АНАЛІЗ СУЧАСНИХ МЕТОДІВ АВТОМАТИЧНОГО ПОШУКУ КОЛОКАЦІЙ

Гольштейн М.М., Бабкова Н.В., Угольнікова Н.С.
Національний технічний університет «Харківський політехнічний інститут», м. Харків

Колокація – це словосполучення, що має ознаки синтаксично та семантично цілісної одиниці, в якому вибір одного з компонентів здійснюється за смислом, а вибір іншого залежить від вибору першого.

Найбільш поширеними статистичними методами пошуку колокацій є PMI та T-score. Розглянемо принципи їх роботи.

Поточкова взаємна інформація (PMI) – це міра пов'язаності, що використовується в теорії інформації та статистиці. Обчислюється за наступною формулою (1):

$$\text{PMI} = \log 2 \frac{p(W, W_1)}{p(W) * P(W_1)}, \qquad (1)$$

де w – головне слово, w_1 – контекстне оточення (колокат), $p(w,w_1)$ – частота спільної зустрічальності 2 слів, p (w) p (w_1) –незалежні частоти зустрічальності 2 слів.

Обмеження PMI полягає в тому, що ця міра схильна до зміщення частоти і буде давати більш низькі або нульові значення частотних термінів у порівнянні з більш частими.

Міра T-score враховує частоту спільної зустрічальності ключового слова і його колоката, відповідаючи на питання, наскільки невипадковою є сила асоціації (зв'язаності) між колокатами. T-score обчислюється за наступною формулою (2):

$$\text{t-score} = \frac{F(w_1, w_2) - \frac{F(w_1) * F(w_2)}{N}}{\sqrt{F(w_1, w_2)}}, \qquad (2)$$

де N – ключове слово; c – колокат; $F(w,w_1)$ – частота зустрічальності ключового слова n в парі з колокатом w_2; $F(w_1)$, $F(w_2)$ – незалежні частоти ключового слова w_1 і колоката w_2 в корпусі (тексті); N – загальне число словоформ в корпусі (тексті). Формула показує, наскільки розподіли ключового слова і колоката в корпусі (тексті) залежать один від одного.

До недоліків використання цієї міри можна віднести те, що вона, в першу чергу, виділяє колокації з дуже частотними словами, зокрема, зі службовими словами. Критерій T-score спрямований на виділення стійких конструкцій, кліше, і загальномовних стійких сполучень та не використовується для вилучення термінологічних сполучень на відміну від PMI.

Проаналізувавши найбільш поширені міри асоціацій, можна сказати, що для досягнення більш точного виявлення колокацій доцільно використовувати ці міри в комбінації, або створити новий алгоритм виявлення колокацій.

На даний момент статистичні методи роботи з колокаціями в російській та українській мовах не розвинуті на такому рівні як в англійській, тому потребується подальші дослідження ефективності цих методів на слов'янській групі мов.