УДК 004.048

А.Ю. Барановська
М.А. Лешок
О.В. Бісікало

ПІДХОДИ ТА МЕТОДИ ЗАСТОСУВАННЯ ТЕХНОЛОГІЙ ШТУЧНОГО ІНТЕЛЕКТУ ДЛЯ РОЗВ'ЯЗАННЯ ЗАДАЧ ОБРОБКИ ТЕКСТОВОЇ ІНФОРМАЦІЇ

Вінницький національний технічний університет

Анотація

У статті описано методи і підходи з галузі штучного інтелекту, які використовуються для розв'язання задач обробки текстової інформації. Окреслені перспективні напрямки розвитку галузі, виділені переваги та недоліки розглянутих методів у актуальних задачах NLP.

Ключові слова:

Штучний інтелект; обробка природної мови; математична лінгвістика; аналіз тексту; машинне навчання; глибинне навчання; інструменти машинного навчання.

Abstract

This article describes methods and approaches that are designed to solve text information processing (NLP) problems. The directions of development of the branch, prospects of advantages and disadvantages of the considered methods in the tasks of NLP, translation of such text are outlined.

Keywords:

Artificial intelligence; natural language processing; mathematical linguistics; text analysis; machine learning; deep learning; machine learning tools.

Вступ

Протягом останніх десятиліть спостерігається значний попит на вирішення задач обробки тексту. Із розвитком ботів і голосових помічників все більше постає необхідність у розумінні та інтерпретації природної мови. Коректність сприйняття тексту суттєво впливає на процес обробки введених даних та отримання результатів. Повне розуміння та відтворення сенсу мови – надзвичайно складне завдання, оскільки людська мова має цілий ряд особливостей.

Обробка природної мови (далі, від англ. абревіатури NLP – Natural Language Processing) – це здатність комп'ютерної програми розуміти природну мову, зокрема процедурно – як зазвичай на ній розмовляють та пишуть тексти цією мовою. NLP є одним із найбільш важливих компонентів штучного інтелекту, оскільки цей актуальний напрямок дослідження існує вже майже 70 років і тісно пов'язаний з предметною галуззю лінгвістики. NLP має важливі практичні застосування у різноманітних областях, включаючи, пошукові системи та бізнес-аналітику, медичні дослідження, контекстну онлайн-рекламу, автоматичний або напівавтоматичний переклад, аналіз емоційного фону у соціальних мережах, маркетинг, розпізнавання мови, чат-боти та голосові помічники.

Принципи побудови додатків NLP

NLP дозволяє комп'ютерам «розуміти» природну мову на зразок того, як це роблять люди. Незалежно від того, чи це розмовна чи письмова мова, обробка природної мови використовує методи штучного інтелекту, щоб приймати вхідні дані, обробляти їх та інтерпретувати їх у спосіб, притаманний комп'ютеру. Подібно до того, як люди мають різні рецептори – наприклад зорові або слухові, – комп'ютери мають засоби для зчитування інформації. Також подібно до мозку людини, який відповідає за обробку отриманих ззовні даних, у комп'ютерів застосовуються програми для обробки певних типів даних. Після закінчення обробки введені дані інтерпретуються у форматі, який може «зрозуміти» комп'ютер [1].

Можна виокремити два основних етапи побудови додатків NLP: попередня обробка природномовних даних і розробка алгоритму для розв'язання певної задачі комп'ютерної лінгвістики. Попередня обробка даних включає підготовку та «очищення» текстових даних для того, щоб забезпечити формальну можливість їх аналізувати. Попередня обробка представляє дані в зручному форматі, зокрема виділяє в тексті особливості, з якими може працювати алгоритм. Це можна зробити кількома способами, у тому числі:

- **Токенізація** – текст розбивається на менші одиниці для роботи;
- **Stop word removal** – службові слова видаляються з тексту, в результаті залишаються унікальні значущі слова, які надають найважливішу інформацію про текст;
- **Лематизація і стеммінг** – слова представляються у кореневих формах для подальшої обробки;
- **Позначення частини мови** – слова позначаються відповідно до частини мови, як-от іменники, дієслова та прикметники.

