УДК 004.93:159.95

О. В. Бісікало, д. т. н., проф.; І. О. Назаров

ОГЛЯД МЕТОДІВ АВТОМАТИЧНОГО АНОТУВАННЯ ТЕКСТІВ

У статті розглянуто методи автоматичного анотування текстів. На основі проведеного огляду запропоновано використання моделі розповсюдження обмежень для вдосконалення методу карт текстових відношень (TRM).

***Ключові слова**: автоматичне анотування, модель розповсюдження обмежень, метод TRM.*

Вступ

Анотація – це короткий виклад тексту, у якому перераховано основні висвітлені питання. Анотації класифікують за змістом і цільовим призначенням, за повнотою охоплення змісту і читацьким призначенням. За першою ознакою виділяють довідкові та рекомендаційні, за другою – загальні і спеціалізовані, як окремий вид існують оглядові анотації [1].

Уперше поняття анотації з'являється у другій половині I ст. н. е., але функціонально анотацію використовували ще в каталогах Олександрійської бібліотеки (III ст. до н. е.). Постійне накопичення і збільшення обсягів текстової інформації в умовах розвитку інформаційних технологій зумовлює актуальність задачі автоматичного анотування природно-мовних текстових матеріалів. Це завдання є одним з головних напрямків комп'ютерної лінгвістики й тісно пов'язане з автоматичним реферуванням. З урахуванням відмінностей у сутності понять анотації та реферату подібні завдання розв'язують за допомогою схожих методів.

Автоматична обробка природно-мовних текстів передбачає труднощі у процесі формалізації поставлених завдань. З іншого боку, на спосіб формалізації впливають наявність різних видів анотації (фактичного результату роботи системи) та підхід до її побудови. Загалом автоматичне анотування передбачає для даного тексту T формування іншого тексту A (анотації), який містить короткий виклад основних, висвітлених у T, питань:

$$T \to A. \tag{1}$$

Завдяки їхній дискретній природі, тексти зручно розглядати як скінченні множини $T = \{t_1, t_2, ..., t_n\}$ і $A = \{a_1, a_2, ..., a_m\}$. Елементами множини T можуть бути різні лексичні одиниці (речення, абзаци, параграфи і т. д.) залежно від її текстового розміру, а елементами множини A – лише речення (через обмеженість її текстового розміру). Якщо розглядати елементами обох указаних множин речення, то з останнього твердження випливає необхідність виконання умови

$$\{A\} << \{T\}. \tag{2}$$

Основною складністю в задачах автоматичного анотування є забезпечення збігу основного змісту тексту T з анотацією A і, власне, пошук такого змісту.

Постановка завдання

Завдання дослідження розглянути основні алгоритми генерації та витягування для розв'язання задачі автоматичного анотування природно-мовних текстів; виходячи з пріоритетності семантичного аналізу, серед алгоритмів генерування обрати метод для модифікації його з використанням підходу на основі моделі розповсюдження обмежень.

Методи автоматичного анотування

На сьогодні існує чимало підходів до розв'язання задачі автоматичного анотування. Їх прийнято ділити на дві групи: методи складання витягів (витягувальні алгоритми) і формування короткого викладу (генерувальні алгоритми). 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. Тому в цій роботі визнаємо пріоритетність генерувальних алгоритмів як таких, які мають перспективи застосування для створення систем автоматичного анотування високого рівня.

На рис. 1 наведено схему наявних методів автоматичного анотування з урахуванням наведеної вище класифікаційної ознаки.

