УДК 004.912

Я. Е. Копецький, О. В. Бісікало, С. В. Голуб, В. М. Севастьянов

МЕТОД АТРИБУЦІЇ АНГЛОМОВНОГО ТЕКСТУ НА ОСНОВІ ВИЗНАЧЕННЯ І АНАЛІЗУ ЧИСЕЛЬНИХ ОЗНАК СЛОВОСПОЛУЧЕНЬ

Вінницький національний технічний університет

Анотація

Запропоновано метод атрибуції англомовного тексту на основі визначення зв'язків між лексемами та інших чисельних ознак словосполучень за допомогою вільно доступних інструментів парсингу. Програмне забезпечення використовує POS тегування, що дозволяє зменшити час обробки текстової інформації та форматувати отримані результати у вигляді таблиці MS Excel.

Ключові слова: метод, атрибуція тексту, модель, англомовний, лексема, словосполучення, чисельні ознаки, парсинг, програмне забезпечення.

Abstract

The software for determining numerical characteristics of phrases of a text file is offered on the basis of freely available parsing tools. The approach uses POS tags, which reduces the processing time of the test information and format the results in the form of a MS Excel spreadsheet.

Keywords: method, text attribution, model, english, token, word combination, numerical signs, parsing, software.

Вступ

Збільшення обсягів медіа та текстових даних у мережі Інтернет у геометричні прогресії висуває нові вимоги до оброблення неструктурованої інформації. Надзвичайно важливим для користувачів Інтернет питанням є довіра до отриманої інформації, особливо, якщо це впливає на прийняття ними рішень. Тому зростає потреба у методах і моделях оброблення великих масивів даних, у першу чергу текстових, з метою атрибуції отриманої інформації, зокрема визначення авторства тексту. Такі методи і технології актуальні не тільки у дистанційному навчанні [1], але й у багатьох інших сферах застосування інформаційних технологій – правоохоронній діяльності, підтримці новинних агенцій та ЗМІ, кібербезпеці тощо. Більшість відомих моделей комп'ютерної лінгвістики [2] базуються на результатах парсингу неструктурованої текстової інформації на основі окремих слів визначеної природної мови та бази знань її граматичних, морфологічних, синтаксичних та інших правил [3]. Проте підвищення точності атрибуції тексту та зменшення часу на проведення парсингу все ще лишаються проблемними питаннями і вимагають застосування нових методів, зокрема на основі моделювання психолінгвістичних процесів мислення людини [4].

Мета роботи полягає у розробленні методу визначення та узагальнення чисельних ознак у словосполученнях англомовних текстових файлів з метою їх атрибуції.

Результати дослідження

Для досягнення мети дослідження було застосовано відому техніку парсингу "зверху-донизу", при цьому Стенфордський парсер [3] оперує такими поняттями:

- речення – це група символів над деяким алфавітом;
- лексема – найнижчий рівень синтаксичної одиниці мови (наприклад, загальна, початкова);
- маркер – певна категорія лексем;
- ключові слова та зарезервовані слова – це ідентифікатор, який використовується як фіксована частина синтаксису оператора; інакше – це зарезервоване слово, яке не можна використовувати як ім'я змінної або ідентифікатор;
- шумові слова – це слова, що не мають вагомого змісту та/або не є обов'язковими, зазвичай вставляються в текст для поліпшення читабельності речення [4];
- розмежувачі – синтаксичні елементи, які позначає початок або кінець деякої синтаксичної одиниці (наприклад, оператор або вираз, "begin", "end" або {});
- ідентифікатор – це обмеження на довжину, яке допомагає збільшити читабельність речення.

В роботі запропоновано новий метод атрибуції англомовного тексту, який, на відміну від існуючих, базується на лінгвістичній моделі визначення зв’язків між лексемами та іншими чисельними ознаками словосполучень у реченнях тексту згідно [5]. У методі були застосовані методи машинного навчання за новими формальними ознаками множини речень тексту [6, 7], що дозволило підвищити якість визначення авторства англомовного тексту. На відміну від моделей образного мислення людини [5, 6], для програмної реалізації методу всі феноменологічні поняиия було спрощено до класифікації [3].

