Ткачик Денис Анатолійович
аспірант кафедри автоматизації та інтелектуальних інформаційних технологій
Вінницький національний технічний університет, Україна

СУЧАСНІ МЕТОДИ ОБРОБКИ ПРИРОДНОЇ МОВИ В КОНТЕКСТІ АНАЛІЗУ ЕМОЦІЙНОГО ЗАБАРВЛЕННЯ ТЕКСТУ

Вступ

Зараз людство проживає в час, коли гігабайти данних генеруються щосекунди, заповнюючи собою ввесь інформаційний простір. Сюди входять тисячі відеозаписів, мільйони фотографій, безліч коментарів на форумах, відгуки під товарами на різних торгових платформах. Після успіху таких соціальних мереж як Facebook та Twitter люди обмінюються щодня мільярдами текстових повідомлень, більшість засобів масової інформації перейшли в онлайн. Завдяки цьому і з розвитком сучасних інформаційних технологій з усіх цих даних можна отримати користь, зокрема можна аналізувати новини для передбачення фінансових ринків, переглядати твіти найбагатших людей аби зрозуміти їхні подальші дії, акумулювати інформацію за однією темою для подальшої обробки, проводити сентиментальний аналіз для розуміння настроїв тих чи інших текстових даних. Саме для цього було розроблено безліч підходів та методів обробки природної мови, що буде розглянуто в даній статті.

Глибинне навчання

Істотна частина технологій обробки природної мови працює завдяки глибинному навчанню – галузі машинного навчання, що ґрунтується на наборі алгоритмів, які намагаються моделювати високорівневі абстракції в даних, застосовуючи глибинний граф із декількома обробними шарами, що побудовано з кількох лінійних або нелінійних перетворень. Доцільність використання глибинного навчання зумовлюється наступними факторами:

• накопичено великі обсяги тренувальних даних;

• розроблено обчислювальні потужності: багатоядерні CPU і GPU;

• створено нові моделі і алгоритми з розширеними можливостями і поліпшеною продуктивністю, з гнучким навчанням на проміжних уявленнях; з'явилися навчальні методи з використанням контексту, нові методи регуляризації та оптимізації [1].

Успішне використання більшості методів машинного навчання було досягнуто завдяки наявності репрезентативних даних, вхідних ознак, та оптимізації ваг, для підвищення точності фінального передбачення Глибинне навчання пропонує гнучкий, універсальний та фреймворк, що може навчатись, для подання інформації за допомогою візульного та лінгвістичного представлень.

Векторне представлення

У традиційних методах обробки природної мови слова розглядаються як дискретні символи, які в подальшому представлені у вигляді одноразових векторів. Недолік такого способу представлення слів полягає у відсутності визначення схожості для одноразових векторів. Вирішити дану проблему можна за допомогою кодування схожості у самих векторах [2].

Векторне представлення – це метод представлення строк у вигляді векторів зі значеннями. Створюється плоский (щільний) вектор для кожного слова так, щоб слова у схожих контекстах мали подібні вектори. Такий спосіб представлення враховує стартову точку для більшості завдань та робить глибинне навчання ефективним на малих датасетах. Техніки векторних представлень Word2vec і GloVe, створених Google, користуються популярністю та часто використовуються для задач обробки природної мови

Word2vec — це одна з методик обробки природної мови. Алгоритм word2vec використовує нейромережну модель для навчання пов'язаностей слів із великого корпусу тексту. Щойно її натреновано, така модель може виявляти слова-синоніми, або підказувати додаткові слова для часткового речення. Як випливає з її назви, word2vec представляє кожне окреме слово певним переліком чисел, званим вектором. Ці вектори ретельно підбираються таким чином, щоби проста математична функція (косинусна подібність векторів) вказувала на рівень семантичної подібності між словами, представленими цими векторами.

GloVe, від англ. Global Vectors (глобальні вектори), — це одна з моделей для розподіленого представлення слів. Ця модель є алгоритмом спонтанного навчання для отримування векторних представлень слів. Це досягається відображенням слів до змістовного простору, в якому відстань між словами пов'язано з семантичною подібністю. Тренування виконується на аґреґованій глобальній статистиці попарної спільної появи слів корпусу, а отримувані в результаті представлення демонструють цікаві лінійні підструктури векторного простору слів. Її розроблюють як відкритий проєкт у Стенфорді. Як логарифмічно-білінійна регресійна модель для спонтанного навчання представлень слів, вона поєднує властивості двох сімейств моделей, а саме, методів глобального розкладу матриць, та локального контекстного вікна.

Машинний переклад

Машинний переклад - перетворення тексту однієї природної мови в еквівалентний за змістом текст на іншій мові. Машинний переклад виконується програмою або машиною без участі людини. У машинному перекладі використовується статистика використання слів по сусідству

У традиційних системах машинного перекладу доводиться використовувати паралельний корпус - набір текстів, кожен з яких перекладено на один або декілька інших мов. Наприклад, маючи вихідних мову А (Французький) і цільову Б (Англійська), потрібно побудувати статистичну модель, що включає вірогіднісне формулювання для правила Байєса, модель перекладу, навчену на паралельному корпусі, і модель мови, навчену тільки на корпусі з англійською мовою [3].

Даний підхід пропускає сотні важливих деталей, вимагає великої кількості спроектованих вручну ознак, складається з різних і незалежних завдань машинного навчання

Висновки.

Обробка природної мови наразі є актуальним завданння в сфері інформаційних технологій. Комбінуючи різні, описані вище, підходи до аналізу текстів можна досягнути значних результатів в дослідженні текстових даних для їх подальшого застосування в різних інформаційних сферах. Зокрема можна виокремити наступні практичні застосування

- ідентифікація різних груп користувачів чи клієнтів, що використовують певний продукт;
- точна детекція та пошук різних категорій відгуків (позитивні та негативні думки, згадки окремих тегів.);
- класифікація тексту відповідно до його змісту (запит елементарної допомоги, термінова проблема).

Розглянуті методи в цій роботі можуть бути розширені, для більш ефективного моделювання мови, образів та знань для більш ширшого та практичного застосування.

Список використаних джерел:

1. J. Le. (2018). ―The 7 NLP Techniques That Will Change How You Communicate in the Future (Part I)‖ [Online]. Available: https://heartbeat.fritz.ai/the-7-nlp-techniques-that-will-change-how-youcommunicate-in-the-future-part-i-f0114b2f0497
2. R. Socher, ― Recursive deep learning for natural language processing and computer vision‖ Stanford University, 2014, pp. 8-120.
3. Iroju, O. G., & Olaleke, J. O. (2015, July 8). A Systematic Review of Natural Language Processing in Healthcare. International Journal of Information Technology and Computer Science, 7(8), 44–50. https://doi.org/10.5815/ijitcs.2015.08.07