Нейронные сети для анализа текстов

Задачи:

* Классификация текста
* Определение эмоциональной окраски текста
* Извлечение сущностей из текста
* Реферирование/аннотирование
* Генерация текста
* Автоматический перевод
* Чат-боты

К изучению:

**Архитектуры нейросетей для обработки текстов**

* Рекуррентные нейронные сети
* Одномерные сверточные сети
* Сети с вниманием
* Перенос обучения для анализа текстов

**Методы представления текста для обработки нейронной сетью**

**Задачи анализа текста**

* Классификация текста
* Определение эмоциональной окраски
* Автоматическая генерация текста

Векторное представление - метод представления строк, как векторов со значениями. Строится плотный вектор (dense vector) для каждого слова так, чтобы встречающиеся в схожих контекстах слова имели схожие вектора. Векторное представление считается стартовой точкой для большинства NLP задач и делает глубокое обучение эффективным на маленьких датасетах.

Техники:

Word2Vec

Word2vec принимает большой корпус (corpus) текста, в котором каждое слово в фиксированном словаре представлено в виде вектора. Далее алгоритм пробегает по каждой позиции t в тексте, которая представляет собой центральное слово c и контекстное слово o. Далее используется схожесть векторов слов для c и o, чтобы рассчитать вероятность o при заданном с (или наоборот), и продолжается регулировка вектор слов для максимизации этой вероятности.

1. **Skip-Gram**: рассматривается контекстное окно, содержащее*k* последовательных слов. Далее пропускается одно слово и обучается нейронная сеть, содержащая все слова, кроме пропущенного, которое алгоритм пытается предсказать. Следовательно, если 2 слова периодически делят cхожий контекст в корпусе, эти слова будут иметь близкие векторы.
2. **Continuous Bag of Words**: берется много предложений в корпусе. Каждый раз, когда алгоритм видим слово, берется соседнее слово. Далее на вход нейросети подается контекстные слова и предсказываем слово в центре этого контекста. В случае тысяч таких контекстных слов и центрального слова, получаем один экземпляр датасета для нашей нейросети. Нейросеть тренируется и, наконец, выход закодированного скрытого слоя представляет вложение (embedding) для определенного слова. То же происходит, если нейросеть тренируется на большом числе предложений и словам в схожем контексте приписываются схожие вектора.

[GloVe](https://nlp.stanford.edu/pubs/glove.pdf) захватывает значения одного word embedding со структурой всего обозримого корпуса. Чтобы сделать это, модель ищет глобальные совпадения числа слов и использует достаточно статистики, минимизирует среднеквадратичное отклонение, выдает пространство вектора слова с разумной субструктурой. Такая схема в достаточной степени позволяет отождествлять схожесть слова с векторным расстоянием.

**Машинный перевод**(Machine translation) — преобразование текста на одном естественном языке в эквивалентный по содержанию текст на другом языке. Делает это программа или машина без участия человека.

В традиционных системах машинного перевода приходится использовать параллельный корпус — набор текстов, каждый из которых переведен на один или несколько других языков. Например, имея исходных язык f (Французский) и целевой e (Английский), требуется построить статистическую модель, включающую вероятностную формулировку для правила Байеса, модель перевода p(f|e) , обученную на параллельном корпусе, и модель языка p(e) , обученную только на корпусе с английским языком.

**Нейросетевой машинный перевод** (Neural Machine Translation) — подход к моделированию перевода с помощью [**рекуррентной нейронной сети**](http://karpathy.github.io/2015/05/21/rnn-effectiveness/) (Recurrent Neural Network, RNN). RNN — нейросеть c зависимостью от предыдущих состояний, в которая имеет связи между проходами. Нейроны получают информацию из предыдущих слоев, а также из самих себя на предыдущем шаге. Это означает, что порядок, в котором подается на вход данные и тренируется сеть, важен: результат подачи “Дональд” — “Трамп” не совпадает с результатом подачи “Трамп” — “Дональд”.

Стандартная модель нейро-машинного перевода является сквозной нейросетью, где исходное предложение кодируется RNN, называемой **кодировщик** (encoder), а целевое слово предсказывается с помощью другой RNN, называемой **декодер** (decoder). Кодировщик «читает» исходное предложение со скоростью один символ в единицу времени, далее объединяет исходное предложение в последнем скрытом слое. Декодер использует обратное распространение ошибки для изучение этого объединения и возвращает переведённый вариант.

