МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ТЕРНОПІЛЬСЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
ІМЕНІ ІВАНА ПУЛЮЯ
ФАКУЛЬТЕТ КОМП'ЮТЕРНО–ІНФОРМАЦІЙНИХ СИСТЕМ
І ПРОГРАМНОЇ ІНЖЕНЕРІЇ

ВАСЬКОВ ВЛАДИСЛАВ ОЛЕКСАНДРОВИЧ

УДК 004.91

МЕТОДИ ТА ЗАСОБИ ОПРАЦЮВАННЯ ПРИРОДНОЇ МОВИ З ВИКОРИСТАННЯМ НЕЙРОННИХ МЕРЕЖ З МЕТОЮ РОЗПІЗНАВАННЯ ВЛАСНИХ НАЗВ

123 «Комп'ютерна інженерія»

Автореферат
дипломної роботи на здобуття освітнього ступеня «магістр»

Тернопіль 2019

Роботу виконано на кафедрі комп'ютерних систем та мереж Тернопільського національного технічного університету імені Івана Пулюя Міністерства освіти і науки України

Керівник роботи: доктор технічних наук, професор
Лупенко Сергій Анатолійович
Тернопільський національний технічний університет імені Івана Пулюя

Рецензент: кандидат технічних наук, доцент кафедри приладів і контрольно-вимірювальних систем
Чайковський Андрій Вікторович
Тернопільський національний технічний університет імені Івана Пулюя

Захист відбудеться 23 грудня 2019 р. о 9^{00} годині на засіданні екзаменаційної комісії №37 у Тернопільському національному технічному університеті імені Івана Пулюя за адресою: 46001, м. Тернопіль, вул. Руська, 56, навчальний корпус №1, ауд. 603

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми роботи. Ми живемо у час, коли об’єми згенерованої людством інформації є більшими, чим коли небудь й кількість таких даних збільшується з кожним днем. Однак велику користь із цих даних можна отримати лише, якщо правильно здійснити опрацювання й аналіз цих даних.

На сьогодні щосекунди по всьому світу генеруються гігабайти нових даних різного виду: створюються нові знімки, відеозаписи, записується сотні відгуків під записами у соціальних мережах й багато іншого. Й більша частина цих даних у «сирому» вигляді є практично безкорисливою. Щоб отримати з цих даних певну користь, їх потрібно відфільтрувати і опрацювати. У часи, коли технології ще не були настільки розвинуті, все це здійснювалось вручну. На це витрачалися години, дні, неділі, а й інколи місяці. А, якщо врахувати, що раніше й самої інформації для опрацювання було у рази менше, то неважко зрозуміти, що сьогодні опрацьовувати такі об’єми інформації вручну просто неможливо. Тому було розроблено велика кількість методів і засобів, які дозволяють здійснювати це з допомогою комп’ютерної техніки.

Метою роботи є обґрунтування та адаптація сучасних методів та засобів опрацювання природної мови в комп’ютерних системах для побудови програмної системи розпізнавання власних назв.

Задачі, які потрібно вирішити у ході даної магістерської роботи полягають у наступному:

– дослідити сучасні методи та засоби опрацювання природної мови з метою розпізнавання власних назв та встановити їх переваги та недоліки;

– проаналізувати методи векторного представлення тексту, що використовуються для задач опрацювання природної мови з метою розпізнавання власних назв;

– обґрунтувати математичне забезпечення програмної системи для розпізнавання власних назв з використанням методів та сучасних засобів рекурентних нейронних мереж;

– проаналізувати сучасні засоби розробки програмних систем для розпізнавання власних назв з використанням нейронних мереж;

– реалізувати програмне забезпечення з метою розпізнавання власних назв;

– здійснити оцінку якості роботи розробленого програмного забезпечення.

Об'єктом дослідження даної магістерської роботи є процес опрацювання природної мови з метою розпізнавання власних назв.

Предметом дослідження є методи, програмні і апаратні засоби для опрацювання природної мови з метою розпізнавання власних назв.

Наукова новизна одержаних результатів:

– проведено компаративний аналіз сучасних методів та засобів опрацювання природної мов, що дало змогу обґрунтувати на основі кількісних та якісних показників найбільш ефективних із них для вирішення задачі розпізнавання власних назв;

– серед множини методів векторного представлення тексту для задач опрацювання природної мови з метою розпізнавання власних назв обґрунтовано використання та адаптацію методів рекурентних нейронних мереж, як найбільш ефективних, що становлять основу математичного забезпечення програмної системи для розпізнавання власних назв;

– з використанням методів інженерії якості програмного забезпечення, а саме із застосування критеріїв таких, як повнота, точність та F-міра здійснено валідацію програмної системи для розпізнавання власних назв.

Методи дослідження. Для виконання задач дипломної роботи використано наступні методи: теоретико-емпіричний, системного аналізу, теорії проектування нейронних мереж й також математичного та комп'ютерного моделювання.

Практичне значення отриманих результатів. Розроблено програмне забезпечення, яке дозволяє опрацьовувати природню мову з метою розпізнавання власних назв.