Програмне забезпечення методу було розроблено на основі технології – мова програмування C# та фреймворк Windows Forms. На початку роботи користувач вносить уривок твору та активує чекбокс, повідомляючи програмі, що даний твір написаний одним автором (в іншому випадку чекбокс залишається пустим). Після натискання на кнопку – через певний невеликий проміжок часу, що залежить від обчислювальних характеристик комп’ютера – в таблиці результатів з’явиться рядок з результатами обробки. Повторюємо даний крок, поки не залишиться необроблених уривків. На рисунку 1 представлено скріншот результатів роботи програми.

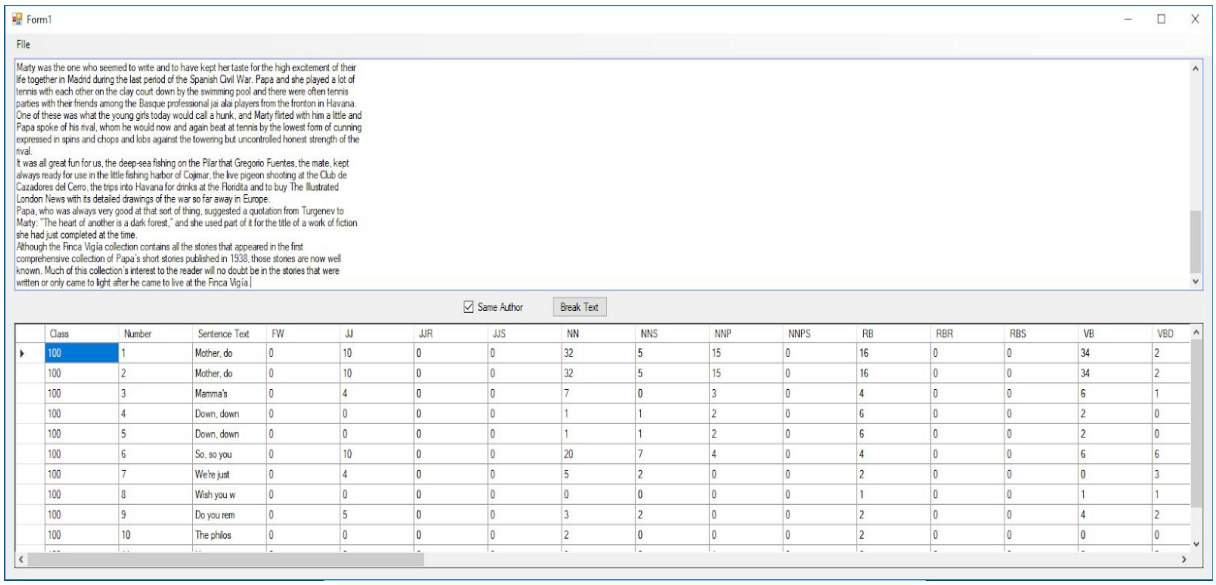

Рисунок 1 – Вікно з результатами роботи програми

Для зручності подальшої обробки, отримані результати експортуються у таблицю Excel з метою уніфікованого внесення у експертну систему. На підставі договору про співпрацю з Черкаським ДТУ результати експерименту було оброблено авторською програмою [7], що забезпечує навчання з вчителем (фактично – класифікацію множини чисельних значень у рядках таблиці Excel) на основі методу групового обліку аргументів.

В наведеному контрольному прикладі уривки з творів Ернеста Хемінгуея відрізняються від розглянутих інших за 2-ма головними класифікаційними ознаками:

- FW – іноземними словами, з вагою 33,33 %;
- QP – складними конструкціями, з вагою 66,67 %.

Висновки

У роботі проаналізовано поняття та структуру парсингу, зокрема запропоновано застосувати модель парсингу з використання POS тегів з метою визначення та узагальнення чисельних характеристик словосполучень англомовних текстових файлів. Внаслідок дослідження запропоновано новий метод атрибуції англомовного тексту, який, на відміну від існуючих, базується на лінгвістичній моделі визначення зв'язків між лексемами та іншими чисельними ознаками словосполучень у реченнях тексту та застосуванні методів машинного навчання за новими формальними ознаками множини речень тексту. Для програмної реалізації методу було обрано технологію – мова програмування C# та фреймворк Windows Forms.

