DOI: 10.34185/1991-7848.itmm.2023.01.103

ЗАСТОСУВАННЯ КОНСТРУКТИВНОГО МОДЕЛЮВАННЯ ПРИ ВИЗНАЧЕННІ АВТОРСТВА ТЕКСТІВ

Демидович І.М. Шинкаренко В.І.
Український державний університет науки і технологій, Україна

Визначення авторства тексту є досить актуальною задачею на сьогоднішній день і охоплює великий спектр цілей: від встановлення автора необхідної статті в інтернеті або уривка художнього твору та аналізу текстів різних направленостей для встановлення запозичень, до досить серйозних наукових та військових цілей [1].

Сучасні експертні методи та прийоми допомагають визначити автора необхідного тексту, оскільки базуються на особливостях мови людини. Тому для визначення справжнього автора тексту часто доводиться звертатися до експертів, які можуть ідентифікувати автора невідомого тексту або визначити належність твору іншому автору за допомогою характерних мовних особливостей та різних стилістичних прийомів [2].

Важливо відзначити, що завдання встановлення авторства текстів (завдання атрибуції) зустрічається у різних галузях і цікавить філологів, літературознавців, юристів, криміналістів, істориків. В даний час для атрибуції текстів застосовуються підходи з теорії розпізнавання образів, математичної статистики та теорії ймовірностей, алгоритми нейронних мереж та кластерного аналізу та багато інших.

Атрибуція тексту – дослідження тексту з метою встановлення авторства або отримання будь-яких відомостей про автора та умови створення текстового документа. Існує чимало методів аналізу стилю. Загалом можна розділити їх на дві великі групи – експертні та формальні.

Експертні методи припускають дослідження тексту професійним лінгвістом-експертом. До формальних належать прийоми що дозволяють автоматизувати процес вирішення задачі встановлення авторства.

Запропонований підхід для встановлення автора тексту спирається на засоби конструктивізму [3], побудований на основі представлення тексту засобами формальної стохастичної граматики для відображення структури речень, що є характерною для особистого стилю письма автора [4].

Підхід на основі конструктора може бути використаний для формалізації тексту, відображення його складових. Опис структури досліджуваного тексту будуються на основі частин мови, як характеристики слова. Для цього створено

конструктор, що формалізує синтаксичну складову тексту та будує на її основі сукупність правил. Для отримання більше інформації про структуру речень та правила їх побудов, характерних для певного автора зчитуванні не тільки частини мови, а й його форма: число та рід. Для кожної частини мови прораховується ймовірність її появи у певному місці певного речення у цьому тексті. Імовірність появи певної частини мови в досліджуваній послідовності дозволить більш точно вловити індивідуальний стиль письма, характерний кожному з авторів, що досліджуються. Імовірність виведення всього речення визначається як добуток ймовірностей, які у ньому послідовностей частин мови. Отриманий конструктор породжуватиме мову, характерну для оброблюваного тексту та структурно подібних творів певного автора.

