АЛГОРИТМ СИСТЕМИ АВТОМАТИЗОВАНОГО РЕФЕРУВАННЯ КОРПУСУ ТЕКСТІВ

Дашкевич О.С.
*Національний технічний університет "Харківський політехнічний інститут",
м. Харків, вул. Пушкінська, 79/2, e-mail: esdashkevich@gmail.com*

Метою роботи є дослідження і створення алгоритму системи автоматизованого реферування корпусу англомовних текстів.

Корпус текстів – це вид корпусу даних, одиницями якого є тексти або їх достатньо значні фрагменти, що включають, наприклад, уривки текстів даної проблемної області.

За останнє десятиліття корпусна лінгвістика швидко прийняла той досвід, який був накопичений в інших достатньо розвинених галузях, що об'єднуються терміном АОТ (автоматична обробка тексту), або ж АРТ (автоматичне розуміння тексту), корпусна лінгвістика сьогодні має дві лінії розвитку – лінгвістичний аналіз тексту і інформаційний аналіз тексту.

Інтелектуальний аналіз даних (ІАД, Data Mining), або добування даних – термін, що застосовується для опису здобуття знань у базах даних, дослідження даних, обробки зразків даних, очищення і збору даних. Це процес виявлення кореляції, тенденцій, шаблонів, зв'язків і категорій.

Пропонується декілька методів обробки повнотекстової інформації. На сьогодні існує чітка класифікація лінгвістичних методів та розбиття існуючих підходів на класи еквівалентності для створення єдиної термінології типових лінгвістичних шаблонів. Створення шкали таких методів дозволило чітко відрізняти одні підходи від інших. В неї входять також статистичні методи, як: тематичний аналіз, реферування, авторубрикація та класифікація.

На сучасному рівні розвитку інформаційних технологій з'явилася задача автоматизованого реферування.

Автоматизоване реферування або квазіреферування – це виявлення в тексті первинного документа фрагментів, що містять заздалегідь заявлені змістові аспекти. Найвищого розвитку формалізація методів реферування набула з автоматизацією цього виду аналітико-синтетичної обробки документів. Необхідність реферування щораз більших обсягів документів і при цьому зменшення суб'єктивізму в наданні інформації зумовили впровадження в реферування електронних технологій.

Складання рефератів формалізованим способом не виключає інтелектуальних дій людини. Насамперед, це стосується процедури редагування тексту реферату, яка передбачає такі інтелектуальні дії, як усунення дублювання в тексті, досягнення зв'язності і логічності тексту реферату-екстракту, чіткості й лаконічності викладення.

Заголовок тексту, що входить у корпус текстів, має містити так звані метадані – загальну інформацію про даний текст. Якщо включити у неї відомості про жанр та тематику тексту, можна оптимізувати корпус та здійснювати реферування текстів необхідної спрямованості.

Таким чином, для текстів однієї тематики можна визначити набір ключових слів, які зумовлюватимуть пов’язаність даних текстів та формуватимуть кістяк майбутнього реферату.

Основним методом, який використовується для реферування корпусу текстів, є статистичний. Це зумовлено тим, що корпус текстів являє собою набір попередньо розмічених текстів.

Принципом статистичного методу є ствердження, що вся необхідна для реферування інформація розташована навколо ключових слів. В даному методі, ключове слово – це знаменне слово тексту, яке з урахуванням синонімів зустрілося в тексті найбільше число разів. Речення, в яких містяться ключові слова, називаються ключовими. Саме вони використовуються для створення реферату.

Найбільш зв'язаними, а тому такими, що мають бути включені до реферату, вважаються речення, які містять найбільшу кількість однаково значущих слів. Тим не менш, у разі використання статистичного методу реферування обсяг і якість рефератів повністю залежать від статистичних характеристик тексту, тому речення, що містять найважливішу інформацію можуть бути взагалі не виділені та не ввійти до реферату.

Є кілька методів для визначення ключових фраз, такі як метод заголовку, метод позиції, метод ваги запиту та латентний семантичний аналіз. Для розробки системи, що здійснюватиме реферування корпусу текстів англійською мовою, було обрано метод ваги терміна.

Так, задача автоматизованого реферування полягає в тому, щоб створити реферат, максимально наближений за якістю до такого, який отримуємо внаслідок людської когнітивної діяльності. На основі цього ствердження нами був створений алгоритм реферування корпусу текстів англійської мови, який спирається на статистичний метод визначення ключових слів.

