УДК 004.421

К.т.н., доцент Петрашенко А.В., к.т.н, ст. викл. Заболотня Т.М., н.с. Михайлюк О.С., магістрант Гончаренко П.С.

Національний технічний університет України «Київський політехнічний інститут»

КОМБІНОВАНИЙ СТАТИСТИЧНО-СЕМАНТИЧНИЙ МЕТОД ОРФОКОРЕКЦІЇ ПРИРОДНОМОВНИХ ТЕКСТІВ

Abstract

Andriy V. Petrashenko, assoc. prof., PhD.; Tatiana M. Zabolotnya, Senior Lecturer, PhD; Olena S. Mykhailiuk, researcher; Pavlo Goncharenko, student

Combined statistical and semantic method for spelling error correction

This paper concerns the task of spelling error correction. There are a lot of different algorithms for spelling checking. But most of them don't use context of the misspelled word. The combined statistical and semantic algorithm for spelling error correction is proposed.

Вступ

На сьогоднішній день в сфері автоматичної обробки текстів різного призначення важливою є задача перевірки орфографії в текстових даних. На даний момент найбільш актуальним є дослідження алгоритмів та методів забезпечення орфокорекції саме з точки зору виправлення помилок.

Більшість сучасних орфокоректорів орієнтовані на виправлення однократних помилок та є недостатньо точними. Крім того, в програмних засобах орфокорекції при виправленні помилок досить рідко враховується контекст спотвореного слова. В даній статті пропонується комбінований алгоритм, який окрім статистичної інформації використовує також і семантичний інструментарій.

Постановка задачі

Задача полягає в розробці нового комбінованого алгоритму, який би використовував теорему Байеса для врахування статистичної інформації та лексико-семантичний словник для якіснішого врахування контексту спотвореного слова. При цьому необхідно забезпечити якнайвищу точність виправлення помилок та швидкодію алгоритму.

Теорема Байеса

Метод, який пропонується, використовує теорему Байеса та семантичний словник. Теорема Байеса визначає апостеріорну імовірність (тобто після того, як подія E доведена) гіпотези H в термінах апріорних імовірностей H і E та імовірності події E, якщо є гіпотеза H. Відносно задачі орфокорекції теорему можна сформулювати наступним чином:

$$\mathrm{P}(c \mid w) = \frac{\mathrm{P}(w \mid c) \cdot \mathrm{P}(c)}{\mathrm{P}(w)},$$

де w – спотворене слово, c – слово-виправлення, $\mathrm{P}(c \mid w)$ – імовірність появи слова c при спотвореному слові w, $\mathrm{P}(c)$ – частотність використання слова c, $\mathrm{P}(w \mid c)$ – імовірність того, що написано слово w, хоча малося на увазі слово c, $\mathrm{P}(w)$ – імовірність появи спотвореного слова w.

Більш ефективним з точки зору точності виправлення є вибір слова-виправлення c з урахуванням контексту спотвореного слова w. В даному випадку під контекстом розуміється $\pm k$ слів, які знаходяться до та після спотвореного слова. Тоді імовірність для кожного слова-виправлення з урахуванням оточуючих слів контексту можна обчислити використовуючи наступну формулу:

$$\mathrm{P}(c_i \mid w_{-k}, \ldots, w_{-1}, w_{-1}, \ldots, w_k) = \frac{\mathrm{P}(w_{-k}, \ldots, w_{-1}, w_{-1}, \ldots, w_k \mid c_i) \cdot \mathrm{P}(c_i)}{\mathrm{P}(w_{-k}, \ldots, w_{-1}, w_{-1}, \ldots, w_k)}.$$

Вираз $\mathrm{P}(w_{-k}, \ldots, w_{-1}, w_{-1}, \ldots, w_k \mid c_i)$ відображує імовірність того, що c_i – це виправлення спотвореного слова w при даному контексті $w_{-k} \ldots w_k$. Цей вираз досить складно обчислити, оскільки необхідно враховувати лише такі випадки, в яких навколо слова w були б усі слова з множини $w_{-k} \ldots w_k$, а не лише деякі. Замість цього можна припустити, що присутність одного слова в контексті не залежить від присутності будь-якого іншого слова. Це припущення дозволяє обчислити вираз $\mathrm{P}(w_{-k}, \ldots, w_{-1}, w_{-1}, \ldots, w_k \mid c_i)$ наступним чином:

$$\mathrm{P}(w_{-k}, \ldots, w_{-1}, w_{-1}, \ldots, w_k \mid c_i) = \prod_{j \in -k, \ldots, -1, 1, \ldots, k} \mathrm{P}(w_j \mid c_i).$$

Слід зазначити, що оцінити кожну імовірність $\mathrm{P}(w_j \mid c_i)$ також досить проблематично. Найбільш простий шлях вирішення цієї проблеми полягає у використанні так званої максимальної оцінки правдоподібності. Для цього необхідно порахувати M_i – загальну кількість входжень слова-виправлення c_i в тренувальний текст та m_i – число входжень слова c_i в текст в межах $\pm k$ слів контексту спотвореного слова. Після цього можливо використовувати відношення m_i / M_i в якості оцінки імовірності того, що слово c_i – це виправлення спотвореного слова з урахуванням його

контексту. Але при цьому слід мати на увазі, що в контексті можуть бути такі слова, які жодним чином не допомагають вибрати слово-виправлення. Такі слова необхідно ігнорувати в одному з двох наступних випадків:

1. $\sum_{1 \le i \le n} m_i < T_{\min}$;
2. $\sum_{1 \le i \le n} (M_i - m_i) < T_{\min}$.

