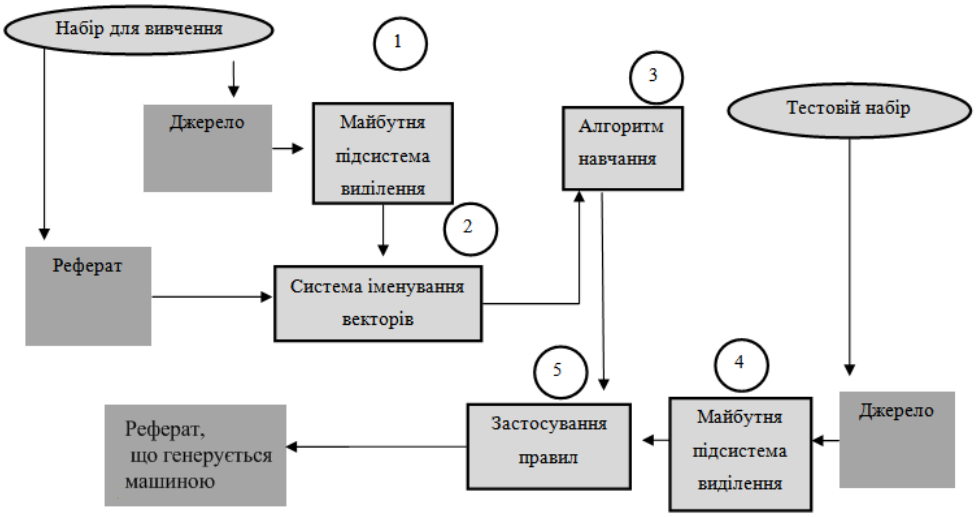

Рис. 2. Автоматичне визначення критеріїв адекватного вибору фрагментів

Цей метод, який використовують системи реферування Inxight, придатний для текстів різних стилів, але для цього користувачі повинні мати в своєму розпорядженні повні тексти і відповідні реферати для кожного стилю.

Головна перевага моделі лінійних коефіцієнтів полягає в простоті її реалізації. Проте виділення речень (або параграфів), яке не враховує взаємозв'язків між ними, призводить до формування незв'язних рефератів. Деякі речення можуть бути пропущені або в них можуть траплятися слова або словосполучення, що «висять» (тобто слова або фрази, які неможливо зрозуміти без іншого слова або фрази). Наприклад, якщо в тексті представлено обґрунтування якогось положення, що складається з декількох фраз, а до реферату потрапляє тільки одна з них, значення може бути втрачено. Можна навести текстовий фрагмент, який ілюструє цю проблему. «Біл Діксон прийшов на роботу в Procter & Gamble в 1994 році. У 1996 році він став її віце-президентом». У цьому фрагменті можна вказати два слова, що потенційно «висять», це слова «він» і «її», які не мають змісту без попередньої фрази, з якої стає зрозумілим, що «він» – це Діксон, а «її» – це компанія Procter & Gamble. Якщо в рефераті перша фраза буде пропущена, текст втратить свою інформативність. Є багато робіт, здійснюються спроби вирішити цю проблему, переважно за допомогою різного роду «латочок». У ряді підходів створюється спеціальне вікно для попереднього речення реферату, за допомогою якого можна визначити наявність смислового розриву або слова, що «висить». У інших випадках речення, що містять слова, які «висять», вилучаються з реферату, або за допомогою короткого лінгвістичного аналізу здійснюються спроби знайти згадку про об'єкт. За такого підходу ступінь стиснення зменшується, оскільки в реферат вноситься стороння інформація. Крім того, коли основний реферат вже сформовано, важко відновити початковий відсоток стиснення.

На відміну від моделі лінійних коефіцієнтів, у методах підбору виписок для підготовки короткого викладу інформації потрібні великі обчислювальні ресурси для систем обробки природних мов (NLP – Natural Language Processing), зокрема граматики і словники для синтаксичного розбору і генерації природномовних конструкцій. Крім того, для реалізації цього методу потрібні онтологічні довідники і поняття, орієнтовані на предметну область, для прийняття рішень під час аналізу і визначення найважливішої інформації. Як показано на рис. 3, метод формування короткого витягу ґрунтується на двох основних підходах [7].

