УДК 004.8

Кліменко В.І., Живілік А.В., Мазурець О.В.

Хмельницький національний університет, Україна

АНАЛІЗ СУЧАСНИХ МЕТОДІВ АВТОМАТИЗАЦІЇ АНОТУВАННЯ ТА РЕФЕРУВАННЯ ТЕКСТІВ

У статті проведено дослідження сучасних методів та підходів до вирішення задачі автоматичного реферування та анотування текстів. Проведено аналіз особливостей застосування розглянутих методів. Встановлено актуальність подальших вишукувань у напрямку розробки ефективних методів автоматизації анотування та реферування текстів.

In this article were researched modern methods and concepts to solve the problem of automatic text summarization. Has been made an evaluation of usage aspects of these methods and determined the actuality of future researches in the branch of development of effective methods for automatization of text annotating.

Постановка проблеми

Застосування 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. Сьогодні 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. 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.

Реферування та анотування документів відносяться до числа основних видів інформаційної діяльності людини в ряду традиційних пошукових технологій. 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 і публікацій

Автоматизоване вилучення знань з тексту є однією з основних задач штучного інтелекту і безпосередньо пов’язане з розумінням текстів на природній мові. Задачу автоматизованої аналітичної обробки текстової інформації намагаються вирішити багато іноземних та вітчизняних вчених.

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 все ж ґрунтується на екстрагуванні речень з оригінального тексту для побудови тексту реферату.

Дослідження і розробки в галузі автоматичної обробки тексту (АОТ) в Європі і США привертають увагу великих приватних фірм і державних організацій найвищого рівня. Європейський Союз вже декілька років координує різні програми в галузі автоматичної обробки тексту. Наприклад, Human Language Technology Sector of the Information Society Technologies (IST) Programme 1998 – 2000 [1]. Основні розробки присвячено автоматизації процесу синтаксичного аналізу для різних систем АОІ, в тому числі й АР.

Синтаксичний аналізатор Ergo Linguistic Technologies Parser [2], розроблений Дереком Бікертоном і Філіпом Браліком з Університету Гонолулу, використовує широко відому 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 [3] створений дослідниками з Гельсінського університету. 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 пропозиції.

Один із найбільш оригінальних підходів до синтаксичного аналізу тексту – Link Parser [4] – розроблено в Carnegie-Melon University. Цей синтаксичний аналізатор – єдиний, чиї початкові коди були опубліковані он-лайн. 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 пропозиції з тексту, використовуючи статистичні алгоритми або слова-підказки. Inxight Summarizer [5] – одна з найбільш відомих комерційно поширюваних систем реферування. Inxight Summarizer був створений у Дослідницькому центрі Ксерокса в Пасло Альто.

Серед комерційних систем також можна відзначити Prosum [6] – систему реферування, розроблену British Telecommunications Laboratories у межах експериментальної комерційної он-лайн платформи TranSend, що являє собою cgi-скрипт, вбудований до веб-сторінки.

Створення систем автоматичного реферування вважається найскладнішим завданням автоматичної обробки тексту, тому що включає в себе необхідність проводити глибокий синтаксичний, семантичний, лексичний і морфологічний аналіз документа з наступним синтезом для видачі коректного результату користувачеві. І хоча поки не існує систем, здатних сформувати повноцінний реферат (вдалося створити лише системи квазіреферування), саме вони, разом з системами автоматичного пошуку і машинного перекладу, допомагають сьогодні орієнтуватися у світовому інформаційному просторі й знаходити потрібну інформацію.

Постановка задачі

Метою цієї роботи є дослідження питання автоматичного реферування та анотування тексту, розглянути основні алгоритми для розв'язання задачі автоматичного стискання текстів, порівняти їх та визначити перспективні напрями подальших досліджень.

Виклад основних матеріалів дослідження

Реферування й анотування займають центральне місце у згортанні інформації. Кожний з них має свої підходи та алгоритми вирішення поставленої проблеми, які можуть суттєво відрізнятись один від одного.