Результати експериментальної апробації показали позитивний результат для атрибуції творів Е. Хемінгуея за 2-ма класифікаційними ознаками, у подальшому автори планують провести більш масштабні експерименти з корпусами англомовних текстів. Окрім того, отримані алгоритми було впроваджено для реалізації окремих задач системи інтеграції електронних ресурсів вищого навчального закладу "Інтегровані електронні ресурси ВНТУ JetIQ" [8].

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Бісікало О.В. Проектування процесів дистанційного навчання на основі формалізації пізнавальної діяльності людини / О.В. Бісікало // Інформаційні технології та комп'ютерна інженерія. – 2005. – № 3. – С. 274–280.

2. Natural language processing [Електронний ресурс]. – 2020. – Режим доступу до ресурсу: https://en.wikipedia.org/wiki/Natural_language_processing.

3. Parsing Basics [Електронний ресурс]. – 2018. – Режим доступу до ресурсу: https://www.ipipe.ru/info/parsing.html. – Electronics and Computer Science of UK, 2002. – 221 p.

4. Psycholinguistics [Електронний ресурс]. – 2018. – Режим доступу до ресурсу: https://en.wikipedia.org/wiki/Parsing. – A Probabilistic Model of Lexical and Syntactic Access and Disambiguation, 2004. — 12 p.

5. Бісікало О.В. Класифікація образного пошуку та моделювання інсайту / О.В. Бісікало // Вісник СумДУ (Серія "Технічні науки"). – 2008. – № 2. – С. 53–59.

6. Bisikalo O. Modeling the phenomenological concepts for figurative processing of natural-language constructions / Oleg Bisikalo, Yuriy Ivanov, Vladyslava Sholota // Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Systems (COLINS-2019). Volume I: Main Conference. – Kharkiv, Ukraine, April 18-19, 2019. – Pp. 1-11.

7. Голуб С. В. Формування показників масиву вхідних даних для ідентифікації авторства текстових повідомлень / С. В. Голуб, О. В. Константиновська, М. С. Голуб // Системи обробки інформації. - 2014. - Вип. 2. - С. 89-92. - Режим доступу: http://nbuv.gov.ua/UJRN/soi_2014_2_21.

8. Грабко В. В. Комп'ютерна програма "Система інтеграції електронних ресурсів вищого навчального закладу "Інтегровані електронні ресурси ВНТУ JetIQ" ("ІЕР ВНТУ JetIQ") / В. В. Грабко, О. Н. Романюк, О. В. Бісікало, М. П. Боцула, Є. А. Паламарчук, О. О. Коваленко // Свідоцтво про реєстрацію авторського права на твір № 72970. – К. : Департамент інтелектуальної власності міністерства економічного розвитку і торгівлі України. – Дата реєстрації : 20.07.2017 р.

Копецький Ярослав Едуардович – студент групи І-15б, факультет комп'ютерних систем та автоматики, Вінницький національний технічний університет, Вінниця, e-mail: kopetsky.yaroslav@gmail.com.

Бісікало Олег Володимирович – д-р техн. наук, професор, декан факультету КСА, Вінницький національний технічний університет, м. Вінниця, e-mail: obisikalo@vntu.edu.ua.

Голуб Сергій Васильович – д.т.н., проф., професор кафедри програмного забезпечення автоматизованих систем Черкаського державного технологічного університету, м. Черкаси, e-mail: s.holub@chdtu.edu.ua.

Севастьянов Володимир Миколайович – к.т.н., доцент кафедри автоматизації та інтелектуальних інформаційних технологій, Вінницький національний технічний університет, м. Вінниця, e-mail: radainaeksu@gmail.com.

Kopetsky Yaroslav E. – student, Faculty of Computer Systems and Automation, Vinnytsia National Technical University, Vinnytsia, e-mail: kopetsky.yaroslav@gmail.com.

Bisikalo Oleh V. – Dr.Sc. (Eng.), Professor, Dean of the Faculty for Computer Systems and Automatic, Vinnytsia National Technical University, Vinnytsia, email: obisikalo@vntu.edu.ua.

Golub Sergiy Vasilovich - Doctor of Technical Sciences, prof., Professor, Department of Automated System Software, Cherkasy State Technological University, Cherkasy, e-mail: s.holub@chdtu.edu.ua.

Sevastyanov Volodymyr M. – Candidate of Technical Sciences, Associate Professor, Department of Automation and Intelligent Information Technologies, Vinnytsia National Technical University, Vinnytsia, e-mail: radainaeksu@gmail.com.