Використання нейронних мереж для визначення схожості текстів українською мовою

Виконав студент 3 року навчання, ІПЗ
Брус А. І.
Науковий керівник
Кандидат тех. наук, старший викладач Шабінська М. О.

ВСТУП

Кількість інформації, у тому числі і текстової, з кожним роком зростає все швидше. Виникає необхідність автоматизувати процеси аналізу та обробки цієї інформації. Одним із найбільш актуальних завдань є аналіз схожості текстів, у тому числі й україномовних.

У зв’язку зі складністю обробки текстів, написаних природними мовами, у галузі NLP широко використовуються методи машинного навчання, у тому числі і нейронні мережі.

Постановка завдання

- Розглянути основні існуючі моделі представлення слів та речень у числовому вигляді;
- Дослідити існуючі архітектури нейронних мереж для порівняння текстів;
- Розглянути існуючі інструменти, розроблені для україномовних текстів;
- На основі обраних алгоритмів реалізувати систему для порівняння україномовних текстів;
- Протестувати систему та проаналізувати результати;

Основні компоненти

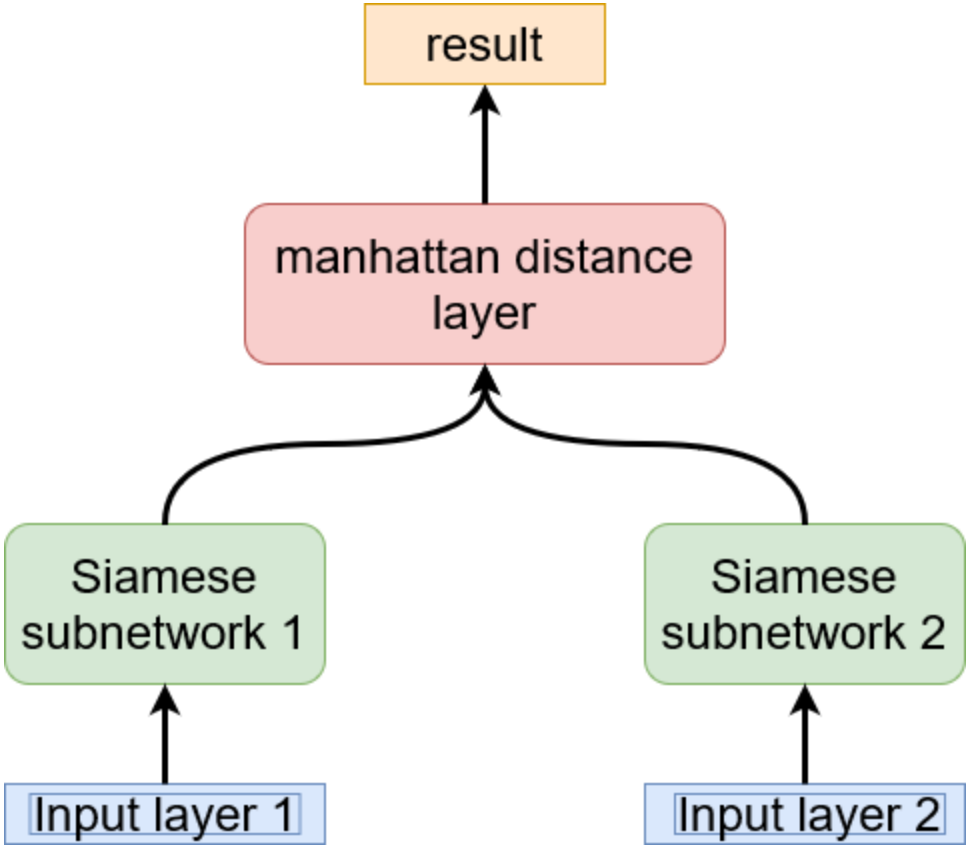
- Word2Vec - двошарова нейронна мережа, що приймає на вхід текст та формує множину нормалізованих векторів, кожен із яких представляє певне слово. Слугує для підготовки тексту для обробки складнішими глибинними нейронними мережами.