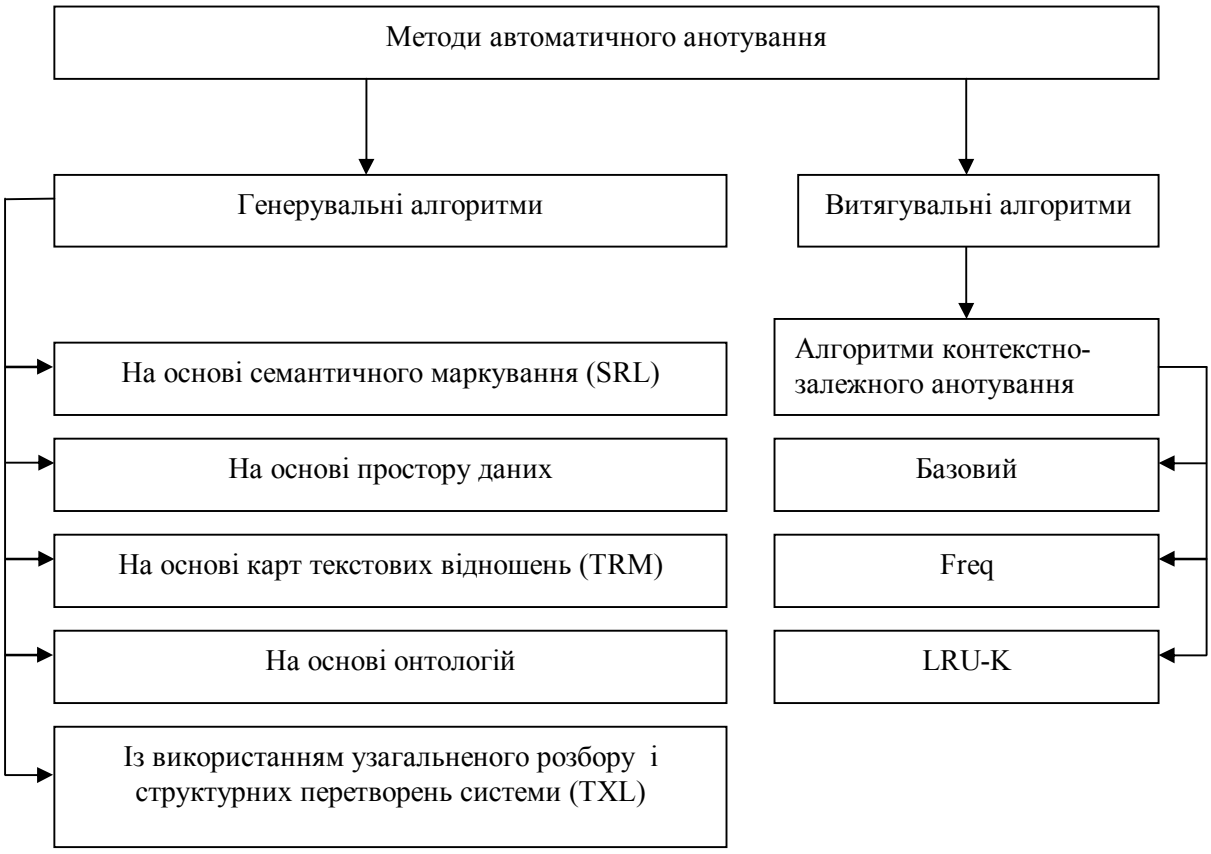

Рис. 1. Класифікація методів автоматичного анотування

Серед витягувальних алгоритмів найпопулярнішими є алгоритми контекстнозалежного анотування HTML-документів. Анотації, складені подібними методами, використовуються пошуковими системами для опису результатів у вигляді коротких неперервних фрагментів тексту відповідно до запиту користувача. Вибір оптимального фрагмента тексту здійснюють на основі розрахунку ваг фрагментів.

Базовий алгоритм

Для розрахунку ваги фрагмента в цьому алгоритмі використовують формулу

$$W = \sum_{i=1}^{n} W_i + K\frac{n}{L}, \tag{3}$$

де W_i – вага i-го слова запиту, яке увійшло до фрагмента; $K = const$; n – кількість слів запиту, які увійшли до фрагмента; L – відстань між першим і останнім словами запиту.

Вагу i-го слова W_i запиту обчислюють як

$$W_i = \frac{\log_2 N_i}{\log_2 N}, \tag{4}$$

де N_i – кількість документів, у яких трапилося i-те слово; N – загальна кількість документів.

До списку результатів пошуку вводять фрагмент тексту з найбільшою вагою. Якщо таких фрагментів більше одного, алгоритм використовує просте правило: до списку вводять найближчий до початку тексту фрагмент.

Проведені експерименти [2] свідчать про те, що базовий алгоритм є найефективнішим за швидкодією, проте якість анотування (за оцінками експертів) поступається іншим алгоритмам.

Алгоритм Freq

Цей алгоритм є вдосконаленням попереднього і, крім кількості слів пошукового запиту, ураховує слова документа, які найбільш часто повторюються. Вагу фрагмента обчислюють за формулою

$$W = W_b + \sum_{i=1}^{n} \log_2 F_i, \tag{5}$$

де W_b – вага, обчислена за базовим алгоритмом; n – кількість слів, які найбільше повторюються; F_i – частота появи i-го слова.

Алгоритм Freq значно поступається базовому за швидкодією, проте забезпечує вищу якість анотування.