Перший спирається на традиційний лінгвістичний метод синтаксичного розбору речень. У цьому методі використовується також семантична інформація для анотування дерев розбору. Процедури порівняння маніпулюють безпосередньо деревами з метою видалення і перегрупування частин, наприклад, зменшенням гілок на підставі деяких структурних критеріїв, таких як дужки або вбудовані умовні або підлеглі речення. Після такої процедури дерево розбору істотно спрощується, стаючи, по суті, структурним «витягом» початкового тексту.

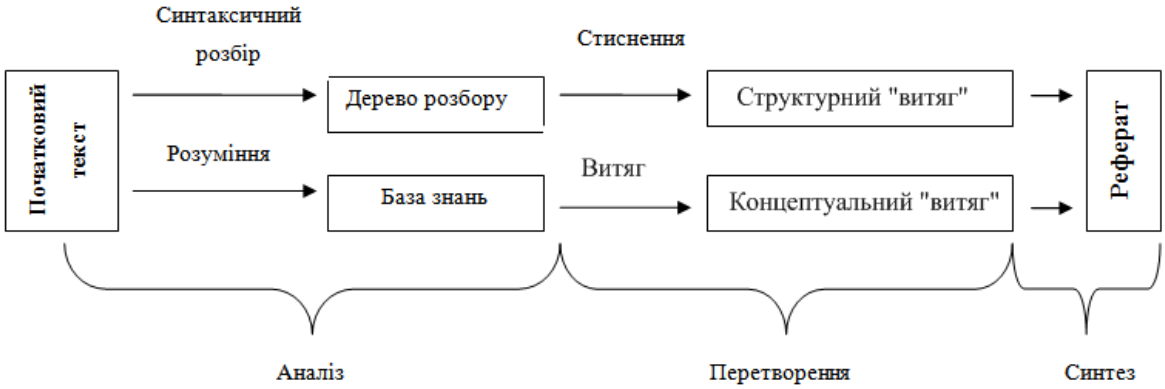

Рис. 3. Два основні підходи методу формування викладу

Другий підхід до складання короткого викладу ґрунтується на підходах штучного інтелекту і спирається на розуміння природної мови [3]. Синтаксичний розбір також є складовою частиною такого методу аналізу, але дерева розбору в цьому випадку не породжуються. Навпаки, формуються концептуальні репрезентативні структури всієї початкової інформації, які акумулюються в текстовій базі знань. Такими структурами можуть бути або формули логіки предикатів, або семантична мережа, або набір фреймів.

Прикладом може слугувати шаблон банківських транзакцій (заздалегідь визначена подія), в якому перерахованої організації та особи, що беруть у них участь, дата, обсяг перерахованих коштів, тип транзакції тощо. Представлене на рис. 3 перетворення концептуального уявлення зазнає декілька змін. Надмірна інформація, що не стосується тексту, усувається шляхом видалення поверхневих суджень або відсікання концептуальних підграфів. Потім здійснюється подальша агрегація інформації, злиттям графів (або шаблонів) або узагальненням інформації, наприклад, за допомогою таксономічних ієрархій відношень підкласів. Для виконання цих перетворень запропоновані методології на базі висновків, такі як макроправила, які маніпулюють логічними припущеннями, або оператори, які виділяють основні шаблони в текстовій базі знань [8].

У результаті перетворення формується концептуальна репрезентативна структура реферату, по суті, концептуальні «витяги» з тексту. Наявність цих формальних репрезентативних шарів (структурних і концептуальних «витягів») відрізняє підхід на основі знань від підходу, що не спирається на знання. Як видно з рис. 3, етап синтезу однаковий для обох підходів: текстовий генератор перетворить структурне або концептуальне уявлення на природномовний реферат. Деякі системи надають користувачеві можливість управляти отримуваними «витягами» і не мають етапу генерації, за умови, що початкові тексти надаються разом з їх коротким викладом. Цей тип

реферування спирається на певні структури знань, які заздалегідь указують системі реферування, яку концепцію вибрати характернішою, або які концептуальні властивості (ролі або поля) має та або інша концепція. Цей спосіб реферування повністю представляє семантичну інформацію у вигляді зв'язків між вузлами у концептуальному графові, як таксономічні (підклас або екземпляр) або метонімічні (частина) відношення. В цьому випадку він також задає напрям і критерії вибору для процедури пошуку або формування висновків. Правила виводу на базі рефератів або загальні схеми виводу (такі як термінологічна класифікація) використовують цю інформацію для того, щоб найточніше відобразити зміст тексту.