- LSTM - рекурентна нейронна мережа “з пам’яттю”, що враховує дані з попередніх проходжень рекурентним шаром та може “пам’ятати” ці дані.

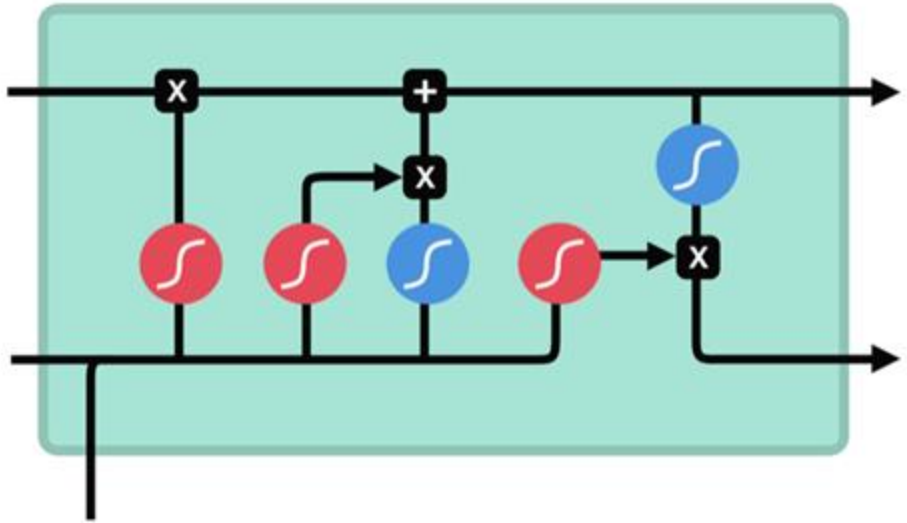
- CNN - мережа, що застосовує операцію згортки для обробки частин текстової послідовності, та агрегації для виділення найважливіших ознак.

Siamese neural network

Архітектура, що містить дві абсолютно ідентичні нейронні мережі. Кожній із підмереж подається на вхід різний набір даних. Обидві підмережі поєднує спільний шар, що визначає, схожі чи несхожі два речення, та продукує результат