На основі результатів попередньої обробки даних будується алгоритм розв'язання конкретної задачі комп'ютерної лінгвістики. Існує досить багато методів і систем обробки природної мови, але зазвичай використовуються два основних типи:

- **Система на основі правил.** Ця система використовує ретельно розроблені лінгвістичні правила для кожної природної мови. Цей підхід часто використовувався на початку розвитку предметної області NLP.
- **Методи та технології машинного навчання.** Алгоритми машинного навчання, як правило, використовують статистичні методи. Вони вчаться виконувати завдання на основі навчальних даних, які отримують, а потім коригують свої моделі за результатами обробки більшої кількості даних. Використовуючи комбінацію машинного навчання, глибокого навчання та нейронних мереж, алгоритми обробки природної мови відточують власні правила (моделі) шляхом багаточисельної обробки й навчання.

NLP та глибинне навчання

Значна частина технологій NLP працює за допомогою методів глибинного навчання. Цей актуальний напрям машинного навчання базується на множині алгоритмів, які є моделями високорівневих абстракцій даних. Зокрема застосовуються глибинні графи із декількома обробними шарами, що побудовані на основі лінійних або нелінійних перетворень. Доречність використання глибинного навчання обумовлюється такими факторами:

- накопичення великих обсягів тренувальних даних;
- розробка обчислювальних потужностей, зокрема багатоядерних CPU і GPU;

- наявність нових моделей і алгоритмів із розширеними можливостями і покращеною продуктивністю, гнучким навчанням на проміжних представлення даних;
- наявність навчальних методів з використанням контексту, нові методи оптимізації та регуляризації.

Ефективність використання більшості методів машинного навчання підвищується завдяки наявності репрезентативних даних, вхідних ознак, а також оптимізації ваг для підвищення точності фінального передбачення. Глибинне навчання на даний час доступне дослідникам через цілий ряд універсальних гнучких фреймворків для подання інформації за допомогою лінгвістичного і візуального представлень.

Векторне представлення

Векторне представлення – це метод представлення текстової інформації у вигляді векторів зі значеннями. Для кожного слова створюється плоский вектор так, щоб слова у схожому контексті мали схожі вектори. Даний спосіб представлення включає стартову точку для переважної кількості задач NLP та робить глибинне навчання ефективним навіть на невеликих наборах даних. Такі техніки векторних представлень як Word2vec і GloVe, розроблені компанією Google, є популярними для завдань NLP [3].

Word2vec – являє собою спосіб побудови стисненого простору векторів слів, що використовує нейронні мережі. Він приймає на вхід великий текстовий корпус та зіставляє кожному слову вектор. Спочатку створюється словник та обчислюється векторне представлення слів. Векторне представлення ґрунтується на контекстній близькості: слова, що зустрічаються в тексті поруч із однаковими словами (рис. 1).

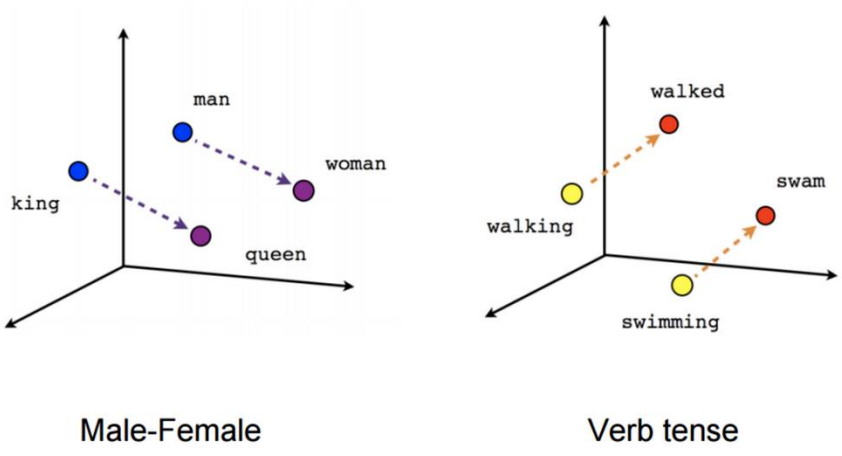

Рис. 1 – Приклад роботи Word2vec.