Методи створення витягів легко побудувати для обробки великих масивів інформації. Оскільки їх діяльність обмежена вибором фрагментів, речень або фраз, у результаті текст реферату буде незв'язним. З іншого боку, метод формування коротких витягів видає складніші анотації, які нерідко містять інформацію, яка доповнює початковий текст. Оскільки вони спираються на формальне представлення інформаційного наповнення документа, їх можна побудувати з дуже високим ступенем стиснення, наприклад, такі, які потрібні для розсилки повідомлень на пристрої PDA (Personal Digital Assistant). Методи заповнення шаблонів підходять тільки для текстів, побудованих за певними шаблонами, хоча засоби реферування можуть використовувати певні статистичні технології на етапі аналізу.

Методи, що ґрунтуються на знаннях, як правило, потребують повноцінних джерел знань. Ця вимога є перешкодою для їх поширення. Останні тенденції в області систем NLP на базі наборів текстів обіцяють в майбутньому вдосконалення синтаксичних аналізаторів, що охоплюватимуть широкий діапазон знань, створення ґрунтовних словників (таких як Wordnet) та онтологічних довідників (таких як Penman Upper Model). Крім того, для навчальних систем NLP напрацьовано великий обсяг текстів, зокрема набір текстових файлів, таких як The Wall Street Journal, або граматично анотованих наборів, таких як Penn Treebank консорціуму Linguistic Data.

Мета роботи

Метою роботи є: аналіз наявних моделей та методів, які використовуються для створення автоматичного реферату; виділення серед розглянутих моделей та методів найперспективніших; розгляд методів TFIDF, TLTF та Text Relationship Map (TRM) для автоматичної побудови реферату для текстів українською та російською мовами; програмна реалізація запропонованого методу.

Видобування ключових слів для побудови анотації (реферату) за допомогою моделі TFIDF

Цей підхід базується на ідеях, запропонованих у роботах Луна ще у 50-ті роки XX ст. В їх основу покладено принципи статистичної лінгвістики, такі як закон Зіпфа, який описує частотний розподіл слів у документі.

Візьмемо слова, що трапляються у тексті. Відсортуємо їх за частотою зустрічальності. Позиція слова в цьому списку називається рангом слова. Згідно з законом Зіпфа добуток рангу слова на його частоту є постійною величиною. Цей результат він отримав, вивчаючи англомовні тексти, проте надалі це було підтверджено й для інших мов [9]. Найпоширенішим методом зважування слів у документі є частота появи слова в документі TF. Частота обчислюється як відношення кількості входження слова до загальної кількості слів документа. Ця оцінка дуже популярна і є основою такого поширеного методу обчислення оцінки міри релевантності, як TFIDF.

Для зменшення значущості слів, які вживаються в багатьох реченнях, вводять інверсну частоту терміна IDF (inverse document frequency) – це логарифм відношення кількості всіх речень $|S|$ до речень, що містять певне слово t. Значення цього параметра тим менше, чим частіше слово зустрічається в документах бази даних. Отже, для слів, вживаних у великій кількості документів, IDF буде близький до нуля (якщо слово міститься у всіх документах, IDF дорівнюватиме нулю), що допомагає виділити важливі слова (формула 4).

$$IDF = \log \frac{|S|}{s_i \ni t}. \qquad (4)$$

Параметр TF (term frequency) – це відношення частоти зустрічальності слова t у документі d до довжини документа (формула 5). Нормалізація довжиною документа здійснюється для того, щоб параметр TF не залежав від довжини документа [10].

$$TF = \frac{|D|}{n_t}. \tag{5}$$

Коефіцієнт TFIDF дорівнює добутку TF и IDF. Тоді ваговими параметрами векторної моделі деякого документа можна прийняти коефіцієнти TFIDF слів, що до нього входять.

Для того, щоб ваги містилися в інтервалі (0, 1), а вектори документів мали рівну довжину, значення TFIDF зазвичай нормалізуються за косинусом. Ця формула оцінює значущість терміна тільки з огляду на частоту входження у документ, тим самим не враховуючи послідовності появи термінів у документі та їх синтаксичну роль; тобто семантика документа зводиться до лексичної семантики термінів, які до нього входять, а композиційна семантика не розглядається.