Архітектура підмереж LSTM

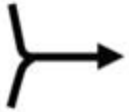


sigmoid

tanh

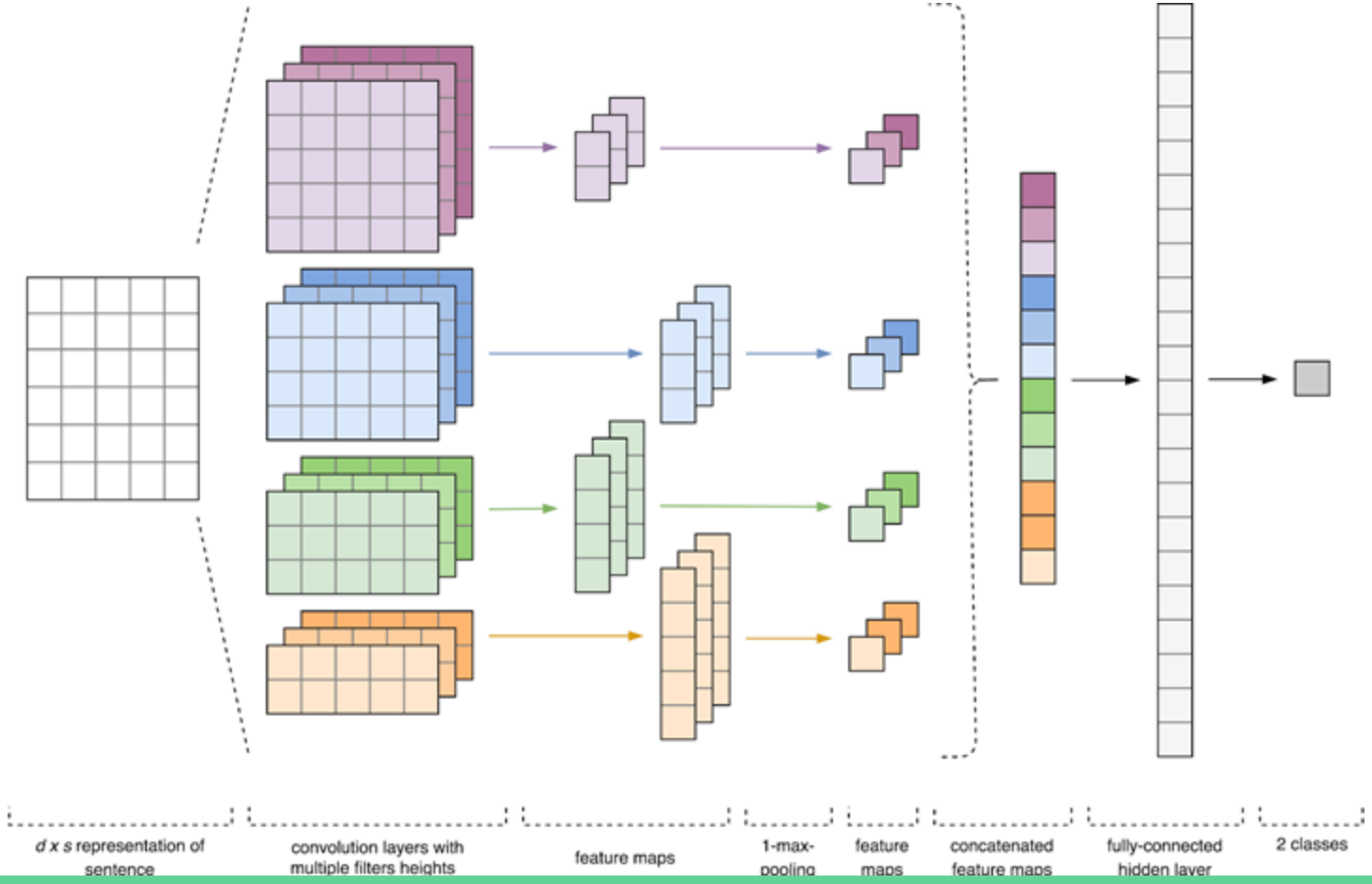
pointwise multiplication

pointwise addition



vector concatenation

Архітектура підмереж CNN



Тестові дані

Датасет сформований на основі колекцій текстів порталу lang-ua

Розмір датасету - 1200 пар речень українською мовою, із текстів категорії “Новини” та “Художня література”

Тестування проводилось для двох випадків - із використанням лематизації слів та без неї. Для кожного випадку обрана відповідний набір векторних представлень слів.

Для кожної моделі проведено 100 епох, після чого обрно найкращу та розблоковано для неї Embedding-шар. Далі проведено ще 20 епох для цієї моделі.

Результати тестування

для нелематизованих речень та векторної моделі lowercase

Модель	*accuracy*
CNN	0.6061
CNN з розблокованим embedding-шаром	0.6132
LSTM	0.6057
LSTM з розблокованим embedding-шаром	**0.6142**

для лематизованих речень та векторної моделі lowercase lemmatised

Модель	*accuracy*
CNN	0.6299
CNN з розблокованим embedding-шаром	0.6386
LSTM	0.6368
LSTM з розблокованим embedding-шаром	**0.6401**

Висновки

- Досліджено основні типи нейронних мереж для аналізу схожості текстів
- Створено та протестовано реазілації двох архітектур - CNN та LSTM
- Продемонстровано ефективність архітектур та вплив лематизації тексту на успішність передбачень
- Покращено розуміння принципів роботи та побудови нейронних мереж, видів та перспектив для NLP

Дякую за увагу!