УДК 004.4

Н. Ю. Олійник
О. К. Колесницький

ІНТЕЛЕКТУАЛЬНИЙ МОДУЛЬ ДЛЯ ПАРСИНГУ РЕЗЮМЕ

[1]Вінницький національний технічний університет

Анотація

У роботі проаналізовано актуальність розробки інтелектуального модуля для парсингу резюме. В даній роботі описується процес класифікації та використання технології NER для класифікації тексту на поіменовані сутності. На основі проведеного аналізу створено UML-діаграму послідовностей інтелектуального модуля.

Ключові слова: інтелектуальний модуль, поіменовані сутності, парсинг.

Abstract

The relevance of the development of an intelligent module for resume parsing is analyzed in the paper. This paper describes the process of classification and use of NER technology to classify text into named entities. Based on the analysis, a UML diagram of intelligent module sequences was created.

Keywords: intellectual module, named entities, parsing.

Вступ

Для вилучення фактів із резюме були запропоновані контрольовані та засновані на правилах методи, але вони сильно покладаються на інформацію про ієрархічну структуру та великі обсяги маркованих даних, які насправді важко зібрати. Це зменшує відсоток успішності рекомендування рекрутів, які відповідають більшості вимог роботодавця і займають занадто багато часу людських ресурсів для підбору роботи.

Мета дослідження – підвищення результативності інтелектуального парсингу резюме за рахунок використання глибокої нейронної мережі на основі технології Deep Learning та класифікаційного алгоритму Support Vector Machine (SVM). На основі проведеного аналізу необхідно створити UML-діаграму послідовностей інтелектуального модуля.

Основна частина

Синтаксичний аналізатор резюме — програмний продукт, який перетворює неструктуровані дані в структуровану форму. Це компонент, який автоматично розділяє інформацію за різними областями та параметрами, як-от контактна інформація, освітня кваліфікація, досвід роботи, навички, досягнення, професійні сертифікати, щоб швидко допомогти вам визначити найбільш релевантні резюме на основі ваших критеріїв. Класифікація тексту – це техніка машинного навчання, яка призначає відкритому тексту набір заздалегідь визначених категорій. Класифікатори тексту можна використовувати для організації, структурування та категоризації майже будь-якого тексту. Класифікація тексту є одним із основних завдань обробки природної мови з широкими застосуваннями, такими як аналіз настроїв, маркування тем, виявлення спаму та виявлення намірів [1].

Деякі з основних причин використання класифікації текстів машинного навчання:

Масштабованість – машинне навчання може автоматично аналізувати мільйони опитувань, коментарів, електронних листів тощо за незначну вартість, часто всього за кілька хвилин. Інструменти класифікації тексту можна масштабувати до будь-яких бізнес-потреб, великих чи малих.

Послідовні критерії – анотатори роблять помилки під час класифікації текстових даних, а людська суб’єктивність створює неузгоджені критерії. Машинне навчання, з іншого боку, застосовує одну й ту саму стратегію та критерії до всіх даних і результатів. Як тільки модель класифікації тексту правильно навчена, вона працює з неперевершеною точністю.

Існує багато підходів до автоматичної класифікації тексту, але всі вони підпадають під три типи систем:

Системи, засновані на правилах – такі підходи класифікують текст на організовані групи за допомогою набору лінгвістичних правил, створених вручну. Ці правила наказують системі використовувати семантично релевантні елементи тексту для визначення відповідних категорій на

основі його змісту. Але у такого підходу є деякі недоліки. По-перше, ці системи вимагають глибокого знання предметної області. Вони також займають багато часу, оскільки генерування правил для складної системи може бути досить складним і зазвичай вимагає багато аналізу. Системи на основі правил також важко підтримувати і погано масштабувати, оскільки додавання нових правил може вплинути на результати вже існуючих правил.

Системи на основі машинного навчання – класифікація текстів машинного навчання навчається робити класифікації на основі минулих спостережень. Використовуючи попередньо позначені приклади як навчальні дані, алгоритми машинного навчання можуть дізнатися про різні асоціації між фрагментами тексту та про те, що певний вихід (тобто теги) очікується для певного введення (тобто тексту). Тег — це заздалегідь визначена класифікація або категорія, до якої може потрапити будь-який текст. Першим кроком до навчання класифікатора NLP з машинного навчання є виділення ознак: метод використовується для перетворення кожного тексту в числове представлення у вигляді вектора. Одним з найбільш часто використовуваних підходів є пакет слів, де вектор представляє частоту слова у попередньо визначеному словнику [3]. Класифікація тексту за допомогою машинного навчання зазвичай набагато точніша, ніж системи правил, створених людиною, особливо для складних завдань класифікації НЛП. Крім того, класифікатори з машинним навчанням легше обслуговувати, і ви завжди можете позначати нові приклади, щоб вивчати нові завдання.

Гібридні системи – поєднують базовий класифікатор, навчений машинним навчанням, із системою на основі правил, яка використовується для подальшого покращення результатів. Ці гібридні системи можна легко налаштувати, додавши спеціальні правила для тих конфліктуючих тегів, які не були правильно змодельовані базовим класифікатором.