Ключове слово в аналізі тексту – слово, що представляє зміст тексту та отримується лінгвістичними і математичними методами (наприклад, аналізуючи частоту появи слова в тексті).

Для знаходження ключових слів у корпусі текстів необхідно:

1. За допомогою інформації про тематику тексту у метаданих визначити колекцію текстів зі спільною тематикою. Подальші дії проводитимуться з текстами однієї колекції.

2. Задля більшої оптимізації визначення ключових слів, на основі розмітки знайти у текстах колекції слова, що є самостійними частинами мови (іменник, прикметник, числівник, займенник, дієслово і прислівник). Інші слова, які не мають самостійного лексичного значення (службові частини мови), цифри, сполучення літер, вигуки і т.д., не оброблятимуться та будуть прийняті за стоп-слова.

3. Вирахувати частоту входження термінів у окремий документ колекції та у всі документи колекції. Частота входження термінів у окремий документ колекції визначається за формулою TF (term frequency – частота слова):

$$\mathrm{TF} = \frac{n_i}{\sum_k * n_k},$$

де n_i – число входжень t_i у документ, а знаменник – загальна кількість слів документу.

4. Для нормалізації df_i визначимо обратну частоту входження терміну у колекцію. Для цього використовуватимо формулу:

$$idf = \log\frac{N}{df},$$

де N – кількість документів у колекції, df – кількість документів із t_i.

5. Щоб вирахувати вагу терміну t_i у колекції, використаємо формулу: TF-idf = TF*idf. Таким чином отримаємо ключові терміни для всієї колекції документів.

6. Знайдемо речення, в яких зустрічаються ключові слова, та сформуємо на їх основі реферат колекції.

Щільність ключових слів – відношення кількості наведених в тексті ключових слів до загальної кількості слів у даному тексті. Оптимальна щільність ключових слів становить 4-6%. Таким чином, стиснення тексту досягатиме приблизно 50%.

Завдяки виділенню цільних ключових речень в реферат, мінімізується можливість порушення синтаксичних зв'язків та дрібнення смислу та ідеї тексту. Ідея полягає в тому, щоб зберегти оригінальну ідею тексту та використані в ньому семантичні одиниці.

Таким чином, в результаті аналізу методів автоматизованого реферування новинних повідомлень була розроблена математична модель реферування текстів природною мовою та запропонований алгоритм для програмної реалізації автоматизованого реферування.

Список літератури

1. Карпіловська Є. А. Вступ до прикладної лінгвістики : комп'ютерна лінгвістика / Є. А. Карпіловська. – Донецьк : Юго-Восток, 2006. – 188 с.

2. Кутузов А. Б. Корпусная лингвистика. Лекция 2 [Електронний ресурс]. – Режим доступу : http://tc.utmn.ru/files/corpus_2.pdf

3. Mitkov R. Towards automatic annotation of anaphoric links in corpora / R. Mitkov // International Journal of Corpus Linguistics 4th ed. – 1999. – P. 261-280.

4. Nenkova A., McKeown K. Automatic Summarization./ A. Nenkova, K. McKeown. – NY.: Springer US, – 2011. – pp. 216.

5. Sparck Jones K. A statistical interpretation of term specificity and its application in retrieval./ K. Sparck Jones. – L.: Journal of Documentation, –1972. – pp. 12.

6. Пятецкий-Шапиро Г.С. Data Mining и перегрузка информацией // Вступительная статья к книге: Анализ данных и процессов / А. А. Барсегян, М. С. Куприянов, И. И. Холод и др. 3-е изд. перераб. и доп. – СПб.: БХВ-Петербург, 2009. – 512 с.

7. Никоненко А.К. Огляд комп'ютерно-лінгвістичних методів обробки природномовних текстів. / А.К. Никоненко – Київ: «Искусственный интеллект» 3'2011. – 8 с.

8. Шемякин Ю.И. Начала компьютерной лингвистики: Учеб. пособие./ Ю.И. Шемякин. – М.: Росвузнаука, 1992. – 322 с.

9. Mitkov R. Towards automatic annotation of anaphoric links in corpora / R. Mitkov // International Journal of Corpus Linguistics 4th ed. – 1999. – P. 261-280.