УДК 019.941 : 004.4

Ірина Коханова,
кандидат педагогічних наук,
доцент кафедри документознавства та книгознавства факультету соціальних комунікацій Харківської державної академії культури

Проблеми та похибки методів автоматизованого реферування документів

У статті наголошено, що засоби та методи автоматизованого реферування прискорюють процеси наукового опрацювання текстів першоджерел, але можуть неточно відтворити їх семантику.

Ключові слова: *автоматизоване реферування, семантика наукових текстів, статистичні, позиційні, індикативні методи.*

Реферування як метод мікроаналітичного згортання інформації, що міститься в первинних документах, передбачає певний алгоритм дій — операцій, які входять до процесу опрацювання, але значною мірою його можна розглядати як наукову творчість обізнаного референта-аналітика.

Тривалий час — майже півстоліття — фахівці з інформаційної діяльності намагалися науково-творчий процес реферування формалізувати, полегшити, замінити інтелектуальну діяльність опрацюванням семантичної складової документу засобами ЕОМ.

Мета статті — проілюструвати переваги та недоліки методів та засобів автоматизованого реферування, довести значущість людського чинника у процесі роботи над змістом першоджерела.

Нині існує низка комп'ютерних програм, які дозволяють створювати вторинні документи у вигляді рефератів-екстрактів, індикативних рефератів тощо. Окремо існують програми перекладів тексту з однієї мови на іншу.

Загалом, методи, що використовуються в автоматичному реферуванні, поділяють на статистичні, позиційні та індикативні. Статистичні методи базуються на розробках американського вченого Г. Луна, який першим у 1958 році отримав машинний реферат. Він запропонував здійснювати відбір речень на основі частоти вживання слів у них (чим частіше зустрічається слово, тим вища його семантична вага), а також зважати на місце розташування значущих слів. При відборі речень до реферату для кожного з них визначається його "змістова вага". Чим більше слів, які часто зустрічаються, опиняються поруч — тим суттєвішу інформацію містить речення. А отже, воно має стати частиною тексту реферату. Подальші розробки з автоматизації реферування, засновані на статистичному аналізі текстів, — це методики російських вчених В. Аграєва, Б. Бородіна [2; 5]. 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 [4]. У разі використання *статистичного методу реферування* обсяг і якість рефератів повністю залежать від статистичних характеристик тексту, тому речення, що містять найважливішу інформацію (наприклад, висновки в наукових статтях) можуть бути взагалі не виділені та не ввійти до реферату. Проте визначені недоліки певною мірою компенсуються завдяки простоті аналізу та однорідності рефератів, які готуються за допомогою ЕОМ. 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 [4].

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 ті, що відібрані за частотним принципом та за цією ознакою визначені ключовими.

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 припущенні, 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% [3].

Подальшого розвитку цей підхід набув під час розробки індикативних методів реферування, порівняно з якими статистичні та позиційні 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 [2].

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 [2].

На початку 1980-х років російські науковці запропонували методику формалізованого реферування з використанням маркерів для текстів з електроніки. Згідно з нею процес автоматичного реферування зведено до двох процедур:

— власне екстрагування (тобто розпізнавання у тексті первинного документа маркерованих речень, їх компіляція та роздрукування);

— постредагування (під час цього процесу відбувається усунення логічних і змістових повторів, зайвих зворотів, а також додавання необхідних змістових зв'язок між фразами).

Постредагування виконується формалізовано, а отже, може здійснюватися не тільки фахівцями з певної галузі знань, а й бібліотекарями, бібліографами, котрі володіють основами реферування.

Слід зазначити, що навіть для автоматизованого процесу реферування потрібні кваліфіковані спеціалісти цієї галузі, щоб створювати словники.

Недолік автоматизованого аналізу тексту — це проблема усунення з повідомлення абсолютно всіх помилок. Практика свідчить, що на 100% автоматизувати редакційний етап неможливо. Бувають випадки опрацювання семантичної складової документу, коли навіть людина-редактор не може усунути з повідомлення ні практично, ні теоретично всіх помилок. Тим паче, цього не зможуть здійснити машини.

У наш час рівень, який забезпечують системи редагування, є значно нижчим, ніж у людей-редакторів. Проте це не повинно вести до припинення досліджень і конструювання таких програм.

Виявляється, формалізація норм редагування — вкрай складне завдання. Формалізованими можуть бути лише норми, подані у вигляді параметрів, списків, шаблонів і моделей, а не положень, які, як правило, стосуються семантичного аспекту редагування повідомлень.

На сучасному етапі існують методи автоматизованого виправлення орфографічних помилок:

1. Частотний, або кількісний, — сутність полягає в тому, що слово вважається значущим, якщо зустрічається в тексті більше ніж три рази. Недолік — якщо текст невеликий, то, відповідно, частотність вживання слова нижча.

2. Метод поліграмного контролю — слово розбивається на поліграми (буквенні сполучення). Машина має поліграмний словник (початок — послідовний перегляд слів та їх зіставлення зі словником. Якщо слів немає у словникові — ЕОМ зазначає помилку).