Апробація результатів дипломної роботи. Результати роботи апробовано на VIII Міжнародній науково-технічній конференції молодих учених та студентів «Актуальні задачі сучасних технологій» м. Тернопіль 27-28 листопада 2019 року та VII науково-технічна конференція «Інформаційні моделі, системи та технології» м. Тернопіль 11-12 грудня 2019 року.

Публікації. Лупенко С.А., Васьков В.О. Аналіз методів для задач опрацювання природної мови. VIII Міжнародна науково–технічна конференція молодих учених та студентів «Актуальні задачі сучасних технологій» 27 – 28 листопада 2019 р.: тези доп. – Тернопіль, 2019. – С.60.

Васьков В.О., Лупенко С.А. Переваги використанням тензорного процесора для роботи з нейронними мережами. VII науково–технічна конференція «Інформаційні моделі, системи та технології» 11 – 12 грудня 2019 р.: тези доп. – Тернопіль, 2019. – С.27.

Структура роботи. Робота складається з пояснювальної записки та графічної частини. Пояснювальна записка складається із вступу, шести розділів, висновків, списку використаних джерел та додатку. Обсяг роботи: пояснювальна записка – 110 аркушів формату А4, графічна частина – 10 аркушів формату А1.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У вступі обґрунтовано актуальність дослідження, мету роботи, задачі, об’єкт, предмет, наукову новизну, практичне значення та публікації дипломних досліджень.

У першому розділі дипломної роботи «Дослідження предметної області опрацювання природної мови з використанням нейронних мереж» здійснено огляд предметної області опрацювання природної мови для задач розпізнавання власних назв. Досліджено особливості опрацювання природної мови. Проаналізовано основні компоненти природної мови, а також проаналізовано методи нормалізації корпусу тексту в NLP.

Розглянуто методи, що використовуються для процесу навчання NLP. Досліджено принципи побудови нейронних мереж. Наведено загальні відомості про нейронні мережі й розглянуто основні функції нейронів.

На основі рекурентної нейронної мережі обґрунтовано математичне забезпечення побудови програмної системи для задач розпізнавання власних назв.

У другому розділі «Аналіз і дослідження методів та засобів опрацювання природної мови з використанням нейронних мереж» досліджено ефективні і дієві методи для представлення слів у вигляді векторів Word2Vec та Glove. Встановлено переваги та недоліки кожного з проаналізованих методів. Розглянуто два основних алгоритми навчання слів – CBOW і Skip-Gram. Також досліджено математичне забезпечення методів Word2Vec та Glove.

Здійснено аналіз апаратних засобів опрацювання природної мови з використанням нейронних мереж. Досліджено ефективність використання тензорних процесорів, в порівнянні із традиційними серверними CPU і GPU. Також, досліджено структуру тензорного процесора, який призначений для задач навчання нейронних мереж.

У третьому розділі «Реалізація програмного забезпечення для опрацювання природної мови з метою розпізнавання власних назв» розглянуто бібліотеку TensorFlow, що використовується для прискорення машинного навчання і дослідження нейронних мереж й її супутню бібліотеку TensorFlow.js, яка призначена для навчання та використання моделей машинного навчання в браузері. Розглянуто високорівневу API Keras, що здатна працювати поверх TensorFlow.

Здійснено опис навчання та функціонування розробленого програмного забезпечення для опрацювання природної мови з метою розпізнавання власних назв. Також, розглянуто принцип навчання нейронної мережі й наведено результати роботи реалізованої програми. Здійснено оцінку якості розробленого програмного забезпечення на основі критеріїв таких, як точність, повнота і F-міра.

У четвертому розділі «Обґрунтування економічної ефективності» проведено обґрунтування, яке надає загальну закінченість дипломного проекту, дозволяє підвищити рівень сприйняття проблеми, зв'язати воєдино технічні та економічні аспекти розв'язуваної задачі і з цієї позиції оцінити проведену роботу в комплексі, і в результаті добитися максимальної повноти і чіткості техніко-економічного опрацювання проекту та в кінцевому підсумку підвищити якість.

П'ятий розділ роботи «Охорона праці та безпека в надзвичайних ситуаціях». У даному розділі проведено аналіз норм праці, шкідливих та небезпечних чинників при дослідження методів та засобів опрацювання природної мови з використанням нейронних мереж з метою розпізнавання власних назв, а також описано параметри і характеристики приміщення, заходи, які були виконані для забезпечення належних умов роботи.

Розглянуто особливості радіаційних аварій, які можуть виникати на радіаційно небезпечних підприємствах, що у своїй виробничій діяльності використовують джерела іонізуючого випромінювання. Проаналізовано особливості робіт пов'язаних з виготовленням, зберіганням, використанням, транспортуванням та похованням об'єктів – джерел іонізуючого випромінювання. Також розглянуто умови і фактори виникнення пожеж та суцільних пожеж. Описано комплекс профілактичних заходів, що спрямований на попередження та (або) усунення чинників виникнення пожеж та суцільних пожеж.

