УДК 004.912

АВТОМАТИЗОВАНИЙ АНАЛІЗ ТЕКСТІВ УКРАЇНСЬКОЮ МОВОЮ

Ковальчук С. В.
к.т.н., професор каф. СПЗ Кунгурцев О. Б.
Одеський Національний Політехнічний Університет, УКРАЇНА

АНОТАЦІЯ. Розроблені алгоритми для автоматизованого аналізу текстів українською мовою. Представлено метод виділення речень з урахуванням переліків в тексті. Описано процес виділення термінів.

Вступ. В основі проектування інформаційної системи (ІС) лежить моделювання предметної області (ПрОб). Під моделлю ПрОб розуміється деяка система, що імітує структуру або функціонування досліджуваної ПрОб і відповідає основній вимозі - бути адекватною цій галузі [1]. Тому основним завданням є створення словника деякої ПрОб, для якої буде розроблятися ІС. І тут виникає проблема щодо аналізу текстів ПрОб і виділення в них ключових слів [2].

Ціль роботи. Зменшення затраченого часу на пошук термінів, для побудови словника предметної області, шляхом автоматизованого аналізу текстів.

Основна частина. Знаходження ключових слів в наукових текстах за допомогою стеммера Портера вже було розглянуто в роботі [3]. Для синтаксичного аналізу тексту українською мовою було використано програмне забезпечення з відкритим вихідним кодом *Language Tool*, а також модуль для *Language Tool*, який призначений для роботи з текстом українською мовою *language-uk* [4]. Даний програмний модуль був модифікований для виконання поставлених цілей. Дана програма підраховує кількість входжень слів, які є іменниками у тексті.

Алгоритм роботи синтаксичного аналізатора української мови:

1. Розбивка тексту на слова (*LanguageTool*).
2. Видалення слів із списку, які повинні ігноруватись.
3. Вибирається слово. Проводиться морфологічний розбір слова (*LanguageTool*).
4. Якщо слово є іменником виконуємо для нього п. 5 Перевірка чи є попереднє і наступне слово прикметником (дієприкметником), якщо так, то узгоджуємо його із іменником і виконуємо для нього п.5.
5. Перевірка чи є вхідне слово (словосполучення) унікальним для результуючого списку, якщо так, то добавляємо його у список, якщо ні, то збільшуємо значення змінної, яка відповідає за кількість входжень слова у текст на 1.
6. Якщо у вхідному спискові є ще слова, то перехід до пункту 3.

Для того, щоб виділити терміни, потрібно здійснити два проходи по тесту. На першому проході потрібно знайти всі іменники, які будуть термінами на даному етапі.

Уявімо аналізований текст T у вигляді множини умовних речень S:

$$T = \{S_i\} i = 1, n \quad (1)$$

а кожне речення - у вигляді послідовності елементів e (слів і знаків пунктуації):

$$e_1, ..., e_l, .., e_m \quad (2)$$

Кожен елемент буде характеризуватися текстом N та множиною атрибутів A:

$$e =< N, A > \quad (3)$$

Визначимо деякі з атрибутів. Нехай $A1$ являє частину мови, $A2$ – число, $A3$ – рід, $A4$ – особа, $A5$ – відмінок, $A6$ – час, $A7$ – заставу, $A8$ – одушевленість.

Знаходимо в кожному реченні іменники $noun$. Якщо виконується умова:

$$e_l \rightarrow A1 = noun \quad (4)$$

то в якості терміну приймається $tx = e_l$.

На другому проході шукаємо частоти спільної появи виділених раніше іменників і сусідніх слів. Потрібно врахувати природні обмеження довжини стійкого словосполучення такі як: розділові знаки; слова, які не можуть входити в термін (наприклад, займенник); обмеження по довжині (не більше 3 слів зліва і з права від раніше виявленого терміну).

Нехай досліджувана група має вигляд:

$$C_{x-3} + C_{x-2} + C_{x-1} + tx_x + C_{x+1} + C_{x+2} + C_{x+3} \quad (5)$$

де $C_{x-3} - C_{x+3}$ – слова (можливі кандидати на входження в словосполучення).

Здійснюємо прохід вліво. Якщо виконується умова:

$$C_{x-1} \to agj \wedge (tx_x \to A2 = C_{x-1} \to A2) \wedge (tx_x \to A3 = C_{x-1} \to A3) \wedge (tx_x \to A5 = C_{x-1} \to A5) \quad (6)$$

то в якості терміну приймається $tx_x = C_{x-1} + tx_x$. Робимо повторні дії до кінця лівого порогу для слів C_{x-2}, C_{x-3}.

По закінченню проходу вліво, здійснюємо прохід вправо. Якщо виконується умова:

$$C_{x+1} \to agj \wedge (tx_x \to A2 = C_{x+1} \to A2) \wedge (tx_x \to A3 = C_{x+1} \to A3) \wedge (tx_x \to A5 = C_{x+1} \to A5) \quad (7)$$

то в якості терміну приймається $tx_x = tx_x + C_{x+1}$. Робимо повторні дії до кінця правого порогу для слів C_{x+2}, C_{x+3}.

Якщо вказані умови не виконуються, то нові терміни не додаються.

Кожен текст може складатися з певних перечислень, які записуються з нового рядку, при цьому не несуть повну інформацію, так як не є повноцінними реченнями. Потрібно проаналізувати текст на наявність в ньому переліків і об’єднати їх в повноцінні речення відповідно до знаків пунктуації.

Виділення повноцінних речень відбувається за допомогою регулярних виразів.

Для коректного виділення речень в тексті потрібно дотримуватися наступних правил:

1. Повноцінно закінченим реченням вважається речення, яке закінчується символом «.» або «!» або «?». Також можна вважати кінцем речення символ кінця рядку (\n).
2. Кожне наступне речення розпочинається з великої літери.
3. Якщо речення розпочинається з нумерації типу «1. Текст.», то нумерацію 1. » потрібно віднести до цього речення.
4. Якщо в реченні зустрічається скорочення типу «і т. д.», то потрібно відносити його до складу даного речення.
5. Якщо в реченні зустрічаються ініціали типу «Т. Г. Шевченко» або «Т. Шевченко», то потрібно відносити їх до складу даного речення.
6. У випадку, коли в реченні зустрічається текст в дужках «( )», потрібно все, що міститься в дужках віднести до складу даного речення.

Висновки. Розроблено метод синтаксичного аналізу тексту української мови. На основі даного методу базується метод побудови словника предметної області в роботі [5], яка дала можливість в 5 раз швидше створювати словник предметної області в порівнянні з створенням словника вручну.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Методологии моделирования предметной области [Електронний ресурс]. - Режим доступу: URL http://www.intuit.ru/studies/courses/2195/55/lecture/1628.

2. Необходимость выделения ключевых слов для свёртывания текста [Електронний ресурс]. – Режим доступу: URL http://www.scienceforum.ru/2014/476/70.

3. *Бісікало* О. В. Виявлення ключових слів на основі методу контент-моніторингу україномовних текстів/ Бісікало О. В.1, Висоцька В. А // Радіоелектроніка, інформатика, управління – №1(36), 2016. – С. 74 – 83.

4. LanguageTool [Електронний ресурс].– Режим доступу: URL https://languagetool.org/uk/.

5. *Кунгурцев*, О. Б. Побудова словника предметної області на основі автоматизованого аналізу текстів українською мовою [Текст] / О. Б. Кунгурцев, С. В. Ковальчук, Я. В. Поточняк, М. В. Широкоступ // Технічні науки та технології – №3 (5), 2016. – С. 164-174.