УДК 004.8

ИСПОЛЬЗОВАНИЕ МАШИННОГО ОБУЧЕНИЯ ДЛЯ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ СТРУКТУРНЫХ ЭЛЕМЕНТОВ НАУЧНЫХ ПУБЛИКАЦИЙ

А.Р. Латипов
Институт проблем управления им. В.А. Трапезникова РАН
Россия, 117997, Москва, Профсоюзная ул., 65
E-mail: latipov257@gmail.com

М.Р. Блашкун
Институт проблем управления им. В.А. Трапезникова РАН
Россия, 117997, Москва, Профсоюзная ул., 65
E-mail: mblashkun@gmail.com

П.А. Кирьянов
Институт проблем управления им. В.А. Трапезникова РАН
Россия, 117997, Москва, Профсоюзная ул., 65
E-mail: lovecoldreams@gmail.com

Ключевые слова: машинное обучение, распознавание структурных элементов, LSTM, GROBID.

Аннотация: Проводится сравнительный анализ применения машинного обучения для автоматического распознавания структурных элементов научных публикаций, с основным фокусом на сравнении машинной модели LSTM (Long Short-Term Memory) с инструментом GROBID, использующим модель Wapiti. Рассматриваются технические характеристики и преимущества каждого метода, а также оценивается их эффективность в контексте точности распознавания структурных элементов. Анализируются результаты экспериментов, проведенных с использованием различных наборов данных, для выявления сильных и слабых сторон каждой модели. Данные для обучения моделей были вручную размечены по структурным элементам в виде тегов, включающих в себя содержание соответствующих элементов.

1. Введение

Необходимо систематизировать и нормализовать данные научных публикаций для создания базы данных с возможностью поиска, построения графов цитирования и использования размеченных данных в обучении языковых моделей. Однако разнообразие структур самих публикаций представляет основную проблему. Использование эвристических алгоритмов, таких как регулярные выражения, может быть решением, но требует разработки множества правил для учета особенностей различных типов структурных элементов. Другие методы, ориентированные на машинное обучение, обеспечивают высокую точность, зависящую от качества данных и выбранной модели.

2. Библиотека GROBID

GROBID [1] (англ. GeneRatiOn of BIbliographic Data) – это библиотека машинного обучения для извлечения, разбора и реструктуризации необработанных документов, таких как PDF, в структурированные документы в кодировке XML/TEI с особым акцентом на технические и научные публикации. GROBID широко используется надежным базовым инструментом для извлечения содержимого структурных элементов из научных статей. Этот инструмент существует уже более десятилетия и считается стандартным инструментом как в академических, так и промышленных кругах [2]. На рис. 1 представлен каскад моделей GROBID.

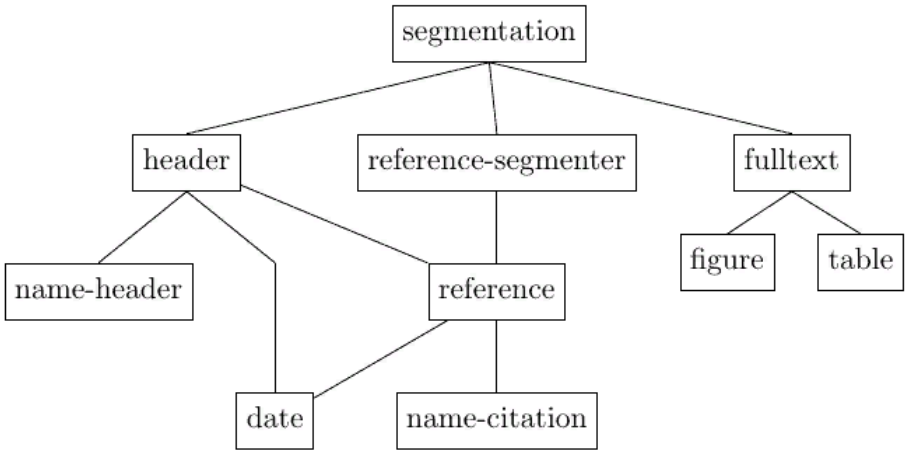

Рис. 1. Каскад моделей GROBID.

3. LSTM-модель

В качестве альтернативного решения была разработана модель, основывающаяся на сетях долгой краткосрочной памяти (LSTM-модель). Изначально была выбрана модель BiLSTM, но в результате тестирования и сравнения с более простой LSTM было выявлено, что первая модель не приносит значительного прироста в точности, при этом требуя значительно больше ресурсов и времени вычислений (время обучения и работы BiLSTM превышает время обучения и работы LSTM примерно на 10-15%). Поэтому в качестве основы была выбрана модель LSTM.

Модель долгосрочной краткосрочной памяти является типом рекуррентной нейронной сети (RNN), спроектированной для эффективной работы с последовательными данными и учета долгосрочных зависимостей в этих данных. Научные публикации обычно имеют структурированный формат с различными разделами, и LSTM хорошо подходят для обработки таких последовательных данных, так как они сохраняют информацию о предыдущих шагах и могут принимать решения, учитывая контекст. Также LSTM способны улавливать долгосрочные зависимости между словами и предложениями. В научных текстах важно учитывать не только текущий контекст, но и связи между различными частями текста, которые могут быть удалены друг от друга.