Першочерговим питанням є визначення різниці між рефератом та анотацією. Вона 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. 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, що робить більш реальним створення автоматичних рефератів.

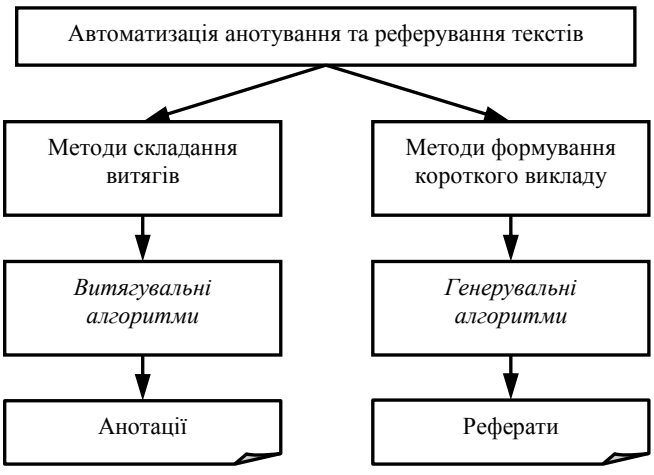

Рисунок 1 – Методи автоматизації анотування та реферування текстів

На сьогодні існує ряд підходів до розв'язання задачі автоматичного анотування (рис. 1). Їх прийнято ділити на дві групи: методи складання витягів (витягувальні алгоритми) і методи формування короткого викладу (генерувальні алгоритми) [7]. Витягувальні алгоритми формують анотацію, використовуючи текстові фрагменти вхідного документа. Для цього виділяють блоки найбільшої лексичної та статистичної значущості. У цьому випадку анотація представляє собою поєднання вибраних фрагментів. Генерувальні алгоритми аналізують вхідний документ для пошуку інформації, на основі якої формують текст анотації. Зрозуміло, що перший зі згаданих підходів є простим в реалізації і не вимагає великих обчислювальних ресурсів, однак не забезпечує достатньої якості складання анотації через відсутність семантичного аналізу

тексту. Другий підхід передбачає низку переваг: відсутність дублювання інформації в основному тексті та в анотації, повноту анотації, урахування семантичних зв'язків у тексті.

Таким чином, реферування – це інтелектуальний творчий процес, що потребує осмислення, аналітико-синтетичної переробки інформації та створення нового документа – реферату, котрий має специфічну мовно-стилістичну форму.

У цілому методи, що використовуються в автоматичному реферуванні, поділяються на статистичні, позиційні та індикативні (рис. 2).

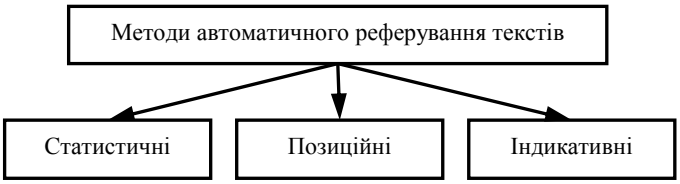

Рисунок 2 – Методи автоматичного реферування текстів

Родоначальником ***статистичних*** методів є Х.П.Лун, який першим у 1958 році отримав машинний реферат. Він запропонував здійснювати відбір речень на основі частоти вживання слів у реченні (чим частіше зустрічається слово у ньому, тим вища його семантична вага), а також зважати на місце розташування значущих слів у реченні [8]. При відборі речень до реферату для кожного з них визначається його „змістова вага". Чим більше слів, які часто трапляються, опиняються поруч, тим суттєвішу інформацію містить речення, що і має бути записане до реферату. Більшість сучасних програм з АР працюють на основі саме статистичних методів.