Ключовими в цьому випадку будуть слова з найбільшою вагою. Слова з малою вагою взагалі можна не враховувати.

Проілюструємо на простому прикладі. Припустимо, що документ складається з трьох речень.

Мама мила милом Машу.

Мама мила милом раму.

В магазині купила мама мило.

Вид словника наведено в таблиці.

Для того, щоб обчислити, яке слово в реченні буде ключовим, ідентифікуємо кожне речення зваженим вектором $s_i = (w_{i1}, w_{i2}, ..., w_{in})$ слів, які з'являються в документі, де n – кількість слів у документі d.

Ключові слова та їх вага

Слово	Вага	Зустрілося у реченні	IDF
Мама	3	3	0
мить	3	2	0,18
мило	2	2	0,47
Маша	1	1	0,47
рама	1	1	0,47
магазин	1	1	0,47
купить	1	1	0,47

Вага w_{ij} слова j залежить від частоти його появи в конкретному реченні i та в усьому наборі речень (у документі), вона визначається за формулою (6).

$$w_{ij} = f_{ij} \log_2(\frac{m}{m_j}),\ i = 1,...,m,\ j = 1,...,n, \tag{6}$$

де m_j – кількість речень, в яких є слово j.

Функція f_{ij} частоти появи слова j у реченні i обчислюється так:

$$f_{ij} = \frac{n_{ij}}{n \cdot len(s_i)}, \tag{7}$$

де n_{ij} – кількість появи слова j у реченні i.

Щоб уникнути зсуву, викликаного довжиною (кількістю слів) речення, функція f_{ij} нормалізується відносно довжини речення, $len(s_i)$ – довжина речення s_i.

Для визначення близькості d_{ip} між реченнями s_i та s_p найчастіше використовується евклідова відстань (формула (8)).

$$d_{ip} = \sqrt{\sum_{r=1}^{n} (w_{ir} - w_{kr})^2}. \tag{8}$$

Ми обчислили частоту появи кожного слова в документі. Проте очевидно, що значущість для слів, які дуже часто зустрічаються, повинна знижуватися, оскільки зазвичай це службові слова-прийменники тощо, для врахування цієї особливості алгоритм обчислення ваги модифікують.

Вводиться список так званих «шумових» або «стоп» слів. Цей список, як правило, формується статично для певної колекції або мови. Потім слова зводяться до нормальної форми. Деякі дослідження (Baker, Mccallum) відзначають зниження ефективності у разі використання морфологічної обробки, хоча зазвичай багато хто звертається до неї, оскільки це сприяє значному скороченню розмірності простору.

Ще одним способом скорочення словника є можливість врахування синонімії, такого, що слова-синоніми позначаються одним терміном словника.

Враховується місце появи слова в тексті. Ця характеристика обґрунтована інтуїтивними міркуваннями, тобто найважливіші з погляду автора слова розміщуються у заголовку документа або його розділів, або на початку тексту.

Звичайно, за такого підходу є ймовірність, що до ключових слів потрапляють випадкові спеціальні терміни, рідкісні слова, власні назви та інший «шум». Тому необхідно обробляти тексти з використанням алгоритмів, які підвищують якість відбору.

Метод TFIDF є найпопулярнішим. При відносній простоті ця характеристика забезпечує непогану якість пошуку. Недоліком є те, що в цьому випадку, навпаки, недооцінюються довгі документи, оскільки в них більше слів і середня частота появи слів в тексті нижча. Для боротьби з цим ефектом використовується доповнена нормалізована частота, яка обчислюється як $0.5+0.5\cdot(TF/ATF)$, де ATF – середня частота появи терміна в документі [11].

Закони Ципфа описують будь-який текст на основі частотного аналізу входження слів до тексту. Проте цього явно недостатньо для оцінки документа у колекції. Модель TFIDF дозволяє перейти до математичної, векторної моделі тексту, виділити список ключових слів.

До переваг методу належить висока продуктивність, гнучкість відносно даних. Проте у цього методу є істотний недолік: під час побудови вектора не враховується порядок слів, контекст, тобто важлива семантична складова тексту. Але за допомогою додаткових алгоритмів, таких як список «стоп» слів, обчислення евклідової відстані, виявлення синонімії можна підвищити продуктивність та ефективність методу.

