ОЦІНКА ЕФЕКТИВНОСТІ ВИДОБУВАННЯ ТЕРМІНІВ ПРЕДМЕТНОЇ ОБЛАСТІ З ТЕКСТІВ

Борисова Н.В.
Національний технічний університет
«Харківський політехнічний інститут»,
м. Харків, вул. Пушкінська, 79/2, тел. 707–63–60,
e-mail: *borisova_nv@mail.ru*

Відповідно до міждержавного стандарту з інформації, бібліотечної та видавницької справи для оцінки ефективності видобування термінів використовуються коефіцієнт точності *Precision*, коефіцієнт повноти *Recall*, коефіцієнт шуму *Fallout*, помилка видобування *Error*.

При проведенні випробувань щодо видобування термінів з текстів предметної області «екологія» інформаційною системою автоматизованого формування лексикографічних ресурсів було проаналізовано близько 400 електронних документів цієї предметної області.

Для проведення розрахунків використовувалися такі параметри:

– загальна кількість термінів предметної області, які містяться в аналізованих документах, відповідно до висновку експерта R;

– кількість термінів-кандидатів, видобутих системою, A;

– кількість термінів-кандидатів, видобутих системою, які дійсно є термінами відповідно до висновків експерта, R_a;

– загальна кількість термінів, які містяться в аналізованих документах, відповідно до висновку експерта, T;

Також було виділено чотири показника по відношенню до включення у перелік видобутих термінів:

$a = R_a$ – кількість видобутих термінів-кандидатів, які є термінами предметної області;

$b = A - R_a$ – кількість видобутих термінів-кандидатів, які не є термінами предметної області;

$c = R - R_a$ – кількість не видобутих термінів-кандидатів, які є термінами предметної області;

$d = (T - R) - (A - R_a)$ – кількість не видобутих термінів-кандидатів, які не є термінами предметної області.

Тоді коефіцієнт точності видобування термінів визначається за формулою

$$Precision = \frac{R_a}{A} = \frac{a}{a+b}.$$

Коефіцієнт повноти видобування визначаємо як

$$Recall = \frac{R_a}{R} = \frac{a}{a+c}.$$

Коефіцієнт шуму системи визначаємо як

$$Fallout = \frac{A - R_a}{A} = \frac{b}{a+b}.$$

Помилка видобування розраховується за формулою

$$Error = \frac{(A - R_a) + (R - R_a)}{T} = \frac{b+c}{(a+b+c+d)}.$$

Для оцінки ефективності вирішення задачі видобування термінів з текстів були отримані такі кількісні значення параметрів: $R = 4037$, $A = 3822$, $R_a = 3593$, $T = 22447$. Відповідно коефіцієнти дорівнюють:

$$Precision = \frac{3593}{3822} = 0,94;$$

$$Recall = \frac{3593}{4037} = 0,89;$$

$$Fallout = \frac{3822 - 3593}{3822} = 0,06;$$

$$Error = \frac{(3822 - 3593) + (4037 - 3593)}{22447} = 0,03.$$

Порівняння отриманих результатів роботи розроблено інформаційної системи автоматизованого формування лексикографічних ресурсів з результатами роботи подібних інформаційних систем за коефіцієнтами точності та повноти представлено у таблиці 1. Результати для інших систем отримано з науково-технічних джерел.

Таблиця 1 – Порівняння ефективності роботи розробленої системи з системами, що вирішують аналогічні задачі

Інформаційні системи, що порівнюються	**Коефіцієнт точності**	**Коефіцієнт повноти**
Розроблена	0,94	0,89
Заснована на LSPL-шаблонах	0,91	0,85
Заснована на методах, що використовують механістичні критерії	0,75	0,69

Результати порівняння показали ефективність роботи розробленої інформаційної системи щодо вирішення розглянутої задачі.

Таким чином, можна стверджувати, що інформаційна система задовольняє потреби користувача щодо видобування термінів з текстів предметної області.