УДК 004.853+004.912

Ісаєнков Я. О., магістр 2 курсу спеціальності 122 «Комп'ютерні науки»
Парамонов А. І., к.т.н., доцент доцент кафедри комп'ютерних наук та інформаційних технологій

ЗАСТОСУВАННЯ АУГМЕНТАЦІЇ ТЕКСТОВИХ ДАНИХ ДЛЯ ПОБУДОВИ СИНТЕТИЧНИХ НАВЧАЛЬНИХ ДАТАСЕТІВ

Донецький національний університет імені Василя Стуса, м. Вінниця

Серед головних проблем машинного навчання [1] можна виділити брак даних та/або їх погану якість. Здавалось б що зараз кількість даних не повинно бути проблемою, але вони є неструктурованими, дуже зашумленими, або знаходяться в недосяжних локаціях. Через те постійно не вистачає наборів даних для якісної підготовки моделей. Якщо ж різні моделі показають поганий результат для задачі і при цьому збір більшого набору даних не є можливим, то один зі шляхів вирішення цієї проблеми – використання техніки аугментації. Тобто збільшення набору даних завдяки різноманітним перетворенням існуючих елементів.

Наприклад, для збільшення датасету зображень, можна використовувати геометричні перетворення: повороти, збільшення, віддзеркалення або більш незвичайні такі як зміна яскравості, вирізання блоків пікселів, розмиття [2]. Але не так багато типів аугментацій існує для текстових даних. В роботі застосовується техніка перекладу: речення з набору даних перекладається з базової мови на любу обрану іншу мову, а потім отримане речення з цієї мови перекладається навпаки до базової. Переводити можна через декілька мов, але завжди потрібно вертатись до базової. Трансформації тексту виконується за допомогою відкритих систем машинного (автоматичного) перекладу.

Наприклад, на рисунку 1 зображено алгоритм аугментації тексту на базі перекладу через англійську мову. В порівнянні з базовим текстом, у аугментованому після першого слова з'явилась кома, фраза «не вміє літати» змінена на «не може літати», а набор слів «має чорно-біле забарвлення» змінився на «має чорно-білий колір». Іноді можуть траплятись більш жорсткі зміни, наприклад речення «Домашня птиця, самка півня» може бути трансформовано до речення «Домашня птиця, півень жіночий», при цьому втрачається лаконічність мови, але деякий зміст залишається, що може виступати для моделей як регуляризація від перенавчання.

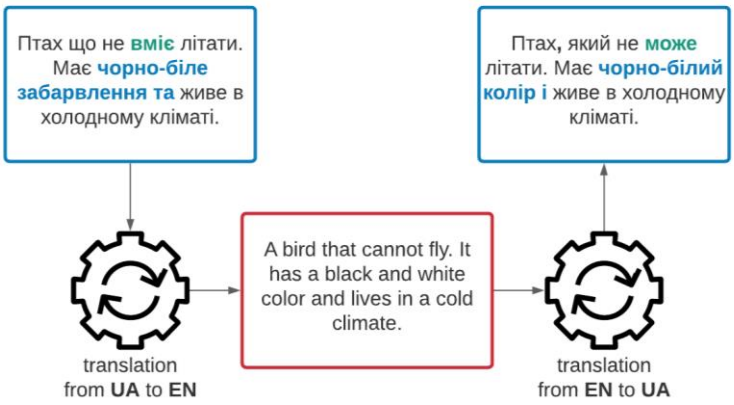

Рисунок 1 - Приклад роботи алгоритму аугментації перекладом

На рисунку 2 показано приклад жорсткіший агументації за допомогою двоетапного перекладу. Вхідне речення перекладається на англійську, потім китайську, знову на англійську а лише потім на українську мову. У такому підході більша вірогідність того, що сенс та деякі слова речення будуть зміненими під час перекладу. В прикладі в першому реченні ціла фраза була замінена на одне слово: «Птах, що не вміє літати» на «Нелітаючий птах».

