УДК 681.12

В. Ю. Бондарчук
І. Р. Арсенюк

ОБРОБКА ТА АНАЛІЗ ПРИРОДНОЇ МОВИ МЕТОДАМИ МАШИННОГО НАВЧАННЯ

Вінницький національний технічний університет

Анотація

Розглянуто основні проблеми обробки природної мови. Проаналізовано основні напрямки обробки, зокрема нейронні мережі з пам'яттю, керовані рекурентні, Tree-LSTM.

Ключові слова: обробка природної мови, аналіз тексту, обробка тексту, аналіз тональності тексту, класифікація, нейронна мережа, збір даних.

Abstract

The main problems of the processing of natural materials are examined. The main strands of sample processing, neuronal measure of memory, memory recurrence, Tree-LSTM were analyzed.

Keywords: Natural Language Processing, text analysis, text processing, sentiment analysis, classification, neural network, data mining.

Лінгвістична обробка природномовних текстів є досить значущим напрямком розвитку інтелектуальних інформаційних технологій. Цій проблемі приділяється значна увага в усьому світі. Зокрема, виділяється значне фінансування на розробку лінгвістичного програмного забезпечення. У зв'язку зі стрімким розвитком Інтернету та ряду інших комп'ютерних технологій лінгвістична обробка природномовних текстів набуває ще більшої популярності [1].

Обробка природної мови (Natural Language Processing – NLP) – міждисциплінарна галузь, що перетинається з комп'ютерними науками, штучним інтелектом та обчислювальною лінгвістикою [2]. Основою NLP є забезпечення взаємодії між людськими мовами та комп'ютером.

Інтелектуальний аналіз тексту (Text mining) – напрям інтелектуального аналізу даних та штучного інтелекту. Його задача – отримання якісної інформації з текстових документів за допомогою застосування методів машинного навчання та обробки природної мови. Основна задача Text mining полягає у виявленні інформації, що може бути прихованою у контексті іншої інформації. Така задача може бути розв'язана із застосуванням методологій аналізу та обробки природної мови [3]. В основі цього лежать, зокрема Tree-LSTM, керовані рекурентні нейронні мережі, нейронні мережі з пам`яттю.

Нейронні мережі з пам'яттю (Memory Networks) можна використовувати з метою отримання відповідей на питання. Дані мережі використовують асоціативну пам'ять для читання та запису. Таку пам'ять не використовують ні CNN, ні Q-Network (для навчання з підкріпленням (reinforcement learning), ні традиційні нейронні мережі. Зокрема, це пов'язано з тим, що завдання формування відповідей на питання в основному покладається на здатність моделювати чи простежувати віддалені залежності, наприклад, запам'ятовувати послідовність подій. У мереж CNN або QNetworks пам'ять вбудована у ваги системи, оскільки навчається від різних фільтрів або за допомогою карток відповідностей станів і дій [5]. Слід зазначити, що LSTM та RNN не здатні запам'ятовувати вхідні дані з минулого, а це означає, що вони не підходять для задач формування відповідей на питання [4].

Керовані рекурентні нейрони (Gated recurrent units, GRU). За допомогою GRU здійснюється обчислення векторів прихованих станів у рекурентній нейронній мережі (Recurrent Neural Networks, RNN), що дозволяє зберігати інформацію про віддаленіші залежності. Під час роботи методу зворотного поширення помилки (back propagation) помилка пересуватиметься по RNN у зворотному порядку від останнього до першого кроку [5]. При досить малому початковому градієнті до третього або четвертого модуля градієнт майже зникне, і тоді приховані стани перших кроків не оновляться.

Найпоширенішим інструментом для задач розпізнавання емоційного забарвлення є Tree-LSTM. Підхід щодо нелінійного розташування компонентів заснований на тезі, що природні мови мають

властивість перетворювати у фрази послідовності слів [6]. Ці фрази, залежно від порядку слів, можуть мати значення, звідки вони були видалені, значення входять до цих компонентів. Щоб відобразити цю властивість, мережа з декількох LSTM-нейронів має бути подана у вигляді дерева, де на кожен нейрон впливають його дочірні вузли. Одна з відмінностей Tree-LSTM від звичайного LSTM полягає в тому, що в останньому прихований стан – функція від поточних вхідних даних та прихованого стану на попередньому кроці. У Tree-LSTM прихований стан – функція від поточних вхідних даних і прихованих станів його дочірніх нейронів.

Висновки

Розглянуто основні методи обробки текстових даних. На даний момент існує досить велика кількість різноманітних методів обробки природних мов: нейронні мережі з пам'яттю, керовані рекурентні нейрони, Tree-LSTM та інші. Ряд методів обробки текстової інформації можуть використовуватись для більшості задач NLP, але існують такі методи, які застосовуються тільки для певних класів задач. Для покращення якості інтелектуального аналізу тестів пропонується комбінувати методи, шо проаналізовано вище.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Барсегян А. А. Анализ данных и процессов: учеб. пособие / А. А. Барсегян, М. С. Куприянов, И. И. Холод, М. Д. Тесс, С. И. Елизаров. – 3-е изд., перераб. и доп. Санкт-Петербург: БХВ-Петербург, 2009. – 512с.
2. Yang Y. A re-examination of text categorization methods / Y. Yang, X. Liu // Proc. of Int. ACM Conference on Research and Development in Information Retrieval (SIGIR-99), 1999. – P. 42 – 49.
3. Вагин В. Н. Достоверный и правдоподобный вывод в интеллектуальных системах / В. Н. Вагин, Е. Ю. Головина, А. А. Загорянская, М. В. Фомина. – Москва: Физматлит, 2004. – 704 с.
4. Quinlan J. R. C4.5 Programs for machine learning. – Morgan Kaufmann, – San Mateo, Californie, 1993.
5. Айвазян С. А. Прикладная статистика: классификация и снижение размерности / С. А. Айвазян, В. М. Бухштабер, И. С. Енюков, Л. Д. Мешалкин. – Москва: Финансы и статистика, 1989.
6. Joachims T. Making large-scale SVM learning practical / T. Joachims // Advances in Kernel Methods Support Vector Learning. – MIT Press, 1999. – 218 p.
7. Бондарчук В. Ю. Порівняння методів аналізу тональності тексту. Бондарчук В. Ю., Арсенюк І. Р. Матеріали XLVIII науково-технічної конференції підрозділів ВНТУ, Вінниця, 13-15 березня 2019 р.

Бондарчук Віталій Юрійович – студент гр. 2КН-19м факультет інформаційних технологій та комп'ютерної інженерії, Вінницький національний технічний університет, м. Вінниця, e-mail: 2kn15b.bondarchuk@gmail.com

Арсенюк Ігор Ростиславович – к. т. н., доцент, доцент кафедри комп'ютерних наук, Вінницький національний технічний університет.

Vitalii Y. Bondarchuk – Student of Department of Information Technology and Computer Engineering, Vinnytsia National Technical University, Vinnytsia, e-mail: 2kn15b.bondarchuk@gmail.com

Igor R. Arsenyuk – Cand. Sc. (Eng), Assistant Professor of the Computer Science Chair, Vinnytsia National Technical University, Vinnytsia.