Щоб досягнути ліпшого результату роботи даного алгоритму, з набору даних видаляються незначущі слова (такі як артиклі та службові слова у англійській мові тощо). Така операція збільшує точність алгоритму та скорочує час тренування моделі.

У word2vec існують дві основні моделі навчання – SKIPGRAM та CBOW (англ. Continuous Bag of Words), схематично представлені на рис 2.

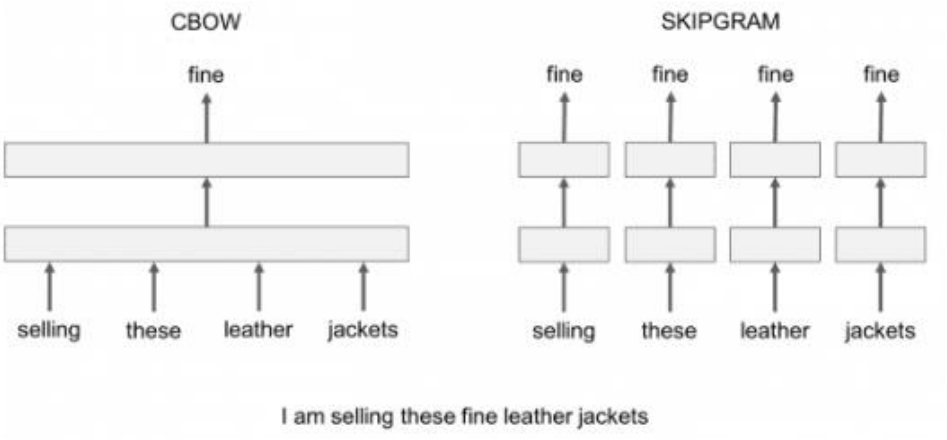

Рис. 2 – Візуальне представлення CBOW та SKIPGRAM.

У моделі SKIPGRAM за словами передбачаються слова з його контексту, а моделі CBOW по контексту підбирається найбільш ймовірне слово. На вихідному шарі використовується функція *softmax* або його варіація, щоб отримати на виході розподіл ймовірності кожного слова. В обох моделях вхідні та вихідні слова подаються в one-hot encoding (процес, за допомогою якого змінні перетворюються у форму, яка може бути надана алгоритмам машинного навчання для кращого прогнозування), завдяки чому при множенні на матрицю W, що з'єднує вхідний та прихований шари, відбувається вибір одного рядка W. Розмірність N є гіперпараметром алгоритму, а навчена матриця W – виходом, оскільки її рядки містять векторні уявлення слів.

Недоліком word2vec є те, що за його допомогою не можуть бути представлені слова, які не зустрічаються в навчальній вибірці. Інша модель fastText вирішує цю проблему за допомогою N-грамів символів. Наприклад, 3-грамами для слова яблуко є яблуко, блу, лук, уко. Модель fastText будує векторні уявлення N-грам, а векторним уявленням слова є сума векторних уявлень всіх його N-грам.

Машинний переклад

Машинний переклад за допомогою нейромереж (Neural Machine Translation, NMT) – це підхід до моделювання перекладу за допомогою рекурентної нейронної мережі (RNN). Це нейромережа із залежністю від попередніх станів, що має зв'язки між ітераціями. Нейрони одержують дані з попередніх шарів та із самих себе на попередньому етапі. Це означає, що порядок, в якому подається на вхід дані і тренується мережа, важливий (рис 3).