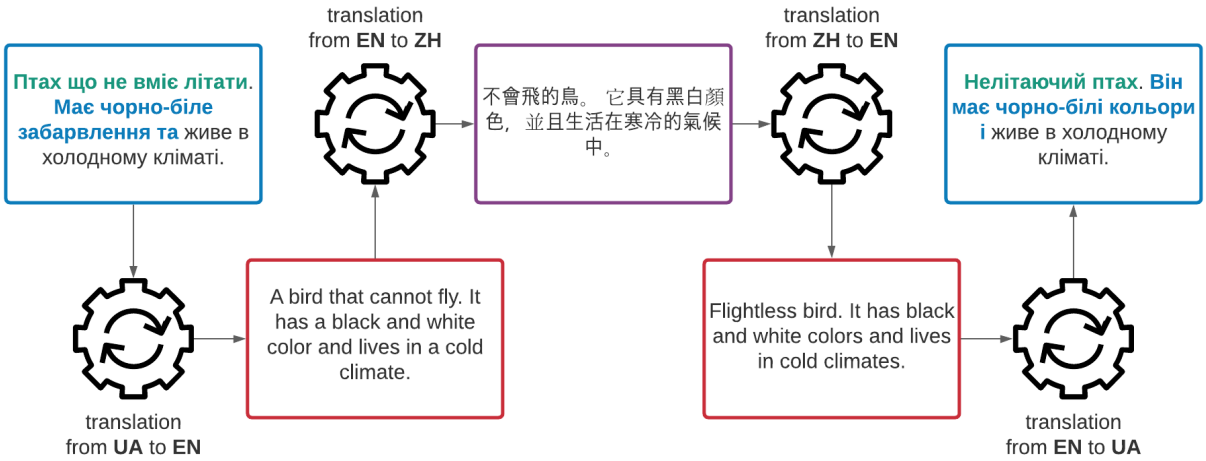

Рисунок 2 - Приклад роботи алгоритму аугментації двоетапним перекладом

Отримані таким способом речення, тепер можуть добавлятися до набору даних, та використовуватись для тренування. В більшості речень ця процедура додає нові слова, які раніше не були присутні, або змінює обороти в мові, але суть речення залишається загалом незміненою.

Описаний підхід дозволив доповнити тренувальний датасет описів слів українською мовою. Наприклад, до застосування аугментацій тренувальний набір містив по сім унікальних описів до кожного з 15 обраних класів, після її застосування до кожного унікального опису додалась його синтетична версія, в результаті кожен клас став мати по 14 описів. Однак, набір для тестування має містити лише по чотири описи на клас, та збільшувати кількість даних трансформаціями не варто – дані для тестування повинні бути як більш

реальними, щоб можна було оцінити поведінку моделі для даних що зустрічаються у справжньому світі.

В ході комп'ютерних експериментів на цьому підготовленому наборі даних над моделями машинного і глибокого навчання, такими як LSTM та RoBERTa [3-4] було відзначено, що чим більше унікальних базових даних ми маємо (без використання аугментацій), тим менше користі від синтетичних даних. Наприклад, при використанні в тренувальному наборі даних тільки від двох експертів, вдалось збільшити точність моделей за рахунок нових синтетичних даних. Але коли число експертів збільшилось та стало дорівнювати сьомі, аугментація перестала приносити збільшення точності моделей. Звісно знайти потрібну кількість експертів не завжди є можливим, саме тому застосування аугментації текстових даних для побудови синтетичних навчальних датасетів є досить зручним, та, як показали експерименти, діючим рішенням цієї проблеми.

Список літературних джерел

1. Haykin S., Neural Networks: A Comprehensive Foundation, Second Edition. Pearson Education, 1999. 842 p.
2. Albumentations. Do more with less data. URL: https://albumentations.ai/ (дата звернення: 15.11.2020).
3. Hochreiter S., Schmidhuber J., "Long Short-Term Memory," Neural Computation, vol. 9, pp. 1735-1780, November 1997.
4. Liu Y., Ott M., Goyal N., Du J., Joshi M., Chen D., Levy O., Lewis M., Zettlemoyer L., Stoyanov V., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv: 1907.11692 [cs.CL], Jul. 2019.