АНАЛИЗ МЕТОДОВ РАСПОЗНАВАНИЯ ПРОФЕССИЙ В РУССКОЯЗЫЧНЫХ ТЕКСТАХ

Иванова М. Д.
Факультет прикладной математики и информатики, Белорусский государственный университет
Минск, Республика Беларусь
E-mail: ivanova.mary76@gmail.com

В работе рассматривается задача распознавания именованных сущностей категории «Профессия» в русскоязычных текстах. Целью исследования является определение наиболее эффективного подхода к извлечению профессий из неструктурированных данных. Проведён сравнительный анализ трёх различных методов: основанных на знаниях (словарь), машинном обучении (CRF) и глубоком обучении (RuBERT). Эксперименты проведены на специально разработанном размеченном корпусе данных. Результаты могут быть использованы для интеллектуального анализа вакансий и данных о рынке труда.

Введение

Распознавание именованных сущностей (Named Entity Recognition, NER) категории «Профессия» представляет собой важное направление в области автоматического анализа текстов, поскольку позволяет извлекать из неструктурированных данных сведения о профессиональной принадлежности и трудовой деятельности людей. Такая информация имеет практическую ценность для широкого круга прикладных задач: автоматического анализа резюме, мониторинга рынка труда, социолингвистических исследований.

В отличие от стандартных категорий именованных сущностей, таких как имена собственные, организации или географические объекты, категория «Профессия» остаётся относительно малоизученной и недостаточно представленной в существующих системах распознавания. Большинство известных NER-моделей не предусматривают выделение профессий как отдельного класса. Дополнительные трудности при обработке русскоязычных текстов создают морфологическое разнообразие наименований профессий и контекстная зависимость их употребления.

Использование универсальных моделей распознавания именованных сущностей, обученных на общих корпусах, не обеспечивает достаточного качества для распознавания профессий, поскольку такие модели не учитывают доменную специфику и контекст употребления профессиональной лексики. В результате возникает необходимость в разработке специализированных решений, ориентированных на данную категорию.

Целью данного исследования является сравнительный анализ различных подходов к распознаванию сущностей категории «Профессия» в русскоязычных текстах и определение метода, обеспечивающего наилучшее качество идентификации. Для достижения этой цели были реализованы и протестированы три подхода: словарный метод, статистическая модель на основе условных случайных полей (Conditional Random Fields, CRF) и нейросетевой метод с использованием предобученной модели RuBERT, адаптированной к доменной задаче путём дообучения на специализированном корпусе данных.

I. Классификация подходов для распознавания именованных сущностей

Подходы для распознавания именованных сущностей можно классифицировать на три основные группы, отражающие их историческую эволюцию [1]. Подробное описание приведено далее.

Методы, основанные на знаниях. Исторически первые системы распознавания именованных сущностей базировались на словарях, лингвистических правилах и регулярных выражениях. Эти методы основаны на экспертном знании предметной области и формализации языковых закономерностей. Они обеспечивают высокую точность в узких доменах и прозрачность интерпретации результатов, однако требуют значительных трудозатрат при создании и адаптации, а также плохо масштабируются на новые области и не способны к обобщению непредусмотренных языковых форм.

Методы, основанные на конструировании признаков. Следующий этап развития NER связан с применением методов машинного обучения. Наиболее эффективными в этой категории показали себя условные случайные поля. Ключевое преимущество CRF перед другими алгоритмами машинного обучения заключается в способности эффективно моделировать зависимости между наблюдениями и метками, учитывая контекстные взаимосвязи [2]. Таким образом, CRF способны обобщать закономерности на основе обучающих данных и сохранять относительную прозрачность модели. Однако их качество существенно зависит от ручного конструирования признаков и полноты обучающего корпуса.

Методы, основанные на глубоком обучении. Современный этап развития NER связан с применением глубоких нейронных сетей. Нейронные сети позволяют автоматически извлекать репрезентативные признаки сущностей из больших наборов данных без явного ручного конструи-

рования. Они способны выявлять сложные лингвистические закономерности и демонстрируют высокую адаптивность к различным языковым контекстам.