Визначення ваги слів методом TLTF

Ідея методу TLTF ґрунтується на тому, що слова, які з'являються часто, переважно є короткими. Такі слова не описують основну тему документа, тобто є «стоп-словами». І навпаки, слова, які з'являються рідко, найчастіше є довгими. Перевагою використання методу TLTF для зважування слів, є те, що цей метод не вимагає ніяких зовнішніх ресурсів, і використовує тільки інформацію в межах документа [12].

У моделі для обчислення кластерів слів у реченні використовується не частота появи термів у тексті (як в багатьох методиках), а складніші правила. Представимо послідовність слів у реченні як: $b=\{w_u,...,w_v\}$, слова включаються до кластера, якщо виконуються такі умови:

- перше w_u і останнє w_v – це значущі слова у реченні;
- значущі слова розділяються заздалегідь визначеною кількістю незначущих слів.
-

Наприклад, ми можемо розділити послідовність слів у реченні так:

$$w_1\cdot[w_2\cdot w_3\cdot w_4]\cdot w_5\cdot w_6\cdot w_7\cdot w_8\cdot[w_9\cdot w_{10}\cdot w_{11}\cdot w_{12}]. \quad (9)$$

В цьому випадку речення складається з 12 слів. Значущі слова – це слова $w_2, w_4, w_9, w_{11}, w_{12}$. У квадратних дужках містяться кластери. Вони сформовані відповідно до такої умови: значущі слова мають бути розділені не більше ніж трьома незначущими словами. Необхідно зауважити, що у реченні може бути декілька кластерів, як у нашому прикладі. Значущість речення визначає найбільше значення кластера. Значення кластера у реченні s_i обчислимо за формулою (10).

$$L_{s_i} = \arg\max_b \frac{ns(b, s_i)^2}{n(b, s_i)}, \tag{10}$$

де $ns(b, s_i)$ – кількість значущих слів у кластері; $n(b, s_i)$ – загальна кількість слів у кластері.

Отже, використовуючи цю модель, за кількістю кластерів у реченні видобувають найважливіші й найбільш значущі речення для реферату.

Метод реферування документів, оснований на використанні карти тестових відношень (TRM)

У роботі використано метод побудови реферату, оснований на використанні карти текстових відношень (TRM – Text Relationship Map). Ідея методу полягає у представленні тексту у вигляді графа [13]:

$$G = (P, E), \tag{11}$$

де $P = \{p_1, p_2, \mathbf{K}, p_k, \mathbf{K}, p_n\}$ – зважені вектори слів, що відповідають фрагментам документа.

Вектор включає ваги слів, що входять до нього. Наприклад, k-й фрагмент буде представлений вектором:

$$\{w_{k,1}, w_{k,2}, ..., w_{k,i}, ...w_{k,m}\}, \tag{12}$$

де $w_{k,1}$ – вага слова, яке міститься у реченні i фрагмента k; E – множина дуг між вузлами графа: $E = \{(p_k, p_b,), p_k, p_b \in V\}$.

На рис. 4 наведено приклад такої карти. Кожен вузол на карті відповідає фрагментові тексту (речення) і представляється зваженим вектором термів.

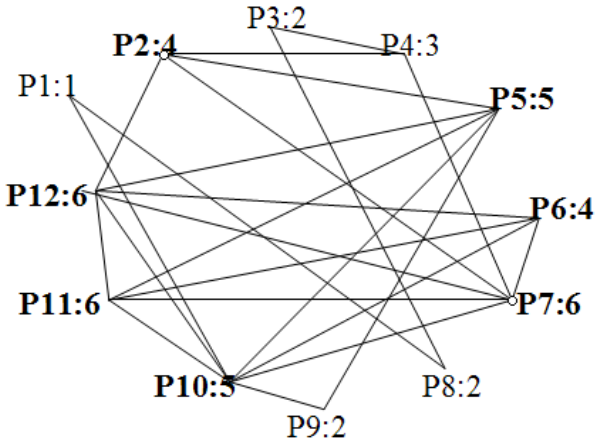

Рис. 4. Приклад карти текстових відношень

Зв'язки створюються між двома вузлами, якщо вони мають високу міру подібності між фрагментами тексту, яка зазвичай обчислюється як скалярний добуток векторів, що представляють ці фрагменти. Якщо є зв'язок між двома вузлами, то говорять, що відповідні фрагменти "семантично близькі". Кількість дуг, що входять до певного вузла, відповідає важливості фрагмента.

