УДК 004.42

ИСПОЛЬЗОВАНИЕ НЕЙРОННЫХ СЕТЕЙ ДЛЯ РАСПОЗНАВАНИЯ ИМЕНОВАННЫХ СУЩНОСТЕЙ

Марков И.Р. – студент группы Пим-251, I курс
Научный руководитель: Глебова Е.А., старший преподаватель
Кузбасский государственный технический университет имени Т.Ф. Горбачева
Россия, г. Кемерово

Обработка естественного языка – это наука о проектировании методов и алгоритмов, которые принимают или порождают неструктурированные данные естественного языка [1]. В условиях экспоненциального роста объемов неструктурированных текстовых данных актуальность задач их автоматического анализа и извлечения смысловой информации существенно возрастает. Одной из фундаментальных подзадач информационного поиска и понимания текста является распознавание именованных сущностей (Named Entity Recognition, NER). Целью NER является идентификация в тексте упоминаний жестко обозначенных объектов (именованных сущностей) и их классификация по заранее заданным категориям.

Типичными категориями для извлечения являются:

1. PER (Person) — имена людей.
2. LOC (Location) — географические названия (города, страны, реки).
3. ORG (Organization) — названия компаний, учреждений, организаций.
4. MISC (Miscellaneous) — прочие сущности (например, названия произведений, события).

Решение задачи NER является критически важным для широкого спектра прикладных систем, таких как: информационный поиск и извлечение фактов, семантический анализ, машинный перевод, биоинформатика, автоматическое реферирование [2].

Для решения задач распознавания и извлечения информации из текстов используются два главных подхода: основанный на правилах, или инженерный, и основанный на машинном обучении [3].

Исторически развитие методов NER прошло несколько этапов: от систем, основанных на рукописных лингвистических правилах и словарях, через статистические подходы с использованием моделей машинного обучения к современным методам глубокого обучения, которые показали революционное повышение точности.

Статистические методы, в частности, CRF, позволили автоматизировать процесс обучения. Модель обучалась на размеченных корпусах текстов и учитывала такие признаки, как регистр букв, наличие цифр, морфологические характеристики и контекстное окружение слова. Однако качество таких моделей сильно зависело от релевантности и полноты набора извлекаемых признаков.

Прорыв в области NER, как и во многих других задачах NLP, связан с широким внедрением глубоких нейронных сетей. Модели на основе рекуррентных нейронных сетей (RNN), с долгой краткосрочной памятью (LSTM) и механизмами внимания, позволили автоматически извлекать сложные контекстные признаки, избавив разработчиков от необходимости их ручного конструирования. RNN предназначены для обработки последовательностей путем рекуррентного применения одной и той же функции к каждому элементу последовательности с передачей скрытого состояния. Это позволяет им учитывать историю предыдущих элементов.

Классическим и лучшим подходом долгое время являлась гибридная архитектура BiLSTM-CRF. Векторы слов, часто наполненные символьно-уровневыми представлениями, подаются на вход BiLSTM-слою, который извлекает контекстные признаки. Затем выход BiLSTM поступает на вход линейному слою, который прогнозирует числовые очки для каждого класса. Далее условное случайное поле (CRF) используется как выходной слой, который моделирует переходные вероятности между последовательными метками, что позволяет учитывать структурные зависимости внутри последовательности меток. Данная архитектура демонстрировала высокую эффективность и оставалась эталоном до появления других систем.

Однако настоящую революцию произвело появление архитектуры трансформер и, в частности, предобученных языковых моделей, таких как BERT. BERT, в отличие от однонаправленных RNN, использует механизм самовнимания, позволяющий модели анализировать контекст слова одновременно со всех сторон — как слева, так и справа. Это позволяет модели напрямую моделировать зависимости между любыми двумя словами в предложении, независимо от их расстояния друг от друга, решая проблему исчезающего градиента, присущую RNN. Предобучение на больших корпусах текстов позволяет BERT иметь глубокие представления о языке, которые затем могут быть эффективно дообучены для решения конкретных задач, включая NER, с относительно небольшими объемами размеченных данных.

Для того чтобы использовать большие языковые модели на подобии BERT для решения конкретных прикладных задач их необходимо дообучить на датасетах с данными из конкретной области. Датасет должен иметь уже размеченные сущности, которые в дальнейшем должна будет предсказывать уже обученная нейронная сеть.

Следующим шаг - токенизация текста, она является критически важным этапом. В отличие от стандартных токенизаторов, необходимо использовать токенизатор, соответствующий предобученной модели BERT. Особенность заключается в том, что одно слово может быть разбито на субтокены. Для корректного выравнивания разметки применяется стратегия, при которой метка сущности присваивается только первому субтокену слова, а последующие маркируются специальным тегом (например, -100), игнорируемым при расчете функции потерь. Этот подход гарантирует, что границы сущностей, определенные на уровне слов, корректно проецируются на уровень субтокенов модели.

Процесс дообучения представляет собой задачу оптимизации с целью настройки весовых коэффициентов как добавленного классификационного слоя, так и (в большинстве случаев) параметров самой предобученной модели BERT. Данный процесс инициируется с весов, полученных в ходе предобучения на больших корпусах текстов, что является реализацией трансферного обучения.

В качестве целевой функции минимизации применяется функция потерь перекрестной энтропии (cross-entropy loss), вычисляемая только по тем позициям в последовательности, для которых определены валидные метки (т.е. по первым субтокенам слов). Вычисленный градиент распространяется обратно через все слои модели, осуществляя тонкую настройку ее параметров в соответствии со спецификой целевой задачи NER.

Оценка качества дообученной модели проводится на отдельном валидационном наборе данных, не участвовавшем в обучении. Поскольку задача NER подразумевает корректное определение как типа, так и границ сущности, стандартные метрики классификации являются недостаточными. Для всесторонней оценки применяются метрики, ориентированные на сущности: точность (Precision), полнота (Recall) и F1-мера (F1-score). Сущность считается корректно распознанной только в случае полного совпадения ее границ и предсказанного класса.

Современные архитектуры глубокого обучения, в особенности, основанные на трансформерах, такие как BERT, демонстрируют превосходство над традиционными методами в задаче распознавания именованных сущностей. Это превосходство обусловлено их способностью извлекать глубокие семантические и синтаксические представления из контекста без необходимости ручного конструирования признаков, а использование предобученных на больших корпусах языковых моделей с последующей их тонкой настройкой на сравнительно небольших размеченных датасетах позволяет достигать лучших результатов, минимизируя затраты на разметку данных.

Перспективы дальнейших исследований в области NER с использованием нейронных сетей видятся в разработке более эффективных и компактных архитектур, адаптации моделей для низкоресурсных языков, а также в интеграции внешних знаний для улучшения распознавания редких и составных сущностей.

Список литературы:

1. Гольдберг Й. Нейросетевые методы в обработке естественного языка. / 2. Гольдберг Й. - Москва: ДМК Пресс, 2019. - 282 с.

2. Хабр [Электронный ресурс]. // URL: https://habr.com/ru/articles/776774/ (дата обращения: 18.11.25)

3. Большакова Е.И. Автоматическая обработка текстов на естественном языке и анализ данных. / Большакова Е.И., Воронцов К.В., Ефремова Н.Э.,

Клышинский Э.С., Лукашевич Н.В., Сапин А.С. - Москва: НИУ ВШЭ, 2017. - 269 с.