Алгоритм LRU-K

Цей алгоритм запропонований в роботі [2] і є варіантом алгоритму «останній, недавно використаний». Автори застосовують оцінку локальної частоти появи слова за умови рівномірного розподілу слів. Експериментальні дослідження показали ефективність застосування запропонованого алгоритму для контекстнозалежних анотацій: якість анотування дещо вища ніж в алгоритму Freq за значно більшої швидкодії.

Алгоритм на основі семантичного маркування (SRL)

Основою алгоритму є блок аналізу семантичних структур аргумент-предикатів. Суть алгоритму полягає в семантичному маркуванні зв'язків у тексті. Для забезпечення зв'язності речень анотації маркування перевіряють предикат-структурою. Анотацію формують трирівневою обробкою:

1. Синтаксичний аналіз.
2. Побудова дерева залежностей.
3. Лексична конструкція.

Основними перевагами цього алгоритму є можливість формування повної і закінченої анотації, відсутність повторів вхідного тексту в анотації. Визначення зв'язків між словами, їх роду, відмінка, числа дозволяє заміну, відкидання, скорочення слів. Недоліками алгоритму на основі семантичного маркування є складність у реалізації, а також необхідність знання всіх зв'язків тексту. Останній недолік – окреме складне завдання, без розв'язання якого практична реалізація алгоритму неможлива.

Алгоритм на основі простору даних

Автоматизоване анотування даних про деяку подію розглянуто в роботі [3]. Задачу генерації анотації розв'язують у два етапи:

1. Інтеграція розрізненої інформації та пошук інформації про подію.
2. Анотація події та обчислення коефіцієнтів підтвердження та спростування інформації.

Для розв'язання першої із наведених вище підзадач пропонуємо підхід на основі простору даних. Простір даних – це структура, яка складається з даних (поданих у вигляді баз даних, сховищ даних, статичних веб-сторінок), локальних сховищ та індексів, засобів пошуку, опрацювання та інтеграції інформації.

Під час розв'язання підзадачі обчислення коефіцієнтів підтвердження та спростування інформації використовуємо побудову адаптивної онтології засобів інформації. Виведені формули дозволяють кількісно оцінити ці коефіцієнти. Знайдені значення використовують для побудови анотації, яка складається з двох абзаців: перший з них підтверджує подію, що розглядається, другий – спростовує. Співвідношення між коефіцієнтами дозволяє з певною ймовірністю визначити: відбулася подія чи ні.

Основним недоліком цього підходу слід визнати відсутність способу побудови абзаців анотування. Цю задачу дослідники визнають як досить складну.

Алгоритм на основі карт текстових відношень (TRM)

Алгоритм ґрунтується на використанні карти текстових відношень (Text Relationship Map - TRM) [4]. Ідея полягає у формалізації тексту у вигляді графу

$$G = (P, V), \tag{6}$$

де $P = \left\{\overline{p_1}, \overline{p_2}, ..., \overline{p_n}\right\}$ – множина вершин графу; $E = \left\{e_1, e_2, ..., e_m\right\}$ – множина ребер між вершинами.

Кожна вершина такого графу представляє фрагмент вхідного тексту і є зваженим вектором, який містить ваги окремих слів фрагменту:

$$\overline{p_i} = \left(p_{i1}, p_{i2}, ..., p_{ik}\right). \tag{7}$$

Ребра з'єднують вершини з великою мірою подібності, яку визначають як скалярний добуток векторів вершин:

$$m_{ij} = \overline{p_i}\,\overline{p_j}. \tag{8}$$

Наявність ребра між парою вершин свідчить про семантичну близькість цих фрагментів тексту. Кількість ребер, пов'язаних із вершиною, визначає важливість фрагмента тексту, представленого цією вершиною. Побудову анотації дозволяє визначення найважливіших фрагментів шляхом їх сортування за кількістю пов'язаних ребер.

Цей алгоритм забезпечує виконання смислового аналізу текстів з метою їхнього анотування. Крім того, він може бути використаний для пошуку близьких за змістом документів, поділу документів на групи за певною тематикою і т. д. Основною складністю в реалізації алгоритму є побудова карти текстових відношень, яка передбачає кількісну оцінку ваг слів фрагментів тексту й міри подібності між фрагментами.

Алгоритм з використанням узагальненого парсингу і структурних перетворень системи TXL

Запропонований у роботі [5] метод семантичного анотування документів використовує узагальнений парсинг і структурні перетворення системи TXL. TXL – це мова програмування, розроблена з метою підтримки аналізу комп'ютерного програмного забезпечення і задач перетворень документів. Процес анотування в цьому випадку складається з трьох етапів:

1. Наявний у TXL інструментарій парсингу використовують для розбору вхідного тексту, одержують апроксимовану (приблизну) структуру фраз.