$$sim(p_i, p_j) = \frac{\sum_{k=1}^{m} p_{i,k} \cdot p_{j,k}}{\sqrt{\sum_{k=1}^{|m|} p^2{}_{i,k}} \cdot \sqrt{\sum_{k=1}^{m} p^2{}_{j,k}}}. \tag{13}$$

Наприклад, на рис. 4 кількість вхідних дуг вузла P_{10} дорівнює 7, оскільки в нього входять дуги від вузлів $P_1, P_5, P_6, P_7, P_9, P_{11}$, P_{12}. Це максимальне значення. Отже, вузол P_{10} своїм змістом може покрити фрагменти, що відповідають вузлам, пов'язаним з ним, і він увійде до реферату.

Основним недоліком цього підходу є те, що враховується тільки один аспект важливості фрагмента, а саме його відношення до інших фрагментів документа. Тут не розглядається інформативність слів, що містяться в окремому фрагменті. В результаті до реферату можуть потрапити фрагменти, які тісно пов'язані з іншими, але не характеризують тематику документа (тобто не містять ключових слів).

Для ліквідування цього недоліку пропонується використовувати поняття локальної і глобальної властивостей фрагмента. Локальні властивості розглядаються як кластери слів усередині речення, вага яких обчислюється методом TLTF. А глобальною властивістю виступає відношення цього речення до всіх інших у тексті, які визначаються методом TRM. Комбінуючи обидві властивості, цей метод визначає ступінь значущості речення і необхідність його включення до реферату.

Описані локальні й глобальні властивості визначають різні аспекти значущості речення. Локальна властивість визначає частину інформації усередині речення, а глобальна – звертає увагу на структурний аспект документа, оцінюючи інформативність всього речення. Для підвищення ефективності пропонується розглядати обидва об'єкти у сукупності, об'єднуючи їх в єдину оцінку інформативності речення, яка може бути використана для прийняття рішення: чи виносити це речення до реферату, чи ні. Для обчислення комбінованої оцінки використовується формула:

$$F(s_i) = l\,G' + (1-l)\,L', \tag{14}$$

де G' – нормалізована глобальна зв'язаність речення, обчислюється за формулою:

$$G' = \frac{d_{s_i}}{d_{\max}}, \tag{15}$$

де $d_{\max}$ – максимальна кількість ребер для одного вузла на карті відношень у тексті; d_{s_i} – кількість ребер для вузла відповідного речення s_i; L' – нормалізоване значення локальної кластеризації речення s_i, обчислюється за формулою:

$$L' = \frac{L_{s_i}}{L_{\max}}, \tag{16}$$

де $L_{\max}$ – максимальна локальна кластеризація в усьому тексті; l – параметр, що змінюється залежно від важливості складових G' або L'.

Отже, враховуючи всі описані методи, отримуємо інтегровану оцінку для всіх речень, за результатами якої можна вибрати речення для реферату або анотації.

Програмна реалізація методу TRM та LSA (Latent Semantic Analysis) для реферування повнотекстового документа

Розроблена система реферування дозволяє працювати з текстом, створювати реферат, вибирати найважливіші речення з тексту, аналізувати частоту появи слів. На рис. 5 представлено інтерфейс програми.

Головне меню представлене декількома пунктами: Файл, Правка, Реферування, Додаткова інформація. Програма призначена для роботи з текстовими файлами з розширенням .txt. Для того, щоб розпочати роботу з програмою, необхідно за допомогою пункту Файл головного меню відкрити файл з текстом, для якого необхідно створити реферат. Аналогічні дії можна реалізувати за допомогою кнопок на панелі інструментів.

Результат виконання наведено на рис. 5.

Під час відкриття файла на вкладці Вихідний текст відображається текст документа. Для подальшої обробки документа необхідно натиснути кнопку Виконати реферування або у головному меню вибрати пункт Реферування. Під час натиснення на кнопку створюється реферат, який відображається на вкладці Реферат.

