

УДК 004.83

SSL модель в задаче распознавания специализированной речи

Безлюдов Михаил Дмитриевич bezlyudovmd@student.bmstu.ru

Корлякова Мария Олеговна mkorlyakova@bmstu.ru

КФ МГТУ им. Н.Э. Баумана, Калуга, Россия

Изучен подвид задачи обработки естественного языка — задачу распознавания специализированной речи, ее актуальность и возможные сферы применения. Особое внимание уделяется обработке устной информации с точки зрения семантической близости: ее обеспечение с применением методов word embedding, а также основные методы и концепции, применяемые из word embedding, которым дается краткое описание с упоминанием преимуществ и недостатков. Также рассмотрен один из подходов к решению задачи распознавания специализированной речи при помощи нейронных моделей класса Self Supervised Learning (SSL). Далее приведен сравнительный анализ SSL модели с моделями, обучающиеся согласно классическому подходу.

Ключевые слова: классификация текстов, Self-Supervised Learning, word embedding

Автоматизация и цифровизация все больше проникает в нашу жизнь во всех сферах жизни и производства. Однако компьютеры и различные системы ввиду своего устройства не понимают человеческую речь так, как ее понимают люди между собой. Прежде всего, голос или текст для компьютера — это обычный файл. Такая прямолинейная интерпретация без учета смыслов делает работу пользователя менее эффективной или вовсе бессмысленной. Отсюда возникает необходимость в алгоритме, который был бы способен распознавать специализированную речь.

В том или ином виде, данная задача сводится к задаче классификации текстов. Задача классификации текстов [1] (англ. Text Classification) — область машинной обработки информации и искусственного интеллекта, заключающаяся в категоризации письменного текста в определенной последовательности и его объединения в упорядоченные словари. Основные подходы при решении данной задачи включают:

- лексический анализ: выявление отдельных слов и фраз, которые используются в речи наиболее часто;
- синтаксический анализ: анализ структуры предложений и грамматических конструкций для определения контекста сказанного;
- семантический анализ: изучение смысловых связей между словами и фразами для более глубокого понимания текста.

При решении данной задачи в целом и при решении распознавания специализированной речи в частности, ключевую роль играет семантическая составляющая. Чтобы получить возможность представить семантическую близость, наиболее подходящим методом решения задачи классификации текстов является использование процесса под названием word embedding [2],

то есть сопоставления слова в тексте некоторому вектору. Краткое описание методов word embedding, а также описание их основных преимуществ и недостатков приведены в таблице.

Описание методов word embedding, применяемых в классификации текстов

Название метода	Краткое описание	Преимущества метода	Недостатки метода
One-hot encoding (OHE)	Создание вектора длиной в словарь, ставится единица в позиции, соответствующей номеру слова в словаре	Простота реализации; Эффективен при малом наборе данных	Неэффективен при больших объемах данных; Семантическая близость слов или учет контекста не обеспечивается
TF	Представление набора текста в виде матрицы «слово-документ»	Эффективность при малом наборе данных; Указание слов, употребляемых в документе	
TF-IDF	Модификация TF метода. В матрице «слово-документ» вес слова пропорционален количеству его употребления в документе и обратно пропорционален частоте употребления в других документах	Эффективность при малом наборе данных; Возможность выявления уникальных слов в документе	
Word2vec	Каждому слову присваивается вектор. Далее идет предсказание появления слова по его окружению: учитываются такие вектора соседей, чтобы вероятность появления слова была равна вероятности встретить его в таком же окружении в реальном тексте	Модель более эффективна при большом наборе данных; Частично реализована классификация текста по семантической близости	Зависимость от качества данных в наборе

Из представленной таблицы очевидно, что OHE, TF и TF-IDF методы имеют практически идентичные преимущества и схожие недостатки, при том для метода word2vec величина данных наоборот является преимуществом. Причина заключается в том, что word2vec метод основан на ином подходе к обучению моделей.

Self-Supervised Learning (или unsupervised) [3, 4] — это метод обучения нейронных моделей, при котором она обучается самостоятельно, используя избыточный набор данных. Сама задача при этом формируется исходя

из внутренней структуры самих объектов в наборе, либо из базовых знаний о них.