Подальші розробки з автоматизації реферування, засновані на статистичному аналізі текстів, – це методики російських вчених В. Аграєва та Б. Бородіна [9, 12]. Вони запропонували спосіб, згідно з яким вибрані з тексту речення пов'язані між собою і мають бути включені до реферату; відповідно, містять найбільшу кількість однаково значущих слів. Також було розроблено метод оцінки та відбору речень за кількістю інформації в них. При цьому тексти підлягають статистичному аналізу для виявлення частоти використання слів. Словами, що найчастіше вживаються у науково-технічній літературі, є терміни. Отже, чим важливіший термін, тим частіше він зустрічається у тексті, а відібрані речення міститимуть максимальну їх кількість. Обсяг одержаного в такий спосіб реферату становить, як правило, не більше трьох речень, незалежно від обсягу первинного документа [10].

У ***позиційних*** методах реферування для ідентифікації найбільш значущих речень використовують розташування речень у тексті. Позиційні методи націлені на вдосконалення технології відбору найбільш значущих речень у текстах із залученням складного математичного апарату. Відбір здійснюється на засадах чотирьох взаємопов'язаних методів: натяку, ключових слів, заголовка, локалізації [10].

Сутність *методу натяку* полягає у використанні під час відбору речень списку слів, в якому заздалегідь виділено слова з позитивною та негативною змістовою вагою, а також "нульові" (нейтральні). При відборі враховуються тільки слова, що передають позитивну й негативну оцінку.

При використанні *методу ключових слів* розглядаються слова, відібрані за частотним принципом та за цією ознакою визначені ключовими, що є аналогічним до запропонованого Луном підходу.

У *методі заголовка* головна роль відводиться словнику термінів, відібраних із заголовка та підзаголовків, які мають більшу "значущість", ніж слова з інших речень тексту. До реферату відбираються речення, в яких зустрічаються терміни, наявні у словнику.

Метод локалізації ґрунтується на припущені, що найсуттєвіша інформація концентрується на самому початку або наприкінці певного уривка чи параграфа тексту.

Зіставлення всіх чотирьох методів засвідчило, що метод ключових слів забезпечує повноту відбиття змісту первинного документа на 15-40%, метод заголовка – на 30-40%, а спільне використання методів натяку, заголовка та локалізації – на 30-60% [11].

Ці дослідження здійснювались за двома напрямами:

- екстрагування з першоджерел найбільш інформативних фрагментів тексту та формування на їх основі рефератів або анотацій (автоматичне екстрагування);
- виявлення в текстах найбільш інформативних фрагментів з наступним синтезуванням із них нових реферативних текстів.

Екстрагування (лат. extrahere – вилучати) – це дослівне алгоритмічне вилучення окремих слів, словосполучень і фраз (речень) з тексту первинного документа за допомогою ЕОМ. Відповідно, отримані внаслідок екстрагування вторинні документи називаються автоматичними екстрактами, або квазірефератами.

Автоматизоване екстрагування передбачає:

- «маркування» по всьому тексту першоджерела на основі лексико-семантичного апарата екстрагування (словника маркерів, індикаторів, конекторів);

- «редагування», що полягає в зменшенні надмірного обсягу екстрактів за рахунок вилучення речень, менш істотних з точки зору пошуку;
- побудова власне реферату-екстракту, тобто вибір із тексту речень, що залишилися після «редагування».

Подальшого розвитку підхід позиційного реферування набув під час розробки індикативних методів реферування, порівняно з якими статистичні та позиційні методи відіграють допоміжну роль.

Індикативні методи дають змогу на підставі синтаксичного аналізу формалізувати виклад основного змісту первинного документа у рефераті телеграфного стилю. Синтаксичному аналізу може підлягати як увесь текст, так і його окремі фрагменти, що містять типові маркери [9].

