ТЕХНОЛОГИИ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ТЕКСТОВ

Е.С. Попова
Научные руководители: В.Г. Спицын, Ю.А. Иванова
Томский политехнический университет
esp9@tpu.ru

Введение

Задачи обработки естественных языков (natural language processing, NLP), находятся на пересечении computer science, искусственного интеллекта и лингвистики. Данная область становится все более актуальной в связи с постоянно растущим объемом информации в интернете и потребностью в ней ориентироваться, а с развитием голосовых интерфейсов и чат-ботов, NLP стала одной из самых важных технологий искусственного интеллекта.

Перечислим основные классы задач, где используются методы анализа текстов на естественных языках [1]:

I. Анализ текста и информационный поиск.

1. Машинный перевод с одного языка на другой.
2. Системы, поддерживающие диалог с пользователем.
3. Поиск текстовой информации по запросу пользователя.
4. Извлечение информации из текстов. Извлечение фактов – переход от текстов к структурированной информации, перенос фактов в базу данных.
5. Вопросно-ответные системы. Поиск точного ответа на вопрос, а не документа как при поиске информации.
6. Автоматическое резюмирование. Построение краткого изложения текста.
7. Поиск близких текстов (документов). Выявление заимствований и плагиата.
8. Кластеризация и классификация текстов. Упорядочивание текстов по группам похожих документов или отнесение документа к предопределенному классу.
9. Контентный анализ: определение характеристик текста и автора, эмоциональной окраски текста, построение психолингвистического портрета автора.

II. Синтез текстов. Автоматическая генерация текстов с заданными характеристиками.

На сегодняшний день помимо классических алгоритмов интеллектуального анализа текстов, большое распространение получили методы, основанные на глубоком обучение нейронных сетей (deep learning), которые предлагают гибкий, универсальный и обучаемый подход для представления мира как в виде визуальной, так и лингвистической информации.

Предобработка теста

Предобработка текста позволяет уменьшить исходное пространство признаков, без потери полезной информации. Ниже приведены основные методы морфологической и синтаксической предобработки текста:

Токенизация – это самый первый шаг при обработке текста. Заключается в разбиении длинных строк текста в более мелкие: абзацы делим на предложения, предложения на слова.

Нормализация – для качественной обработки текст должен быть нормализированным. Все слова приводятся к одному регистру, удаляются знаки пунктуации, расшифровываются сокращения, числа приводятся к их текстовому написанию и т.д. Нормализация необходима для унификации методов обработки текста.

Стэмминг – это устранение придатков к корню, то есть отделение суффикса, приставки, окончания и приведение слова к основе.

Лемматизация – близка к стеммизация. Отличие в том, что лемматизация приводит слово к смысловой канонической форме слова (инфинитив для глагола, именительный падеж единственного числа – для существительных и прилагательных). Например: зафрахтованный – фрахтовать, ценами – цена, лучший – хороший.

Удаление стоп-слов. Стоп-слова – слова, которые не несут никакой смысловой нагрузки. Их еще называют шумовыми словами. Например, в английском языке это артикли, в русском – междометия, союзы, маты и т.д.

Перевод текста в векторное представление

Векторное представление считается стартовой точкой для большинства NLP задач. Это метод сопоставления текстовому слову некоторого числового вектора фиксированной размерности. Векторное представление может строиться не только для слов, но и для произвольных объектов.

Так же векторы могут обладать разнообразными полезными свойствами, например, отражать семантическую близость между словами.

Способы получения векторных представлений:

- One-hot encoding
- SVD
- Topic modeling
- word2vec, GloVe, FastText, StarSpace

Рассмотрим подробнее технику векторного представления Word2vec от Google, которая пользуется популярностью, и часто используются для задач NLP.

Word2Vec – предназначена для статистической обработки больших массивов текстовой информации. W2V собирает статистику по совместному появлению слов в фразах, после чего методами

нейронных сетей решает задачу снижения размерности и выдает на выходе компактные векторные представления слов, в максимальной степени отражающие отношения этих слов в обрабатываемых текстах.

Для достижения лучшего результата Word2vec удаляет из набора данных бесполезные слова (или слова с большой частотой появления, в английском языке — a, the, of, then). Это поможет улучшить точность модели и сократить время на обучения.