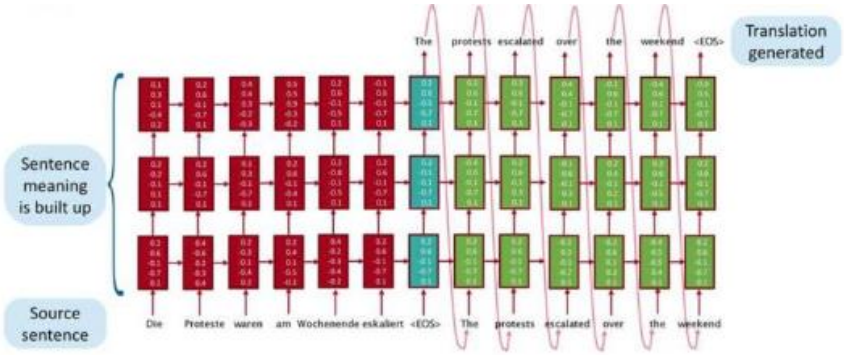

Рис. 3 – Принцип дії NMT.

Недоліки: даний підхід не враховує сотні важливих деталей, вимагає великої кількості спроектованих вручну ознак, складається з різних і незалежних одна від одної задач машинного навчання. Головна проблема алгоритму RNN – це затирання градієнту, коли інформація втрачається з часом. Як наслідок, RNN моделі будуть зазнавати труднощів у запам'ятовуванні слів, що стоять далі в послідовності, а передбачення будуть робитися на основі крайніх слів.

Нейронна машина для відповідей

Нейронна машина для відповідей (NRM – Neural Responding Machine) – це генератор відповідей для недовгих розмов, створений із використанням кодер-декодер фреймворка (рис 4).

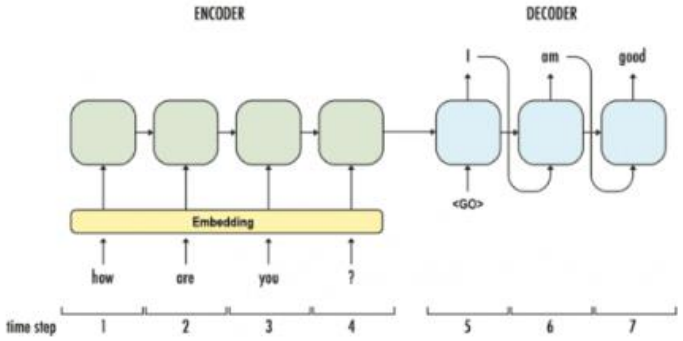

Рис 4. – Нейронна машина із використанням кодер-декодер фреймворку.

Спочатку формалізується створення відповіді, як процес розшифрування на основі прихованого представлення вхідних даних, у той час як кодування та декодування здійснюються на основі рекурентних нейронних мереж. Навчання NRM здійснюється на великих об’ємах даних із однозначними діалогами, зібраними з мікроблогів. Встановлено, що NRM генерує граматично коректні, контекстуально доречні відповіді у 75% наданих для обробки розмов, що перевищує показники інших сучасних моделей подібної конфігурації [4].

Через обмеженість можливостей штучного інтелекту у NLP, поки що розробка повноцінного розмовного асистента залишається відкритим питанням.

Системи «Question - answer»

Системи «питання-відповідь» отримують дані безпосередньо з джерела (документа, розмови, онлайн пошуку, тощо). Такі системи надають короткі і лаконічні відповіді замість розгорнутого тексту. Більшість NLP задач можна розглядати як задачі типу «питання-відповідь», де користувач надсилає запит (наприклад, через інтегрованого чат-бота) та одержує відповідь від системи.

Задля вирішення задач «питання-відповідь» існує спеціальна оптимізована архітектура глибинного навчання – Мережа Динамічної Пам'яті (Dynamic Memory Network, DNM). DNM навчається на тренувальному наборі вхідної інформації та питань і формує епізодичні «спогади» про них, які потім використовуються для генерації доречних відповідей [5].