Нетекстова інформація (таблиці, графіки, схеми, рисунки) вилучається під час інтелектуального реферування, що передує введенню відомостей до ЕОМ. Відібраним реченням після аналізу надається позитивна чи негативна семантична вага. Крім того, визначається семантична цінність окремих елементів речення. Індикатором для виділення таких елементів виступають розділові знаки всередині речення. Наприклад, іменник-підмет має пріоритет над іменником в іншій ролі. Електронний словник з алфавітним переліком термінів і фраз є покажчиком спеціального коду, що визначає семантичну вагу, або семантичну цінність речення. Така побудова словника дає змогу після незначного редагування вводити документи різноманітної тематики та з різноаспектним висвітленням змісту. Обсяг одержаних рефератів становить у середньому до 35% обсягу оригіналу [9].

Наведені методи автоматизованого реферування постійно розвиваються й удосконалюються, проте ще досі не створено досконалого алгоритму автоматичного стискання текстів. Кожен метод має свій ряд недоліків та переваг, що дає можливість підбирати під конкретну задачу оптимальний алгоритм для її вирішення.

Висновки

Реферування текстів є однією з найважливіших галузей сучасних інформаційних технологій, оскільки кількість інформації, з якою доводиться мати справу людині, постійно зростає і настає час, коли опрацювати весь необхідний матеріал стає просто неможливим. Таким чином, розробка алгоритмів автоматичного реферування текстів не тільки не втрачає своєї актуальності, а й навпаки, стає все більш необхідною у зв'язку з постійно зростаючим обсягом актуальних текстових даних.

Таким чином, у даній статті були досліджені сучасні методи та підходи до рішення задачі автоматичного реферування та анотування текстів. Проведено аналіз їх переваг та недоліків, а також встановлено актуальність подальших вишукувань у даній області. Загалом, для вирішення завдання автоматичного реферування текстів доречно застосовувати методи позиційного підходу, а саме спільне використання методів натяку, заголовка та локалізації.

Література

1. *Human* Language Technology Sector of the Information Society Technologies (IST) Programme 1998 – 2000 / Сайт відділу природно-мовних технологій Товариства інформаційних технологій. – [Електронний ресурс]. – 2015. – Режим доступу: http://www.linglink.lu.
2. *ERGO* Linguistic Tecnologies. – [Електронний ресурс]. – 2015. – Режим доступу: http://www.ergo-ling.com.
3. *Functional* Dependency Grammar. – [Електронний ресурс]. – 2015. – Режим доступу: http://www.conexor.fi
4. *Link* Grammar Homepage . – [Електронний ресурс]. – 2015. – Режим доступу: http://bobo.link.cs.cmu.edu/link.
5. *Inxight* Summarizer . – [Електронний ресурс]. – 2015. – Режим доступу: http://www.inxight.com.
6. *Prosum* Summarizer . – [Електронний ресурс]. – 2015. – Режим доступу: http://transend.labs.bt.com/cgi-bin/prosum/prosum.
7. Ильичева Н. В. Аннотирование и реферирование / Н. В. Ильичева, А. В. Горелова, Н. Ю. Бочкарева. – Самара: Вид-во Самарського держуніверситету, 2003. – 100с.
8. *Luhn H. P.* The automatic creation of literature abstracts / H. P. Luhn // Advances in automatic text summarization. – The MIT Press, 1999. – P. 15-21.
9. Ляшенко Т. В. Преобразование информации средствами информационно-аналитических систем / Т. В. Ляшенко // Научно-техническая информация. Серия 1. — 2003. — № 6. — С. 23—25.
10. Скороходько Э. Ф. Роль системно- и текстообусловленных характеристик термина в частотном индексировании научных текстов / Э. Ф. Скороходько // Научно-техническая информация. Серия 2. — 2002. — № 8. — С. 1—6.
11. Ненич Л. Про принципи відбору ключових слів у рефератах / Л. Ненич // Вісник Книжкової палати. — 2000. — № 9. — С. 22—23.
12. Станкевич А. Ю. Формирование системы лингвистической піддержки автоматического реферирования / А. Ю. Станкевич // Научно-техническая информация. Серия 2. — 2002. — № 4. — С. 24—30.