**Обробка природної мови (Natural Language Processing, NLP)** - загальний напрямок штучного інтелекту і математичної лінгвістики. Воно вивчає проблеми комп'ютерного аналізу і синтезу природних мов. Стосовно до штучного інтелекту, аналіз означає розуміння мови, а синтез - генерацію грамотного тексту. Вирішення цих проблем буде означати створення більш зручної форми взаємодії комп'ютера і людини. [1] До основних завдань обробки природної мови відносять такі, як:

• Формування відповідей на питання (Question Answering)

• Аналіз емоційного забарвлення висловлювань (Sentiment Analysis)

• Знаходження тексту, відповідного зображенню (Image to Text Mappings)

• Машинний переклад (Machine Translation)

• Розпізнавання мови (Speech Recognition)

• Морфологічна розмітка (Part of Speech Tagging)

• Витяг сутностей (Name Entity Recognition)

**Способи аналізу**

**Word2Vec**

В основі даної технології лежить уявлення слів у вигляді векторів заданої розмірності, маючи в своєму розпорядженні схожі слова близько один до одного. Тобто, відстань між векторами слів, що позначають схожі речі буде значно менше, ніж між словами, значення яких мають мало спільного. Дана особливість дозволяє більш гнучко представляти дані, які в подальшому можуть бути використані в навчанні нейронних мереж, різних класифікаторів. Для створення бази відповідностей "слово - вектор", алгоритм спочатку переглядає весь виданий йому текст, складаючи "словник", який в наступних ітераціях роботи алгоритму, буде використаний для визначення відповідних векторів. Існує два основні підходи: CBOW (Continuous Bag of Words) і Skip-gram. CBOW - «безперервний мішок зі словами" модельна архітектура, яка передбачає поточне слово, виходячи з навколишнього його контексту. Архітектура типу Skip-gram діє інакше: вона використовує поточний слово, щоб передбачати навколишні його слова. [3]

**Визначення структури тексту**

Всі тексти на природній мові мають велику кількість слів, які не несуть інформації про даний текст. Наприклад, в англійській мові такими словами є артиклі. Дані слова називають шумовими або стоп-словами. Для досягнення кращої якості класифікації на першому етапі попередньої обробки текстів зазвичай необхідно видаляти такі слова. Другий етап попередньої обробки текстів - приведення кожного слова до основи, однаковою для всіх його граматичних форм. Це необхідно, тому що слова несуть один і той же сенс можуть бути записані в різній формі.

**Нейронні мережі**

Штучні нейронні мережі являють собою систему з'єднаних і взаємодіючих між собою простих процесорів - штучних нейронів. Алгоритм роботи таких процесорів найчастіше вкрай простий.

**RNN / LSTM** - рекурентні нейронні мережі [4], які відрізняються від іншого типу мереж тим, що крім зв'язків, що переходять від одного нейрона до іншого безпосередньо, як в мережах прямого поширення, а також зв'язку, що проходять "в часі". Тобто, сигнал від одного нейрона на етапі t перейде до іншого (або цього ж) нейрона на етапі t + 1. Таким чином рекурентні нейронні мережі можуть зберігати інформацію в часі, тим самим "запам'ятовуючи" деякі дані. Дана їх особливість як раз дуже сильно допомагає в перекладі, класифікації та обробки природного тексту в цілому, так як наша мова влаштована таким чином, що деякі дані на початку блоку тексту, можуть вплинути на розуміння і / або переклад в його кінці.

**CNN СНС** - надточні нейронні мережі найкраще показали себе в розпізнаванні об'єктів і образів на картинках, класифікації зображень, виділення особливостей і стисненні даних. Однак, їм знайшлося застосування і в обробці тексту.

**Seq2Seq** - універсальна бібліотека для Tensorflow [2], яка може використовуватися для машинного перекладу, визначення змісту тексту, моделювання діалогів, опису змісту зображень і т. д. Seq2Seq дозволяє створювати і навчати моделі нейронних мереж виду 'sequence to sequence'.

ЛІТЕРАТУРА

1. Обробка природної мови - Режим доступу: https://ru.wikipedia.org/wiki/Обработка\_естественного\_языка

2. Neural Machine Translation (seq2seq) Tutorial - Режим доступу: https://www.tensorflow.org/tutorials/seq2seq

3. Word2vec - Режим доступу: https://ru.wikipedia.org/wiki/Word2vec

4. LSTM - мережі довгої короткостроковій пам'яті - Режим доступу: <https://habrahabr.ru/company/wunderfund/blog/331310/>