Інакше кажучи, слово контексту w_j ігнорується, якщо воно практично ніколи не з'являються в оточенні будь-якого зі слів-виправлень, або якщо воно практично завжди присутнє в оточенні кожного зі слів-виправлень c_i.

Міра семантичної близькості слова до його контекстного оточення

Для визначення міри семантичної близькості понять або термінів в більшості задач автоматизованої обробки текстів використовують онтології. Формально модель онтології можна представити у вигляді упорядкованої трійки $O =< X, R, \Phi >$, де X – скінченна множина концептів (понять, термінів) предметної галузі, яку представляє онтологія O; R – скінченна множина відношень між концептами (поняттями, термінами) заданої предметної галузі; Φ – скінченна множина функцій інтерпретації, заданих на концептах та/або відношеннях онтології O. Онтологію доцільно розглядати у вигляді семантичної мережі, яка являє собою орієнтований граф, вершинами якого є поняття (X), а дугами – відношення між ними (R). Поняття, які пов'язані за змістом, є сусідніми вершинами такого графу [1].

Для задачі орфокорекції використовується лексико-семантичний словник, який має вигляд орієнтованого графу $G = (W_{dict}, E)$, вершинами якого є лексеми природної мови W_{dict}, поєднані між собою лексико-семантичними відношеннями з множини E. При цьому вважається, що всі дуги графа є рівноважними [1, 2].

Міру семантичної близькості слова до заданого контексту визначимо як суму мінімальних довжин найкоротших шляхів від заданого слова до кожного зі слів контексту за структурою семантичного словника.

Поєднання імовірнісної оцінки та міри семантичної близькості

Для більш точного вибору слова виправлення з множини варіантів доцільним є використовувати одночасно імовірнісну оцінку на основі теореми Байеса та міру семантичної близькості слова до його контекстного оточення. Вище було виведено вираз, який характеризує імовірність того, що c_i – це слово-виправлення при даному контексті:

$$S = \mathrm{P}(c_i \mid w_{-k}, \ldots, w_{-1}, w_{-1}, \ldots, w_k) =$$

$$= \frac{\mathrm{P}(w_{-k}, \ldots, w_{-1}, w_{-1}, \ldots, w_k \mid c_i) \cdot \mathrm{P}(c_i)}{\mathrm{P}(w_{-k}, \ldots, w_{-1}, w_{-1}, \ldots, w_k)} = \frac{m_i}{M_i} \cdot \mathrm{P}(c_i)$$

Для того, щоб додати до цього виразу міру семантичної близькості, необхідно ввести відповідний коефіцієнт K_i, який би виражав суму мінімальних довжин шляхів між словом-варіантом c_i та усіма словами контексту спотвореного слова. Після цього отримаємо наступний вираз:

$$S_i = \frac{m_i}{M_i} \cdot \mathrm{P}(c_i) \cdot \frac{1}{K_i}.$$

Оскільки необхідно вибрати таке слово c_i, для якого б значення виразу S було б максимальним, остаточний вираз матиме наступний вигляд:

$$S_i = \arg\max_{ci} \frac{m_i}{M_i} \cdot \mathrm{P}(c_i) \cdot \frac{1}{K_i}.$$

Саме цей вираз доцільно використовувати при виборі слова-виправлення із множини варіантів.

Висновки

Запропоновано комбінований статистично-семантичний метод виправлення орфографічних помилок шляхом поєднання імовірнісної оцінки, зробленої на основі теореми Байеса, та міри семантичної близькості слова-варіанта та контекстного оточення.

Запропоновано формулу, за результатами обчислення якої можливо приймати рішення щодо того, яке зі слів-варіантів найбільше підходить для виправлення спотвореного слова.

Також слід зазначити, що перспективним напрямком для подальших досліджень в цьому методі є вивчення впливу параметру k (половина контекстного вікна) на ефективність орфокорекції.

Література

1. *Заболотня Т.М.* Оптимізація процесу контекстноорієнтованої орфокорекції шляхом спрощення обчислення міри семантичної близькості слів // Проблеми інформатизації та управління. Збірник наукових праць. Випуск 3(21). – К.: НАУ, 2007. – С.55-59.
2. *Гаврилова Т.А., Хорошевский В.Ф.* Базы знаний интеллектуальных систем. – СПб.: Питер, 2000. – 384с.