Морфологический анализ текста с помощью нейронных сетей

А.Н. Жданова
Самарский национальный исследовательский университет им. академика С.П. Королева
Самара, Россия
zhdan.aleksandra@gmail.com

А.В. Куприянов
Самарский национальный исследовательский университет им. академика С.П. Королева
Институт систем обработки изображений - филиал ФНИЦ «Кристаллография и фотоника» РАН
Самара, Россия
akupr@ssau.ru

Д.С. Шеренков
Самарский национальный исследовательский университет им. академика С.П. Королева
Самара, Россия
dsherenkov000@gmail.com

***Аннотация*—Статья посвящена применению технологии нейронных сетей для решения задачи морфологического анализа текста. Для обучения был использован банк размеченного текста, где каждому слову поставлены в соответствие его часть речи и форма слова. Было проведено сравнение точности разметки с доступными сервисами, использующими как нейросетевой подход, так и библиотечный метод морфологического анализа.**

Ключевые слова— морфологический анализ, рекуррентная нейронная сеть, автоматическая обработка текста.

1. Введение

Наука о естественном человеческом языке, лингвистика, имеет множество отраслей и разделов, одним из которых является морфология. Морфологический анализ – одна из базовых задач автоматической обработки текста, относящаяся к типу классификации последовательностей. Такие задачи, как автореферирование текста, определение эмоциональной характеристики, машинный перевод, распознавание сущностей и многие другие, первоочередно включают в себя задачу сопоставления каждого слова, исследуемого текста, с его морфологическими признаками, как один из первых этапов [1]. Небольшие ошибки морфологического анализа могут привести к более серьезным последствиям в дальнейших этапах обработки текста, что обуславливает значимость и актуальность задачи морфологической разметки [2].

2. Сравнение со словарными методами

Словарный метод морфологического анализа текста заключается в определении морфологических признаков отдельно взятого слова с помощью нескольких словарей. К таким словарям относятся: словари основ существительных, окончаний существительных, основ прилагательных, основ глаголов, словарь служебных частей речи и т. д. [3]. Это является первым преимуществом в пользу нейросетевого подхода, для которого не нужно использовать настолько объемные и строгоструктурированные данные. В виду постоянного расширения языков, словари периодически должны дополняться новыми словами. Нейронная сеть же способна обрабатывать не только новые слова, но и несуществующие, опираясь на контекст и общую структуру слова.

Другой проблемой, с которой тяжело справляться словарным методам, являются различные виды омонимии. Обычная омонимия – это совпадение произношения и написания слов, совершенно разных по значению [4]. Частичная лексическая омонимия – совпадение различных форм одного слова [4]. Например, слово «физики» может обозначать как группу ученых, так и являться родительным падежом слова «физика». Также стоит отметить проблему омографии, когда слова имеют одинаковое написание, но разное ударение и смысл [4]. Например, словосочетание «большая часть» имеет разный смысл в зависимости от того на какой слог поставить ударение в слове «большая», на первый или на второй.

Одно из главных преимуществ нейронных сетей – это возможность нахождения скрытых зависимостей между входными и выходными данными, на что не способны методы с использованием словарей, лингвистические методы и т.д.

3. Реализация модели

Для реализации модели была использована библиотека Keras. С помощью нее можно спроектировать нейронную сеть, задав параметры для ее оптимальной настройки под канкретную задачу. Параметр Dropout был задан 0.2, что означает, что 20% случайно выбранных нейронов игнорируются во время обучения на каждом цикле обновления весов. Это поможет избежать быстрой перенастройки сети. Количество эпох было установлено равным восьми, так как при слишком большом количестве итераций многослойный перцептрон начинает переобучаться. Был задан параметр оптимизации «Adam», реализующих метод стахостического градиентного спуска. Параметр shuffle задан как true, что обеспечивает перетасовку тренировочных данных перед началом каждой эпохи.

4. Обучение нейронной сети

Нейронная сеть обучается на наборе данных, который состоит из 200 000 слов, каждому из которых поставлено в соответствие тэг, определяющий часть речи и форму слова. Пример тренировочного набора представлен на рисунке 1.

```
[('Повязанный', 'ADJ'), ('вокруг', 'ADP'), ('шеи', 'NOUN'),
('шелковый', 'ADJ'), ('платок', 'NOUN'), ('придавал', 'VERB'),
('его', 'DET'), ('довольно', 'ADV'), ('будничному', 'ADJ'),
('костюму', 'NOUN'), ('некоторую', 'DET'), ('элегантность', 'NOUN'),
('.', 'PUNCT')]
```

Рис. 1. Часть тренировочного набора данных

Тестовый набор данных имеет ту же структуру для автоматического определения точности полученных результатов.

В результате обучения модели была достигнута точность разметки 93.5%. Результаты проведения экспериментов для русскоязычной модели представлены на рисунке 2.

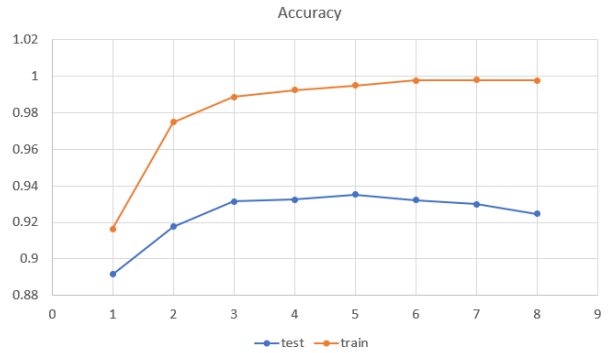

Рис. 2. Графики зависимостей точности разметки текста от количества эпох

Из графиков, представленных на рисунке видно, что максимальная точность разметки для тестового набора данных достигается после пяти эпох обучения и составляет 93.5%. Далее точность разметки для тренировочных наборов продолжает возрастать, а для тестовых начинает убывать, из чего можно сделать вывод о том, что начался процесс переобучения.

По некоторым данным, точность ручной частеречной разметки составляет примерно 98%, что достаточно близко к полученным 93.5%. Можно сказать, что применение модели многослойного перцептрона для решения задачи частеречной разметки оказалось успешным.

5. ЗАКЛЮЧЕНИЕ

В ходе исследования было проведено сравнение нейросетевого и словарного методов для решения задачи морфологической разметки текста. Реализована модель многослойного перцептрона для решения данной задачи.

Достигнутая точность морфологической разметки текста близка к сервисам, использующим словарный метод для решения данной задачи, а в некоторых случаях оказалась выше. Это доказывает перспективность использования нейронных сетей для решения как отдельной задачи морфологической разметки, так и в качестве начального этапа для решения более сложных и практических задач автоматической обработки текста.

ЛИТЕРАТУРА

[1] Кочконбаева, Б.О. О морфологическом анализе в приложениях автоматической обработки текста / Б.О. Кочконбаева // Бюллетень науки и практики. – 2018 – Т. 4, № 12. – С. 608-612.

[2] Цитульский, А.М. NLP – Обработка Естественных Языков / А.М. Цитульский, А.В. Иванников, И.С. Рогов // StudNet. – 2020. – № 6. – С. 467-475.

[3] Бажанова, А.И. Разработка морфологического анализатора для построения понятийного аппарата электронной библиотеки кафедры АСУ / А.И. Бажанова // Информатика и компьютерные технологии. – 2011. – С. 326-330.

[4] Бочаров, В.В. Прикладная и компьютерная лингвистика / В.В. Бочаров, И.С. Николаев, О.В. Митренина, Т.М. Ландо. – М.: URSS, 2017.