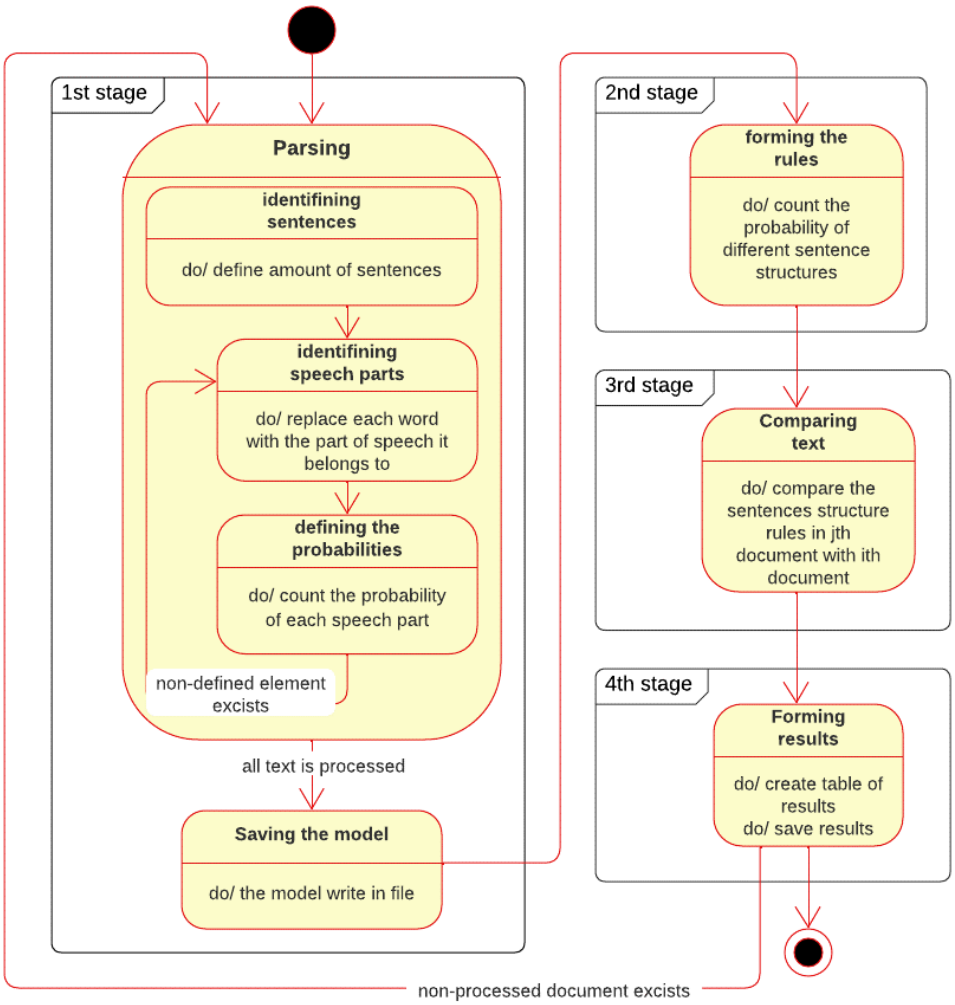

Рисунок 1 – Послідовність виконання пошуку запозичень
у природньомовних текстах за допомогою конструктора речень

Визначення подібності текстів буде прораховуватись на основі подібності їх конструкторів. При існування однакових правил або правил з частково схожою структурою ступінь їх статистичної структурної подоби

визначатиметься як добуток мінімальної різниці ймовірностей застосування відповідного правила.

Завдяки використання такого метода з'являється можливість говорити про подібність двох текстів без великої кількості необхідних обчислень та використання великої бази текстів для порівняння. Особливості конструктивного підходу дозволять більш повно відобразити авторський стиль завдяки широкому коли його можливостей та дозволить врахувати значно більше коло різних характеристик слів та побудови речень ніж існуючі методи.

Література

1. Foltýnek, Tomáš & Meuschke, Norman & Gipp, Bela. Academic Plagiarism Detection: A Systematic Literature Review. ACM Computing Surveys. 52. 2019. 1-42. 10.1145/3345317.
2. Ahuja, Lovepreet & Gupta, Vishal & Kumar, Rohit. A New Hybrid Technique for Detection of Plagiarism from Text Documents. Arabian Journal for Science and Engineering. 2020. 45. 1-14. 10.1007/s13369-020-04565-9.
3. В. І. Шинкаренко, В. М. Ільман. Конструктивно-продукційні структури та їх граматичні інтерпретації. I. Узагальнена формальна конструктивно-продукційна структура. Кібернетика та системний аналіз. – Київ, 2014. – том 50, №5 – с. 8 – 16
4. V. I. Shynkarenko, I. M. Demidovich. Natural Language Texts Authorship Establishing Based on the Sentences Structure, in: Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2022), Volume I: Main Conference, Gliwice, Poland, May 22-23, 2022, pp. 328-337.

CONSTRUCTIVE MODELING APPLICATION IN THE TEXTS AUTHORSHIP DETERMINING

Demidovich Inna, Shinkarenko Viktor

Abstract. The development of the constructor, which allows displaying the sentences construction peculiarities for different authors, is presented. This approach takes into account the sentences structure characteristic of an individual author and can be used to detect plagiarism and establish the authorship of texts in various genres and styles. Thanks to which, the usual paraphrasing of the work or changing the order of sections, sentences or words will not become an obstacle for authorship determining. The proposed approach is promising and low-cost in terms of calculation capacity, unlike the existing ones. This way of the sentences construction representation is presented for the very first.

Keywords: natural language texts, authorship determination, statistical analysis, classification, correlation coefficient, constructive-production modeling

References

1. Foltýnek, Tomáš & Meuschke, Norman & Gipp, Bela. Academic Plagiarism Detection: A Systematic Literature Review. ACM Computing Surveys. 52. 2019. 1-42. 10.1145/3345317.

2. Ahuja, Lovepreet & Gupta, Vishal & Kumar, Rohit. A New Hybrid Technique for Detection of Plagiarism from Text Documents. Arabian Journal for Science and Engineering. 2020. 45. 1-14. 10.1007/s13369-020-04565-9.

3. V. I. Shynkarenko, V. M. Ilman. Constructive-Synthesizing Structures and Their Grammatical Interpretations. I. Generalized Formal Constructive-Synthesizing Structure / Cybernetics and Systems Analysis, 2014, Volume 50, Issue 5, pp 8-16.

4. V. I. Shynkarenko, I. M. Demidovich. Natural Language Texts Authorship Establishing Based on the Sentences Structure, in: Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2022), Volume I: Main Conference, Gliwice, Poland, May 22-23, 2022, pp. 328-337.