Шостий розділ роботи «Екологія». В даному розділі розглянуто питання теоретичних основ екології, яке стосується застосування екологічних знань у різних галузях соціально-політичного життя, також розглянуто вимоги до моніторів (ВДТ) і ПЕОМ.

ВИСНОВКИ

У даній магістерській роботі досліджено методи та засоби опрацювання природної мови з використанням нейронних мереж з метою розпізнавання власних назв. Основні результати та висновки проведених досліджень такі:

– досліджено сучасні методи та засоби опрацювання природної мови з метою розпізнавання власних назв та встановлено їх переваги та недоліки;

– проаналізовано методи векторного представлення тексту, що використовуються для задач опрацювання природної мови з метою розпізнавання власних назв, що дало змогу обґрунтувати математичне забезпечення програмної системи для розпізнавання власних назв з використанням методів та сучасних засобів рекурентних нейронних мереж;

– на основі попередньо обґрунтованого математичного забезпечення розроблено архітектуру програмної системи для розпізнавання власних назв;

– обґрунтовано бібліотеку TesorFlow.js та API Keras, як засобів розробки програмної системи для розпізнавання власних назв, що уможливило ефективну її реалізацію;

– здійснено оцінку якості роботи програмного забезпечення за такими параметрами, як: точність, повнота та F-міра.

Проведено економічні розрахунки, які спрямовані на визначення економічної ефективності та вартості проведення дослідження.

Здійснено опис вимог з охорони праці й техніки безпеки відповідно до нормативних документів щодо: організації робочого місця, електробезпеки, шуму та вібрації, освітленості, мікроклімату та пожежної безпеки.

Розглянуто питання екології, що стосуються магістерської роботи.

СПИСОК ОПУБЛІКОВАНИХ АВТОРОМ ПРАЦЬ ЗА ТЕМОЮ РОБОТИ

1. Лупенко С.А., Васьков В.О. Аналіз методів для задач опрацювання природної мови. VIII Міжнародна науково–технічна конференція молодих учених та студентів «Актуальні задачі сучасних технологій» 27 – 28 листопада 2019 р.: тези доп. – Тернопіль, 2019. – С.60.

2. Васьков В.О., Лупенко С.А. Переваги використанням тензорного процесора для роботи з нейронними мережами. VII науково–технічна конференція «Інформаційні моделі, системи та технології» 11 – 12 грудня 2019 р.: тези доп. – Тернопіль, 2019. – С.27.

АНОТАЦІЯ

Васько В. О. Методи та засоби опрацювання природної мови з використанням нейронних мереж з метою розпізнавання власних назв.

Дипломна робота на здобуття освітнього ступеня магістра 123 – Комп'ютерні системи та мережі. – Тернопільський національний технічний університет імені Івана Пулюя 2019.

Мета роботи полягає у обґрунтуванні та адаптації сучасних методів та засобів опрацювання природної мови в комп'ютерних системах для побудови програмної системи розпізнавання власних назв.

У дипломній роботі досліджено сучасні методи та засоби опрацювання природної мови з метою розпізнавання власних назв та встановлено їх переваги та недоліки.

Проаналізовано методи векторного представлення тексту, що використовуються для задач опрацювання природної мови з метою розпізнавання власних назв, що дало змогу обґрунтувати математичне забезпечення програмної системи для розпізнавання власних назв з використанням методів та сучасних засобів рекурентних нейронних мереж.

На основі попередньо обґрунтованого математичного забезпечення розроблено архітектуру програмної системи для розпізнавання власних назв.

Обґрунтовано бібліотеку TesorFlow.js та API Keras, як засобів розробки програмної системи для розпізнавання власних назв, що уможливило ефективну її реалізацію.

Здійснено оцінку якості роботи програмного забезпечення за такими параметрами, як: точність, повнота та F-міра.

Ключові слова: розпізнавання власних назв, опрацювання природної мови, нейронна мережа, word2vec, glove, тензорний процесор.

ANNOTATION

Vaskov V. O. Methods and tools of natural language processing due to neural networks use aimed at proper names recognition.

The diploma paper for obtaining the Master's degree 123 – Computer systems and network – Ternopil Ivan Puluj National Technical University 2019.

The purpose of the work is to substantiate and adapt modern methods and means of natural language processing in computer systems to build a software system for proper name recognition.

In the thesis investigates modern methods and means of natural language processing for the purpose proper names recognition and their advantages and disadvantages.

The methods of vector representation of text used for natural language processing tasks for the purpose of recognition of proper names are analyzed, which made it possible to substantiate the mathematical software of the system for recognition of proper names using methods and modern means of recurrent neural networks.

Based on pre-grounded mathematical software, a software system architecture was developed for recognizing proper names.

The TesorFlow.js library and API Keras have been substantiated as a means of developing a software system for identifying their own names, which has made it possible to implement it effectively.

The quality of the software was evaluated using such parameters as precision, recall and F-measure.

Keywords: proper name recognition, natural language processing, neural network, word2vec, glove, tensor processing unit.