Прежде чем рассмотреть Self-Supervised Learning, рассмотрим классический подход к обучению нейронных сетей. Он подразумевает случайную инициализацию весов модели и использование немалых объемов данных, размеченных для лучшей точности. Поскольку параметры в современных нейронных моделях могут исчисляться миллионами, случайная инициализация является единственным выходом для обеспечения сходимости.

Из приведенного выше описания исходят три главные проблемы классического метода обучения нейронных моделей:

- необходимость инициализировать модель перед началом каждого обучения;
- большая ресурсоемкость;
- сильная зависимость от обучающей выборки, прежде всего — от ее размеров.

Источник первой и второй проблем кроется в необходимости каждый раз инициализировать модель. Это происходит довольно часто, так как во время эксперимента может поменяться архитектура модели, либо параметры, либо набор данных. Как следствие, процесс обучения модели приходится каждый раз начинать сначала для обеспечения лучшего результата. Частичным решением данных проблем может послужить дообучение уже предтренированной модели для ускорения сходимости. Такое решение заключается в тренировке модели на гораздо больших данных, чем тех, что связаны с исходной задачей.

Это усугубляет третью проблему. Если для обучения будет дан большой объем тренировочных данных, то модель не сможет эффективно выполнить свою задачу на тестовых. Вопрос с малым объемом исходной выборки стоит особенно остро в контексте задачи распознавания специализированного текста, т. к. очевидно, что объем реальных данных к задаче априори меньше.

Из описания выше можно сделать вывод, что классический метод обучения нейронных моделей малоприменим для распознавания речи в специализированных лексических областях.

Как уже было упомянуто, SSL метод обучения основан на использовании информации, которая и уже известна про каждый конкретный объект, следовательно, необходимости делать разметку данных нет [5]. Согласно данному подходу, сначала выполняется два важных этапа:

- на основе имеющейся информации модель делает искусственную (псевдо) метку, на которой обучается модель, чтобы сформировать представления об объекте;
- модель проверяет качество полученных представлений. Почти во всех экспериментах в статьях по самообучению на полученных признаковых представлениях в последующих задачах обучают простые модели: логистическую регрессию или метод ближайшего соседа.

Методы SSL начали массово появляться в нейронных сетях именно после успехов в задачах естественной обработки языка. Примером этого успеха яв-

ляется как раз word2vec модели. Они берут неразмеченный набор текстов, затем сами ставят метку на определенном слове или словах и начинают обучение на этой задаче.

Далее такая обученная модель может применяться в другой задаче, которая отличается от изначальной. В задачах распознавания речи эта техника обучения позволяет быстро адаптировать модель к специализированным терминам и особенностям актуального речевого окружения. Проективные слои позволяют значительно изменять свойства модели без существенных затрат ресурсов.

Литература

- [1] *Плавное введение в Natural Language Processing (NLP)*. URL: <https://datastart.ru/blog/read/plavnoe-vvedenie-v-natural-language-processing-nlp> (дата обращения 22.02.2025).
- [2] *Чудесный мир Word Embeddings: какие они бывают и зачем нужны?* URL: <https://habr.com/ru/companies/ods/articles/329410/> (дата обращения 22.02.2025).
- [3] *Self-Supervised Learning. Проблематика и постановка задачи*. URL: <https://habr.com/ru/articles/704710/> (дата обращения 22.02.2025).
- [4] *Self-Supervised Learning and Its Applications*. URL: <https://neptune.ai/blog/self-supervised-learning> (дата обращения 22.02.2025).
- [5] *Анализ малых данных*. URL: <https://alexanderdyakonov.wordpress.com/2020/06/03/%D1%81%D0%B0%D0%BC%D0%BE%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5-self-supervision/> (дата обращения 22.02.2025).

SSL model in specialized speech recognition task

Bezlyudov M.D bezlyudovmd@student.bmstu.ru

Korlyakova M.O. mkorlyakova@bmstu.ru

BMSTU (KB), Kaluga, Russia

This work studies a subtype of the natural language processing problem — the problem of specialized speech recognition, its relevance and possible fields of application. Special attention is given to the processing of oral information in terms of semantic proximity: its provision with word embedding methods, as well as basic methods and concepts applied from word embedding, which are briefly described with reference to the advantages and disadvantages. Also, in this article we consider one of the approaches to solve the problem of specialized speech recognition — by means of neural models of the class Self Supervised Learning (SSL). Next, a comparative analysis of SSL models with models learning according to the classical approach is performed.

Keywords: text classification, Self-Surveyed Learning, word embedding