1. **Сквозное обучение**: параметры в NMT (Neural Machine Translation) одновременно оптимизируются для минимизации функции потерь на выходе нейросети.
2. **Распределенные представления**: NMT лучше использует схожести в словах и фразах.
3. **Лучшее исследование контекста**: NMT работает больше контекста — исходный и частично целевой текст, чтобы переводить точнее.
4. **Более беглое генерирование текста**: перевод текста на основе глубокого обучения намного превосходит по качеству метод параллельного корпуса.

Главная проблема RNN — проблема исчезновения градиента, когда информация теряется с течением времени.

[**Сети краткосрочной-долгосрочной памяти**](https://neurohive.io/ru/osnovy-data-science/lstm-nejronnaja-set/) (Long/short term memory, далее [LSTM](https://neurohive.io/ru/osnovy-data-science/lstm-nejronnaja-set/)) пытаются бороться с проблемой градиента исчезновения вводя гейты (gates) и вводя ячейку памяти. Каждый нейрон представляет из себя ячейку памяти с тремя гейтами: на вход, на выход и забывания (forget). Эти затворы выполняют функцию телохранителей для информации, разрешая или запрещая её поток.

* Входной гейт определяет, какое количество информации из предыдущего слоя будет храниться в этой ячейке;
* Выходной гейт выполняет работу на другом конце и определяет, какая часть следующего слоя узнает о состоянии текущей ячейки.
* Гейт забывания контролирует меру сохранения значения в памяти: если при изучении книги начинается новая глава, иногда для нейросети становится необходимым забыть некоторые слова из предыдущей главы.

[**Закрытые рекуррентные блоки**](https://arxiv.org/pdf/1412.3555v1.pdf) (Gated recurrent units, далее GRU) отличаются от LSTM, хотя тоже являются расширением для нейросетевого машинного обучения. В GRU на один гейт меньше, и работа строится по-другому: вместо входного, выходного и забывания, есть **гейт обновления** (update gate). Он определяет, сколько информации необходимо сохранить c последнего состояния и сколько информации пропускать с предыдущих слоев.

В случаях, где не требуется особая выразительность, GRU показывают лучше результат, чем LSTM.

Помимо этих трех главных архитектур, за последние несколько лет появилось много улучшений в нейросетевом машинном переводе. Ниже представлены некоторые примечательные разработки:

* [Sequence-to-Sequence Learning with Neural Networks](https://arxiv.org/pdf/1409.3215v3.pdf) доказали эффективность LSTM для нейронного машинного перевода. Статья представляет общий подход для последовательного обучения, для которого характерны минимальные предположения о структуре последовательности. Этот метод использует многослойную LSTM, чтобы отобразить входящую последовательность в виде вектора с фиксированной размерностью, далее идет применение другой LSTM для декодирования целевой последовательности из вектора.
* [Neural Machine Translation by Jointly Learning to Align and Translate](https://arxiv.org/pdf/1409.0473v6.pdf) представили механизм внимания (attention mechanism) в NLP (который будет рассмотрен в следующей части). Признавая факт, что использование вектора фиксированной длины является узким местом в улучшении результативности NMT, авторы предлагают разрешать модели автоматически искать части исходного предложение, которые релевантны к предсказанию целевого слова, без необходимости явного формирования этих частей.
* [Convolutional over Recurrent Encoder for Neural Machine Translation](https://ufal.mff.cuni.cz/pbml/108/art-dakwale-monz.pdf) усиливают стандартный RNN кодировщик в NMT с помощью дополнительного сверточного слоя, чтобы захватывать более широкий контекст на выходе кодировщика.
* Google создала собственную NMT систему, называемую [Google’s Neural Machine Translation](https://arxiv.org/pdf/1609.08144.pdf" \t "_blank), которая решает задачи точности и простоты применения. Модель состоит из глубокой LSTM сети с 8 кодирующими и 8 декодирующими слоями и использует как остаточные связи, так и attention-связи от декодер- до кодер-сети.
* Вместо использования рекуррентных нейросетей, Facebook [AI](https://neurohive.io/ru/osnovy-data-science/iskusstvennyj-intellekt-voprosy-i-otvety/) Researchers используют [сверточную нейронную сеть](https://arxiv.org/pdf/1705.03122.pdf" \t "_blank) для задач sequence-to-sequence обучения в NMT.