УДК 004.934

ОБРОБКА ПРИРОДНОЇ МОВИ ТА КОМП'ЮТЕРНА ЛІНГВІСТИКА

Юськів Є.В.
Науковий керівник – д.т.н., доц. Вечур О.В.
Харківський національний університет радіоелектроніки, каф. ПІ
м. Харків, Україна
тел. +38(095) 240-52-19,

This report is devoted to the topic of Natural Language Processing (NLP) and Computational Linguistics (CL). It provides an overview of these fields, including their history, key concepts, and applications. The report also discusses the challenges facing these fields, such as the ambiguity of human language and the need for large amounts of annotated data to train models. One particular problem addressed in this report is the difficulty of handling rare or unseen words in NLP models, and it presents various ways to resolve this problem, including the use of sub word units, character-level models, and transfer learning techniques. The report concludes with a comparison of these techniques and their effectiveness in addressing the challenge of rare words. The report highlights the importance of NLP and CL in processing natural language text and presents solutions to some of the challenges facing these fields.

Обробка природної мови (ОПМ) і комп'ютерна лінгвістика (МЛ) — це галузі, які націлені на те, щоб комп'ютери могли розуміти й аналізувати текст природною мовою. Останніми роками ці сфери набувають дедалі більшого значення через вибух цифрових даних і потребу в більш ефективних і ефективних способах їх обробки. Однією з є труднощі обробки рідкісних або невидимих слів у моделях ОПМ. Ми представляємо різні способи вирішення цієї проблеми, включно з використанням підслівних одиниць, моделей рівня символів і методів навчання переносу. ОПМ і МЛ мають широкий спектр застосування в різних сферах:

- методики ОПМ можуть бути використані для аналізу тексту та вилучення відповідної інформації з великих баз даних і колекцій документів;
- моделі ОПМ можна навчити ідентифікувати почуття, виражені в тексті, що має застосування в таких сферах, як маркетинг і політичний аналіз;
- техніки ОПМ можна використовувати для автоматичного перекладу тексту з однієї мови на іншу.

Незважаючи на численні застосування ОПМ і МЛ, ці галузі стикаються з кількома проблемами:

- як згадувалося раніше, людська мова за своєю суттю неоднозначна, що може ускладнити точну інтерпретацію та аналіз тексту моделями ОПМ;
- моделі ОПМ потребують великої кількості анотованих даних для навчання, але цих даних часто бракує, особливо в спеціалізованих областях;

- мовні дані можуть бути упередженими щодо певних груп або точок зору, що може призвести до несправедливих або неточних результатів моделей ОПМ;

- моделі ОПМ важко впоратися з рідкісними або невидимими словами, що може вплинути на їх здатність точно опрацьовувати текст.

Існують різні способи вирішення цих проблем у ОПМ та МЛ, зокрема:

- розробка більш складних алгоритмів, які можуть краще впоратися з неоднозначністю та рідкісними словами;
- збільшення доступності та різноманітності анотованих мовних даних для підвищення точності моделей ОПМ;
- розробка більш надійних методів обробки упередженості в мовних даних;
- ынтеграція багатьох джерел інформації, таких як бази знань і контекст, для підвищення точності моделей ОПМ.

Підводячи підсумок, у цьому звіті підкреслюється важливість ОПМ і МЛ для обробки тексту природною мовою та представлені деякі проблеми, з якими стикаються ці сфери. Він також пропонує вирішення однієї з цих проблем, а саме труднощів обробки рідкісних або невидимих слів у моделях ОПМ. Порівнюючи ці рішення, цей звіт має на меті допомогти дослідникам і практикам прийняти обґрунтовані рішення під час вибору найбільш прийнятної техніки для їхніх конкретних потреб.

Список використаних джерел:

1. Джурафскі, Д., Мартін, Дж. Г. (2019). Обробка мови та мовлення.

2. Меннінг, К. Д., Рагхаван, П., Шютце, Г. (2008). Вступ до інформаційного пошуку

3. Берд, С., Кляйн, Е., & Лопер, Е. (2009). Обробка природної мови з Python