4. Обучение

4.1 Обучение модели GROBID

Общее количество статей в базе составляло 16 662. После исключения файлов с шифрованием, поврежденной кодировкой и отсутствием текстового слоя, осталось 14 434 статьи, включая 10 461 статью на русском языке.

Метод обучения библиотеки GROBID, рассмотренный в данной работе, основывается на предположении о стабильности структуры статей журналов в течение времени. В соответствии с этим предположением произведен анализ статей из разных источников, с последующей группировкой и выявлением наиболее часто встречающихся шаблонов. Начальная фаза обучения включает 400 таких шаблонов, применяемых для обучения моделей, включая header, segmentation, affiliation-address, name-header и reference-segmenter.

В ходе экспериментов использовались PDF-файлы статей для создания обучающих данных при помощи инструмента GROBID. Внесение корректировок вызывало трудности из-за ограничений в методике исправления обучающих файлов. Если в обучающем файле отсутствовала часть текста из статьи, ее добавление было невозможно, что делало данный файл не пригодным для обучения. В результате некоторые шаблоны были исключены из тренировочного набора в связи с вышеуказанными ограничениями.

4.2 Обучение модели LSTM

В ходе обучения за элементарную единицу данных было взято слово. На этом уровне были проведены следующие операции:

- преобразование слов в нижний регистр;
- лемматизация (приведение слов к нормальной форме);
- удаление стоп-слов (слова, знаки, символы, которые самостоятельно не несут никакой смысловой нагрузки);
- подготовка общего словаря для всех документов;
- преобразование слов в вектора (с помощью фреймворка pytorch), с которыми умеет работать модель.

Проведена разметка, результат которой – txt файлы, которые содержат первые 1000 слов статей, и файлы json с конечной разметкой по данным из txt файла. Всего получилось 628 пар.

Набор подготовленных данных был разбит на обучающую и валидационную выборку в соотношении 99 к 1. Для обучения был выбран алгоритм поиска минимума ADAM с параметром learning rate = $1e^{-3}$. В качестве функции потерь была выбрана функция NLL loss (The negative log likelihood loss). Обучение длилось 10 эпох.

5. Сравнение моделей

5.1 GROBID

После завершения процесса обучения были получены значения метрики f1 (показанные в таблице 1), которые отражают эффективность моделей.

Таблица 1. Метрика f1 для модели GROBID.

Модель GROBID	f1 all (micro avg.)	f1 all (macro avg.)

header	80,95	73,25
segmentation	83,95	70,33
affiliation-address	82,71	78,29
name-header	93,97	87,71
reference-segmenter	94,55	95,25

Имея доступ к заголовкам и аннотациям в базе статей, был проведен анализ сходства с применением метрик, таких как расстояние Джаро-Винклера, Левенштейна и косинусное расстояние.

Анализ результатов подтвердил высокую точность соответствия заголовков и аннотаций в рамках выбранных метрик. В частности, модели, такие как header и segmentation, достигли значительных успехов. Тем не менее, другие модели, в частности affiliation-address, name-header и reference-segmenter, выявили потребность в дополнительном улучшении для достижения желаемых результатов.

5.2 LSTM-модель

Также была проведена оценка качества работы модели на валидационной выборке. Получена следующая таблица:

Таблица 2. Метрики precision, recall, f1 для LSTM-модели.

Структурный элемент	precision	recall	f1
author_position	1.00	1.00	1.00
affiliations	0.97	1.00	0.98
author_degree	1.00	1.00	1.00
annotation	1.00	1.00	1.00
author_mail	1.00	1.00	1.00
udc	1.00	1.00	1.00
keywords	1.00	1.00	1.00

author_name	1.00	0.95	0.97
title	1.00	1.00	1.00
address	1.00	1.00	1.00

Как видно из таблицы, LSTM-модель в среднем хуже справляется с классификацией следующих меток:

- «affiliations» (аффилиации) - так как в аффилиации могут присутствовать слова, которые имеют географическое значение. Например, в «**Московский** технологический университет» слово московский, из-за схожести со словом Москва классифицируется нейронной сетью как адрес;
- «author_name» (ФИО автора) - ввиду особенности алгоритма по разделению текста на предложения, ФИО определяется за отдельное предложение, поэтому нейронной сети становиться труднее корректно классифицировать ФИО ввиду отсутствия контекста.

6. Выводы

С учетом выявленных потребностей в доработке, планируется активно продолжать обучение моделей, сосредотачиваясь, в первую очередь, на тех, которые продемонстрировали менее удовлетворительные результаты. При этом также рассматриваются возможные улучшения в методологии, направленные на повышение эффективности обучения и достижение более высоких показателей точности.

В контексте общей значимости исследования, следует подчеркнуть его вклад в область обработки текста и структурирования научных статей на русском языке. Работа направлена на решение актуальных проблем в этой области, в том числе повышение точности и эффективности инструментов автоматической обработки текста на русском языке, что может значительно облегчить работу исследователей и обеспечить более точные результаты в анализе научных материалов.

Список литературы

1. https://github.com/kermitt2/grobid (дата обращения 20.12.2023).
2. Lipinski M., Yao K., Breitinger C., Beel J., Gipp B. Evaluation of header metadata extraction approaches and tools for scientific PDF documents // 2013. P. 385-386.