Отже, користувач програми отримає короткий зміст документа, найважливіші речення з тексту, які допомагають зрозуміти сутність тексту. Після виконання програми користувач має можливість також подивитися інформацію про текст, побачити числове значення міри схожості речень та зрозуміти, які речення і чому програма винесла до реферату.

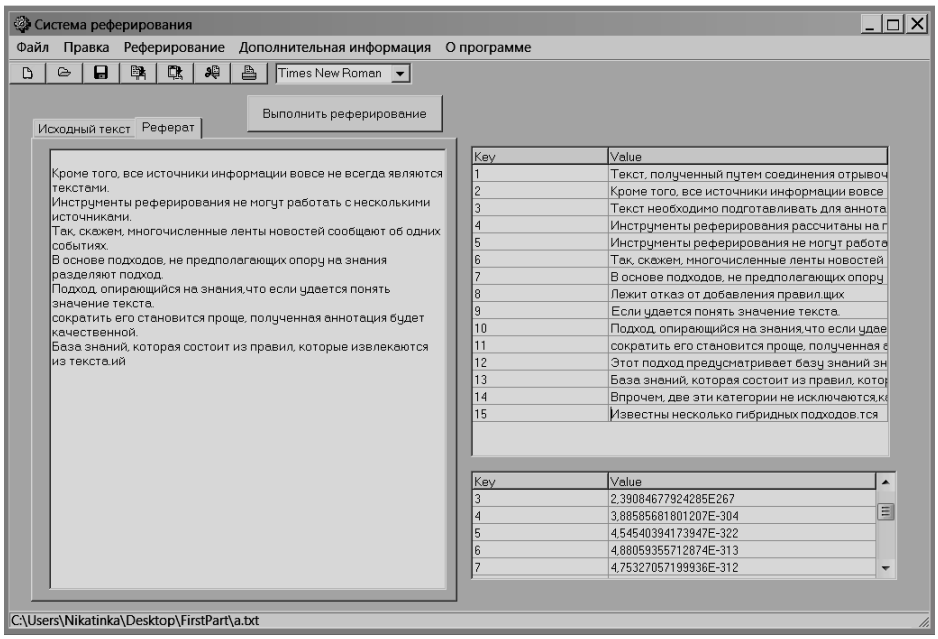

Рис. 5. Результат виконання програми

Висновки та аналіз отриманих результатів

Задача автоматичного реферування – це задача видобування змісту тексту. Існує багато програмних інструментів для створення реферату документа. Однак вони не завжди дають необхідний користувачеві результат, тому автоматичне реферування залишається однією з пріоритетних задач штучного інтелекту.

Огляд літератури, аналіз методів та підходів до проблеми автоматичного реферування показує, що для розв'язання цієї задачі важливим моментом є виділення ключових слів, словосполучень, інформаційно насичених речень тексту, штучно побудованих речень, які характеризують основний зміст тексту.

У роботі проаналізовано відомі методи та моделі для побудови автоматичного реферату. Виконано огляд промислових систем, які реалізують функції автоматичного реферування. Отже, проведені дослідження дали змогу розробити математичну модель побудови автоматичного реферату для російських та українських повнотекстових документів, основану на використанні методів латентно-семантичного аналізу (LSA – Latent Semantic Analysis), карти текстових відношень (TRM – Text Relationship Map) та метрики TFIDF(Term Frequency Inverse Document Frequency) для видобування з тексту ключових слів.

На основі цієї моделі побудовано алгоритм, який програмно реалізовано за допомогою мови програмування C++ у середовищі Borland Builder 6.0 та бази даних, створеної в Microsoft Access.

1. Михайлов А. И. Основы информатики / А. И. Михайлов, А. И. Черный, Р. С. Гиляревский. – М.: Наука, 1968. 2. Леонов Б. П. О методах автоматического реферирования (США 1958–1974 гг.) / Б. П. Леонов // Научно-техническая информация, сер.2. – 1975. – № 6. – С. 16–20. 3. Пащенко Н. А. Проблемы автоматизации индексирования и реферирования / Н. А. Пащенко, Л. В. Кнорина, Т. В. Молчанова и др. // Итоги науки и техники. Сер. Информатика. – М.: ВИНИТИ, 1983. – Т.7. – С. 7–164. 4. Севбо И. П. Структура связного текста и автоматизация реферирования / И П. Севбо. – М.: Наука, 1969. – 135 с. 5. Белоногов Г. Г. Компьютерная лингвистика и перспективные информационные технологии / Г. Г. Белоногов, Ю. П. Калинин, А. А. Хорошилов. – М: Русский мир, 2004. – 248 с. 6. Башмаков А. И. Интеллектуальные информационные технологии: учеб. пособие / А. И. Башмаков, И. А. Башмаков. – М.: Изд-во МГТУ им. Н.Э. Баумана, 2005. – 304 с. 7. Borko H. Abstracting Concepts and Methods / H. Borko, C. L. Bernier. – Academic Press, New York, 1975. 8. Iatsko V. Linguistic Aspects of Summarization / V. Iatsko // Philologie in Netz. – 2001. – № 18. –