3. Метод словникового контролю (початок — послідовний перегляд слів — є слово? — ні (кінець), так — чи є слово з такою основою у машинному словникові — ні (видати на друк як помилку — 2), так — чи відсічна частина входить до парадигми допустимих закінчень — так чи ні).

Значні ускладнення та похибки відтворення змісту виникають у процесі опрацювання текстів з абревіатурами. Вони бувають ініціальні, літерні, складові (колгосп), змішані (коли поєднані великі та малі літери), галузеві, загальновживані, текстові (лише у межах певної статті). Абревіатури — найлегший вид помилок для машини, але не для людини.

Черговим ускладненням для ЕОМ під час опрацювання тексту є автоматизоване виявлення помилок, пов'язаних із милозвучністю. Наприклад, вживання з-зі-із: "з" — після будь-якої літери і перед голосним; після голосної чи паузи і перед приголосним (крім с, ш); "із" — після шиплячих і свистячих (з, с, ч, ш, щ) і перед шиплячим або свистячим; після групи приголосних і перед групою приголосних; "зі (зо)" — після будь-якої літери і перед сполученням приголосних, коли початковий з, с, ш, щ. Алгоритм: початок — послідовний перегляд слів — є слово? (ні — кінець) — це слово "з"? — чи наступне слово починається на голосний (так — 2) — чи наступне слово починається на з, с, ш, щ, ч (ні — чи стоїть "з" після паузи чи на початку речення — ні, тоді до 10, так, то до 2), так — замінити на зі. Автоматизовані методи визначення морфологічної інформації:

1. Морфологічний метод — коли на основі суфіксів і афіксів машина визначає морфологічну інформацію.

2. Словниковий. Також є метод квазіфлексій. Деякі науковці визначають його як окремий метод, інші — підметод словникового. Він полягає у тому, що 3-4-буквенні сполучення дають 80% правильної інформації. Словниковий метод має два словники: перший — незмінні частин мови — слова, які складаються з трьох букв (прийменники, займенники, частки); другий — вміщує квазіфлексії, починаючи від двох кінцевих букв, іноді навіть ціле слово. Спочатку слово перевіряють на наявність у першому словникові, йому приписується код інформації (яка частина мови). Якщо немає у першому словнику, то перевіряють у другому.

Автоматичне редагування тексту на морфологічному рівні передбачає визначення машиною морфологічної інформації слів, зокрема з категоріально-морфологічними характеристиками. Вирізняють три способи:

1. Словниковий (найефективніший, але вимагає великих обсягів пам'яті — вмістити багатотомний словник, галузеві словники; параметри — частина мови, характеристики, словозміни, афікси; принцип роботи: початкова форма, парадигми флексій) — відбувається звірка слова із словником, в якому зазначено приналежність до частин мови. Не підходить, якщо є багато помилок.

2. Власне морфологічний (80% ефективності, словник слів відсутній, є лише словник закінчень української мови, відсікається від слова найдовша з можливих флексій, підраховується кількість основ у тексті, якщо > 3, слово вважається правильно визначеним).

3. Квазіфлексійний (нетрадиційний: абревіатури, скорочення, цифри виносяться в окрему категорію, укладається словник квазіфлексій — останніх 2—3 літер слова, за ними

визначається частина мови, 90—100% ефективності, метод розроблений у НАН України) [1; 2; 5].

Отже, виходячи із зазначеного, стає зрозуміло, що якісно опрацювати семантичну складову первинного документа здатний лише висококваліфікований референт-аналітик, редактор, перекладач. Комп'ютер може випередити людину-фахівця лише у швидкості обробки введених даних, при цьому виникають суттєві похибки, перекручення змісту, висвітлення основної теми не в тому контексті, якого дотримувався автор первинного документа тощо.

Список використаної літератури

1. *Коханова І. О.* Автоматизоване реферування як засіб підвищення релевантності інформаційного пошуку / І. О. Коханова // Культура та інформаційне суспільство XXI століття : матеріали Всеукр. наук.-теорет. конф. молодих учених (Харків, 23—24 квіт. 2009 р.) / [редкол.: В. М. Шейко та ін.]. — Харків, 2009. — С. 238—239.
2. *Ляшенко Т. В.* Преобразование информации средствами информационно-аналитических систем / Т. В. Ляшенко // Научно-техническая информация. Серия 1. — 2003. — № 6. — С. 23—25.
3. *Ненич Л.* Про принципи відбору ключових слів у рефератах / Л. Ненич // Вісник Книжкової палати. — 2000. — № 9. — С. 22—23.
4. *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.
5. *Станкевич А. Ю.* Формирование системы лингвистической піддержки автоматического реферирования / А. Ю. Станкевич // Научно-техническая информация. Серия 2. — 2002. — № 4. — С. 24—30.

В статье делается акцент на то, что средства и методы автоматического реферирования позволяют ускорить процессы научной обработки текстов первоисточников, но при этом могут неточно отразить их содержание.

The article is dedicated the means and methods of automatic reviewing, whish hasten of process of scientific information activity, but they can alter the content of primary documents.

Надійшла до редакції 26 серпня 2014 року