Також буде використовуватись один із підрозділів класифікації, а саме – розпізнавання іменованих сутностей (Named-Entity Recognition). Завдання NER – виділити спани сутностей у тексті (спан – безперервний фрагмент тексту). Припустимо, є текст новин, і ми хочемо виділити в ньому сутності (деякий заздалегідь зафіксований набір — наприклад, персони, локації, організації, дати тощо). Що таке іменовані сутності? У першій, класичній постановці, яка була сформульована на конференції MUC-6 у 1995 році, це персони, локації та організації. З того часу з'явилося кілька доступних корпусів, у кожному з яких свій набір іменованих сутностей. Зазвичай до персон, локацій та організацій додаються нові типи сутностей. Без розв'язання задачі NER важко уявити собі розв'язання багатьох задач NLP, припустимо, розв’язання займенникової анафори або побудова запитально-відповідних систем. Займенникова анафора дозволяє зрозуміти, якого елементу тексту належить займенник [2].

Вхідними даними до задачі розпізнавання іменованих сутностей є набір текстів із сутностями, що налічує більше 685000 екземплярів [4]. Екземпляром є словник, формату ключ-значення, що містить мітку одного з 6 класів. А також для задачі класифікації «скілів» буде використовуватись окремий набір даних , що налічує більше 300 екземплярів, Екземпляром є словник, формату ключ-значення, що містить мітку «скіл» або «не скіл».

UML діаграма послідовностей розроблюваного інтелектуального модуля представлена на рис. 1.

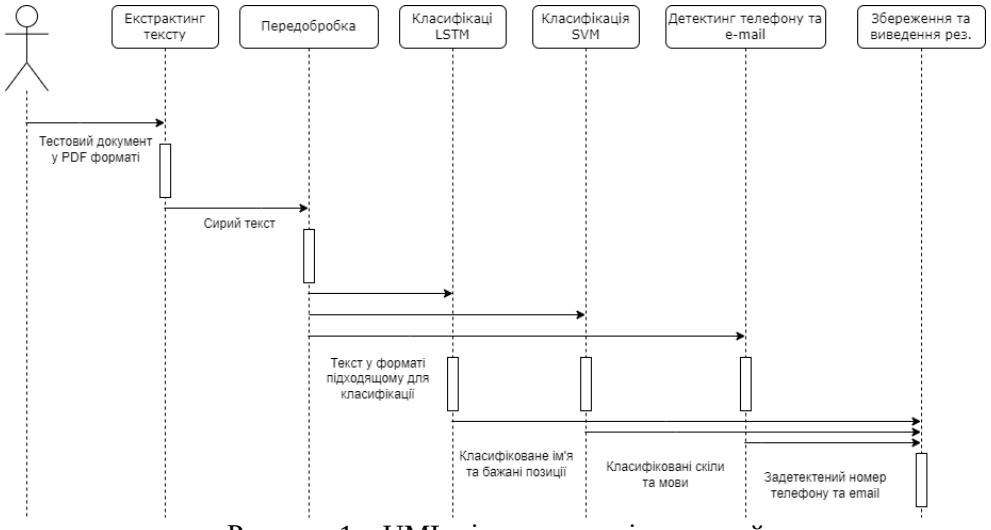

Рисунок 1 – UML діаграма послідовностей

Висновки

У ході проведеного аналізу було доведено актуальність створення інтелектуального модуля для парсингу резюме. У результаті аналізу визначено адгоритми класифікації та обґрунтовано доцільність інтелектуального парсингу резюме. На основі проведеного аналізу було створено UML-діаграма класів послідовностей модуля.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. The Art of Software Testing / Glenford J. Myers, Revised and Updated by Tom Badgett, Todd M.Thomas, Corey Sandler. - 2nd ed. - Hoboken, New Jersey.: John Wiley & Sons, Inc., 2004 - 234 p..
2. [Електронний ресурс]. Режим доступу: https://en.wikipedia.org/wiki/Named-entity_recognition.
3. Computer Vision Technology for Food Quality Evaluation (Second Edition) [Електронний ресурс]. Режим доступу: https://www.sciencedirect.com/book/9780128022320/computer-vision-technology-for-food-quality-evaluation#book-description.
4. Глибокі нейронні мережі для вирішення завдань розпізнавання і класифікації зображення [Електронний ресурс]. – Режим доступу: http://itcm.comp-sc.if.ua/2017/Sineglazov.pdf.

Олійник Нікіта Юрійович – студент групи 1КН-18б, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, м. Вінниця, email: delmark1904@gmail.com.

Колесницький Олег Костянтинович – канд. техн. наук, доцент кафедри комп'ютерних наук, Вінницький національний технічний університет, email: kolesnytskiy@vntu.edu.ua.

Oliinyk Nikita Y. – Department Intelligent Information Technology and Automation, Vinnytsia National Technical University, Vinnytsia, email: delmark1904@gmail.com.

Kolesnytskiy Oleg k. — Cand. Sc. (Eng), Assistant Professor, Vinnytsia National Technical University, Vinnytsia, email: kolesnytskiy@vntu.edu.ua.