СТАТИСТИЧНА МОДЕЛЬ ВСТАНОВЛЕННЯ ОРИГІНАЛЬНОСТІ ТЕКСТІВ

Попова В. В.
Національний технічний університет "Харківский політехнічний інститут", м. Харків, вул. Пушкінська, 79/2, тел. 707–63–60, e-mail: tamoya@ukr.net

Необхідність побудови статистичної моделі встановлення оригінальності текстів виникає, зокрема, в юриспруденції. Побудова такої моделі обумовлена необхідністю формалізованого опису портрету авторської мови, яка складається з авторської лексики та синтактики. Такий портрет складатимуть найбільш вживані лексичні одиниці (лексика) та синтаксичні обороти (синтактика), які використовує автор. Тому необхідні статистичні характеристики вказаних мовних явищ.

Для побудови такої моделі скористуємося формалізованим представленням тексту, запропонованим у роботі [1], де текст T розглядається як множина фраз, що його складають: $T = \{\theta_\alpha\}$, де θ_α – фрази тексту; $\alpha = \overline{1, n}$ – порядковий номер фрази в тексті; n – кількість фраз у тексті. Кожна фраза описується кортежем з множини синтаксичних схем і їх відображень на алфавіт слововживань, використаних в тексті: $\theta = \langle C_c, \psi \rangle$, де C_c – синтаксична схема фрази, яка є кортежем з двох множин $C_c = \langle M, A \rangle$, де $M = \{x_i\}$ – множина слововживань x_i, що входять до фрази, i – порядковий номер слововживання у фразі, $A = \{a_j\}$ – множина лінгвістичних відношень у фразі; ψ – відображення множини M на алфавіт U, $\psi : M \to U$, де U – алфавіт, множина слововживань, що складають авторський текст.

Фрази тексту складаються зі слововживань, послідовність яких знаходиться у відношенні строгого порядку і граматична форма яких встановлюється відповідно до правил природної мови, утворюючи синтаксичні схеми фраз або їх частин. Синтаксичні схеми відображають лінгвістичні відношення $A = \{a_j\}$ між слововживаннями фрази (узгодження, керування і т.п.).

У флективних мовах (російська, українська, білоруська та ін.) розрізняють п'ять типів лінгвістичних відношень: слідування, узгодження, граматичного керування, входження до складових та однорідності.

Недоліком наведеної вище моделі є ті обставини, що вона не передбачає врахування статистичної інформації про частоту зустрічальності лексики та синтактики в авторських текстах, яка необхідна для побудови образів авторської мови. Тому слід модифікувати наведене вище формалізоване представлення тексту, яке враховувало б ймовірнісні характеристики авторської лексики та синтаксичних схем (синтактики) фраз у тексті.

Особливістю лінгвістичної експертизи встановлення авторства є те, що необхідний опис образів авторської лексики і синтактики, які складають авторський образ тексту. Зазначені образи повинні містити ймовірнісні характеристики. У зв'язку з цим слід модифікувати вище наведену модель тексту шляхом введення статистичних (ймовірнісних) характеристик. Тоді, в результаті модифікації, множина слововживань авторського тексту $M = \{x_i\}$ трансформується в представлення: $M = \{x_i, k_j\}$, де x_i – слововживання, що мають значення, наприклад, граматичної категорії «іменник», «дієслово» або іншої частини мови, вжиті в авторському тексті, а k_j – ймовірність появи цих слововживання в авторському тексті. Через те, що немає необхідності у визначенні змісту лінгвістичних відношень між словоформами, множина лінгвістичних відношень $A = \{a_j\}$ модифікується в множину синтаксичних конструкцій авторського тексту наступним чином. Кожному елементу a_j ставлять у відповідність послідовність елементів x_i, де значення індексу *i* може змінюватися від 1 до 5. Вибір інтервалу індексу обґрунтовано в роботі [2] і означає, що аналізу синтаксичних конструкцій підлягають оточуючі елемент x_i слововживання. Наприклад, якщо ми аналізуємо конструкцію, в якій елементу x_i передує елемент x_{i-1} в тексті, то це відповідає синтаксичній конструкції a_{j-1}. Якщо ми аналізуємо конструкцію з елементом x_i в тексті, то це відповідає синтаксичній конструкції a_j, яка виражена одним з найчастотніших слів авторської лексики. Тоді множина авторських синтаксичних схем (синтактики) модифікується в наступне представлення: $A = \{a_j, k_f\}$, де a_j – авторська синтаксична конструкція (синтаксема) слідування, k_f – ймовірність появи цієї конструкції в авторському тексті.

Виходячи з вищенаведеного, авторський образ тексту може бути представлений кортежем наступного вигляду: $O = \langle M, A \rangle$, де $M = \{x_i, k_j\}$ – множина слововживань з високим значенням величини ймовірності вживання в авторському тексті; $A = \{a_j, k_f\}$ – множина авторських синтаксичних конструкцій (синтаксем). Значення величин ймовірності та появи в авторському тексті значущих слів і характерних синтаксичних конструкцій можуть бути отримані з використанням частотного аналізу авторського тексту з використанням закону Ципфа.

Список літератури

1. *Федорченко Л.А.* Формализованное представление фрагментов текста учебно-методического материала / Л.А. Федорченко // Вестник Международного Славянского университета. Харьков. Серия «Технические науки», т. X., 2007, № 1. с 44 – 52.

2. Кибернетическая педагогика: онтологический инжениринг в обучении и образовании [Текст] : монография / К.А. Метешкин, О.И. Морозова, Л.А. Федорченко, Н.Ф. Хайрова. – Х. : ХНАГХ, 2012. – 207 с.