Сверточные нейронные сети для задачи классификации текстов

Одной из распространённых задач NLP является классификации (категоризация) текстов.

Примерами задач классификации текстов являются такие задачи, как фильтрация спама, анализ тональности, определение авторства и т.д.

Для решения данной задачи последнее время активно используется сверточные нейронные сети (convolutional neural network, CNN), которые исходя из недавно вышедшей статьи [2] от коллектива авторов из Intel и Carnegie-Mellon University подходят для этого даже лучше, чем рекуррентные нейронные сети (recurrent neural network, RNN), которые безраздельно властвовали в эта области на протяжении последних лет.

Далее опишем основные подходы использования сверточных нейронных сетей для задачи классификации текстов.

Посимвольный подход

Посимвольных подход для классификации текстов с помощью сверточных нейронных сетей был предложен в статье [4]. Опишем данный метод подробнее. Назовем алфавитом упорядоченный набор символов. Пусть выбранный алфавит состоит из m символов. Каждый символ алфавита в тексте закодирован с помощью 1−m−кодировки. (т. е. каждому символу будет сопоставлен вектор длины m элемент которого равен единице, в позиции равной порядковому номеру символа в алфавите, и нулю во всех остальных позициях.)

Если в тексте встретится символ, который не вошел в алфавит, то необходимо закодировать его вектором длины m состоящим из одних нулей. Из текста выбираются первые l символов. Параметр l должен быть большим, чтобы в первых l символах содержалось достаточно информации для определения класса всего текста.

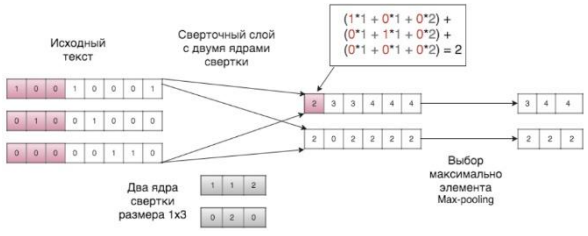

Рис. 1. Посимвольный подход

Далее полученные векторы составляются в матрицу размера $m \times l$, в которой в каждый столбец будет иметь не более одной единицы. Каждая строка полученной матрицы используется как отдельная карта признаков. На вход светрочной нейронной сети подается m карт признаков размера $1 \times l$ аналогично изображению. Архитектуру сети необходимо выбирать исходя из задачи. На рисунке 1 приведен пример посимвольного подхода для l = 6, m = 3. В примере показан один сверточный и один субдискретизирующий слой.

Подход с использованием кодирования слов

В статье [3] описан подход, где каждому слову в тексте сопоставляется вектор фиксированной длины, затем из полученных векторов для каждого объекта выборки составляется матрица, которая аналогично изображениям подается на вход сверточной нейронной сети. На рисунке 2 приведен пример сверточной нейронной сети с использованием кодирования слов.

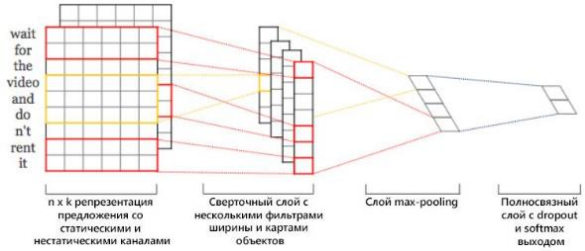

Рис. 2. Кодирование слов

Заключение

В результате проведенного исследования были выявлены основные группы задач NLP, рассмотрены методы предобработки и векторизации текстов. Так же в ходе исследования была изучена возможность применения сверточных нейронных сетей для задачи классификации текстов. Работа поддержана грантом РФФИ № 18-08-00977 А.

Список использованных источников

1. Федюшкин Н.А., Федосин С. А. Понятие, проблемы и разновидности интеллектуального анализа текста — Проблемы и достижения в науке и технике. Сборник научных трудов по итогам международной научно-практической конференции - № 3 - г. Омск, 2016 - 206 с.
2. Bai, S., Kolter, J. Z., & Koltun, V. (2018). An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. arxiv.org/abs/1803.01271
3. Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), 1746–1751.
4. Zhang, X. Character-level convolutional networks for text classification / Xiang Zhang, Junbo Zhao, Yann LeCun // In Advances in Neural Information Processing Systems. - 2015. - Feb. - 649-657pp.