Р. 33–46. 9. Скороходько Э. Ф. Семантические сети и автоматическая обработка текста / Э. Ф. Скороходько. – К.: Наукова думка, 1983. – 219 с. 10. Чугреев В. Л. Модель структурного представления текстовой информации и метод ее тематического анализа на основе частотно-контекстной классификации: автореф. канд. техн. наук / В. Л. Чугреев. – С-Пб., 2003. – 24 с. 11. Hahn U., Mani I. The Challenges of Automatic Summarization / U. Hahn, I. Mani // IEEE Computer Cociety. – 2000. – vol. 33, no. 11. – pp. 29-36. 12. Барсегян А. А. Технология анализа данных: Data Mining, Visual Mining, Text Mining, OLAP / А. А. Барсегян, М. С. Куприянов, В В. Степаненко, И. И. Холод. – СПб.: БХВ-Петербург, 2007. – 384 с.

УДК 811.162.1=162.2'374.822:004.65

Natalia Kotsyba
Faculty of „Artes Liberales", Warsaw University, 69, Nowy Świat str., Warsaw, 00-046, Poland

OVERVIEW OF THE UKRAINIAN LANGUAGE RESOURCES WITHIN THE MULTILINGUAL EUROPEAN MULTEXT-EAST PROJECT, V.4

© Natalia Kotsyba, 2013

Подано огляд комп'ютерних ресурсів для української мови, створених у межах багатомовного європейського проекту MULTEXT-East (MTE, http://nl.ijs.si/ME/V4), доступних безкоштовно для дослідницьких цілей від травня 2010 року. Ресурси охоплюють формальну репрезентацію морфологічно-синтаксичних специфікацій 1239 унікальних граматичних тагів у форматі XML, згідно з вимогами TEI-5, та морфологічно-синтаксичний лексикон на понад 200000 словоформ разом з лемами та тагами.

Ключові слова: комп'ютерні мовні ресурси, обробка природної мови, TEI (Ініціатива Кодування Текстів), стандарти, українська мова, морфологічно-синтаксичні специфікації, граматичний таг, лема, морфологічно-синтаксичний лексикон.

The article presents an overview of computational resources for the Ukrainian language within a multilingual European MULTEXT-East project (MTE, http://nl.ijs.si/ME/V4) freely available for researchers since May 2010, including a formal representation of morphosyntactic specifications consisting of 1239 unique grammatical tags in the XML, TEI-5 compatible, format and a morphosyntactic lexicon covering over 200000 wordforms with lemmas and morphosyntactic codes.

Key words: computational language resources, NLP, TEI, Text Encoding Initiative, standards, Ukrainian language, morphosyntactic specifications, morphosyntactic lexicon.

Introduction. Aim of the article

Due to historical reasons, developing of computational resources for the Ukrainian language was discouraged in the times of their rapid growth for widely used world languages like English or Russian, which is the reason why at present there is still no solid computational linguistic base for Ukrainian in terms of both materials and original theoretical works, cf. [14:4]. One of the consequences of this situation is a continuing strong orientation at the modern Russian corpus linguistics, which, notwithstanding the strong post-Soviet scientific heritage, itself is largely influenced by the developments of English linguistic resources. Hence, there is a considerable gap between the modern Ukrainian corpus and computational linguistics and the most recent work in this field done in the Western world. This is the reason why worldwide initiatives involving Ukrainian are beneficial for following good practices in the field, and knowledge about them should be disseminated among present and potential researchers. Thus, the purpose of the article is to present to a wide audience the existing linguistic resources for Ukrainian developed within a recent international project in a possibly accessible way, shedding also the light on some linguistic