II. Методология исследования

Корпус данных. Для проведения экспериментов был создан специализированный корпус русскоязычных текстов, содержащий аннотированные упоминания профессий. Корпус был сформирован на основе текстов, собранных с онлайн-платформ по поиску работы, профессиональных форумов, объявлений о вакансиях и резюме, а также дополнен синтетическими данными.

Модели и методы. В исследовании сравниваются три подхода:

1. Словарный подход. Словарный подход основан на использовании специально разработанного словаря профессий, который включает несколько тысяч наименований профессий с учётом морфологических вариаций.
2. CRF. В качестве представителя подхода, основанного на конструировании признаков, используется модель CRF. Модель учитывает последовательные зависимости между токенами и использует совокупность лексических, морфологических, орфографических и контекстуальных признаков.
3. RuBERT. В качестве нейросетевого решения применяется предобученная языковая модель RuBERT, адаптированная к задаче распознавания профессий путём дополнительного обучения на специально размеченном корпусе русскоязычных текстов.

Метрики оценки. Качество моделей оценивалось с помощью стандартных метрик: $Precision = \frac{TP}{TP+FP}$; $Recall = \frac{TP}{TP+FN}$; $F1 = 2 * \frac{Precision*Recall}{Precision+Recall}$, где TP, FP и FN обозначают истинно положительные, ложно положительные и ложно отрицательные случаи распознавания соответственно [3].

III. Результаты

Таблица 1 демонстрирует сравнительные результаты для трёх методов: словарного подхода, CRF и RuBERT.

Таблица 1 – Сравнение методов распознавания профессий

Метод	**Precision**	**Recall**	**F1-score**
Словарь	0,98	0,55	0,70
CRF	0,84	0,72	0,78
RuBERT	0,87	0,85	0,86

Результаты показывают, что словарный подход обеспечивает высокую точность за счёт жёстких правил, но не охватывает вариативность естественного языка. CRF-модель достигает более сбалансированных показателей за счёт учёта контекста, однако её эффективность сильно зависит от качества используемых признаков.

Модель RuBERT демонстрирует наилучший результат среди рассмотренных подходов. Она превышает CRF по F1-мере на 8%. RuBERT корректно обрабатывает многословные профессии и учитывает морфологическую вариативность. Качественный анализ ошибок показал, что оставшиеся проблемы связаны с редкими или новыми наименованиями профессий, метафорическим употреблением и шумом в обучающих данных.

Заключение

Проведённый сравнительный анализ подтвердил гипотезу о необходимости специализированных решений для идентификации сущностей категории «Профессия» в русскоязычных текстах. Эволюция подходов от методов, основанных на знаниях, к методам, основанным на глубоком обучении, демонстрирует устойчивое улучшение качества.

Тонкая настройка предобученной модели RuBERT показала наилучший результат по F1-мере (0,86). Данная модель демонстрирует высокую способность к обобщению и возможность эффективно обрабатывать сложные случаи: многокомпонентные наименования, морфологическую вариативность и контекстно-зависимые упоминания профессий.

Практическая ценность работы заключается в возможности применения разработанных решений в системах анализа текстовых данных: при мониторинге рынка труда, обработке вакансий и резюме, а также в социологических исследованиях. Дальнейшее развитие исследования предполагает расширение корпуса данных, совершенствование методов разметки и создание более лёгких и энергоэффективных моделей, пригодных для внедрения в реальные информационные системы. Таким образом, исследование способствует развитию технологий обработки естественного языка и подтверждает эффективность нейросетевых моделей для решения прикладных задач извлечения информации.

1. Keraghel, I. Recent Advances in Named Entity Recognition: A Comprehensive Survey and Comparative Study / I. Keraghel, S. Morbieu, M. Nadif // arXiv preprint arXiv:2401.10825. — 2024.
2. Cariello, M. C. A Comparison between Named Entity Recognition Models in the Biomedical Domain / M. C. Cariello, A. Lenci, R. Mitkov // Proceedings of the Translation and Interpreting Technology Online Conference. – 2021. – P. 76–84.
3. Метрики классификации и регрессии [Электронный ресурс]. – Режим доступа: `https://education.yandex.ru/handbook/ml/article/metriki-klassifikacii-i-regressii`. – Дата доступа: 24.10.2025.