2. Вказують позитивні та негативні показники семантичних категорій для списку слів, одержують первинну семантичну анотацію документа.

3. Використовують розмічений XML-текст для наповнення бази даних XML.

Цей алгоритм призначений для напівавтоматичного анотування природномовних текстів. Одержану автоматично на другому етапі початкову анотацію в ході наступного етапу корегує експерт. Таке обмеження є суттєвим недоліком подібного підходу.

Висновки

Через обмеженість обсягу цієї статті в ній висвітлено не всі наявні на сьогодні підходи до розв'язання задачі автоматичного анотування природномовних текстів. Існує велика кількість методів напівавтоматичного анотування, які потрібно виділяти в окрему групу, адже вони вимагають участі експерта в процесі складання анотації. У ході проведення дослідження автори зазначили схожість підходів до автоматичного анотування й реферування, тому для складання анотацій можна використовувати модифіковані відповідним чином методи реферування.

Як випливає з проведеного огляду, у генерувальних алгоритмах тою чи іншою мірою використовують семантичний аналіз вхідного тексту. Основним недоліком більшості наявних підходів є недосконалість проведення такого аналізу і, як наслідок, відсутність помітних успіхів у розв'язанні задачі. Тому доцільним є використання методу визначення змісту текстової інформації на основі моделі розповсюдження обмежень, запропонованому в [6]. Ефективний семантичний аналіз може бути використаний для покращення розглянутого вище алгоритму на основі карт текстових відношень (TRM).

З метою адаптації моделі текстових відношень до запропонованого підходу доцільним є представлення речень у вершинах графу, на відміну від класичного використання абзаців у якості вершин. Це дозволить удосконалити прийняття рішень у процесі анотування. На відміну від класичного підходу, пропонуємо можливість побудови графу не для окремого тексту, а для колекції документів. В такому випадку граф будуємо так: кожне речення тексту представляємо вершиною; ребра між вершинами визначають міру семантичного зв'язку речень. Раціональним є представлення тексту в адаптованому вигляді, позбавленому мовних одиниць, які не несуть семантичного значення. У якості лексичної міри подібності речень можливе використання косинусної міри [7]. У якості перспективного напрямку дослідження необхідним уважаємо вдосконалення математичного апарату цього методу.

СПИСОК ЛІТЕРАТУРИ

1. Ильичева Н. В. Аннотирование и реферирование / Н. В. Ильичева, А. В. Горелова, Н. Ю. Бочкарева. – Самара: Изд-во Самарского госуниверситета, 2003. – 100 с.

2. Губин М. В. Эффективный алгоритм формирования контекстно-зависимых аннотаций / М. В. Губин, А. И. Меркулов // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог-2005» (Звенигород, 1-6 июня 2005 г.). – 2005. – С. 116 – 120.

3. Про задачу автоматичного анотування події на основі простору даних [Електронний ресурс] / Шаховська Н. Б., Литвин В. В. // Науковий вісник Чернівецького національного університету ім. Юрія Федьковича. Збірник наук. праць. – Вип. 426: Фізика. Електроніка. – 2008. Режим доступу до журн.:

http://www.nbuv.gov.ua/portal/natural/Nvchnu_ks/2008_426/426_09_Shakhovska.pdf.

4. Митрофанов М. С. Автоматическое аннотирование документов в многокомпонентной системе поиска и анализа естественно-языковой инфорации / М. С. Митрофанов, И. Е. Чижевский // Научная сессия МИФИ-2010. Ч. 1. XIV выставка-конференция. Телекоммуникации и новые информационные технологии в образовании. – С. 156 – 159.

5. Kiyavitskaya N. Text Mining through Semi Automatic Semantic Annotation / N. Kiyavitskaya, N. Zeni, L. Mich, J. Cordy, J. Mylopoulos // PAKM 2006. LNCS (LNAI). – vol. 4333. – 2006. – P. 143 – 154.

6. Квєтний Р. Н. Визначення сенсу текстової інформації на основі моделі розповсюдження обмежень / Р. Н. Квєтний, О. В. Бісікало, І. О. Назаров // Вимірювальна та обчислювальна техніка в технологічних процесах. – 2012. – № 1. – С. 93 – 96.

7. Salton G. Automatic text processing / G. Salton. – Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, – 1988. – 450 p.

Бісікало Олег Володимирович – д. т. н., професор кафедри автоматики та інформаційно-вимірювальної техніки.

Назаров Ігор Олександрович – студент кафедри автоматики та інформаційно-вимірювальної техніки.

Вінницький національний технічний університет.