Анотування тексту

Анотування (реферування) тексту (Text Summarization) – інструмент інтерпретації текстових даних, що дозволяє створювати лаконічні підсумки великих текстових фрагментів. Скорочення інформації формується у декілька етапів: підрахунок частоти появи слова в текстовому документі; визначення 100 найбільш частих слів; сортування визначених слів; оцінювання кожного речення із найбільш частими словами, із наданням більшого вагового коефіцієнту словам, що зустрічаються частіше; сортування перших X речень з урахуванням їхнього положення в оригінальному тексті.

Виділяють два основних підходи до скорочення тексту: видобувний та абстрактний.

- видобувний підхід добуває слова та фрази з оригінального тексту для створення стислого перекладу.
- Абстрактний підхід вивчає внутрішнє мовне представлення тексту, щоб створити подібну до людської інтерпретацію шляхом перефразування.

Висновок

Дослідження в галузі обробки природної мови спрямовані на створення програмних продуктів, які не тільки розуміють і реагують на текст або голосові дані, але й відповідають власним текстом або мовленням – майже так само, як люди. Розглянуті підходи і методи штучного інтелекту дозволяють розв'язувати актуальні та корисні задачі комп'ютерної лінгвістики.

Моделі рекурсивного глибокого навчання можуть вирішувати цілий ряд мовних завдань, що включають передбачення на рівні слова та речення як безперервного, так і дискретного характеру. Одним із багатьох можливих рішень для досягнення цієї мети є застосування рекурсивних або рекурентних методів. Інша задача для глибинних моделей – це логічні міркування першого порядку, які можуть знадобитися для отримання коректних даних з баз знань за допомогою питань природної мови.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. J. Le. (2018). ―The 7 NLP Techniques That Will Change How You Communicate in the Future (Part I). Режим доступу: https://heartbeat.fritz.ai/the-7-nlp-techniques-that-will-change-how-you-communicate-in-the-future-part-i-f0114b2f0497

2. M. Bates (1995). ― Models of natural language understanding. Режим доступу: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC40721/

3. R. Socher, ―Recursive deep learning for natural language processing and computer vision. Stanford University, 2014, pp. 8-120.

4. Afanasieva I. Data exchange model in the Internet of Things concept / I. Afanasieva, N. Golian, O. Hnatenko, Y. Daniiel, K. Onyshchenko // Telecommunications and Radio Engineering, New York, 2019. – 10(78). – p. 869-878

5. Onyshchenko A. Adaptive method of training neural networks / A. Onyshchenko, K. Onyshchenko // Technique and technology. Science, Research, Development #29. Gdansk, 2020. – p. 9-1

Науковий керівник – Бісікало Олег Володимирович – професор, в. о. завідувача кафедри Автоматизації та інтелектуальних інформаційних технологій, Факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, Вінниця, e-mail: obisikalo@vntu.edu.ua

Анастасія Юріївна Барановська – студентка групи 1ІСТ-19б, Факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, Вінниця, e-mail: xktsumst@gmail.com

Максим Андрійович Лешок – студент групи 1ІСТ-19б, Факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, Вінниця, e-mail: max6leshok@gmail.com

Supervisor – Bisikalo Oleh Volodymyrovych – professor, senior of department of Automatization and Intellectual Informational Technologies, Faculty of Intellectual Informational Technologies and Automation, Vinnytsia National Technical University, Vinnytsia, e-mail: obisikalo@vntu.edu.ua

Baranovska Anastasiia Y. – student of 1IST-19b, Faculty of Intellectual Informational Technologies and Automation, Vinnytsia National Technical University, Vinnytsia, e-mail: xktsumst@gmail.com

Maksym Andriyovych Leshok – student of 1IST-19b, Faculty of Intellectual Informational Technologies and Automation, Vinnytsia National Technical University, Vinnytsia, e-mail: max6leshok@gmail.com