Применение нейросетевых методов извлечения ключевых слов для составления резюме студента по рабочим программам

М. В. Сорочина[1], П. В. Корытов[2], И. И. Холод[3]
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)
[1]mvsorochina@stud.eltech.ru, [2]thexcloud@gmail.com, [3]iiholod@etu.ru

Аннотация. **В рамках работы рассматриваются применение методов обработки естественного языка (NLP) для решения задачи выявления основных навыков, полученных студентами в процессе обучения. Описано применение ансамбля из моделей WikiNEuRal, KBIR-inspec, ruGPT3Large и word2vec, а также других методов NLP, для извлечения и обработки ключевых слов на русском языке. Описанный метод разработан для автоматизации составления резюме студента в ИС «Личный кабинет партнера» СПбГЭТУ «ЛЭТИ».**

Ключевые слова: обработка естественного языка, ключевые слова, резюме

I. Введение

У ВУЗов имеется большой объем информации о студентах, в том числе насколько успешно были пройдены дисциплины, включенные в учебный план. По большей части, информация о студентах хранится в формате документов на естественном языке. Также, университет может предложить студентам много активностей, таких как участие в конференциях, олимпиадах или соревнованиях, дополнительных занятия по множеству направлений. Кроме того, учебные заведения имеют вакансии, подходящие студентам, как внутри заведения, так и вакансии от компаний-партнеров.

Качество рекомендаций данных активностей студентам можно повысить, если учитывать особенности каждого отдельного студента, а не только целой группы или направления. Для решения этой задачи в данной работе предлагается с помощью методов обработки естественного языка составить резюме студента – выявить основные навыки, полученные в ходе обучения. Навыки включают в себя изученные технологии, понятия, методы и другие сущности.

Сформированный список навыков является основой для цифрового профиля студента, который учащийся может редактировать на свое усмотрение. Данный профиль играет роль резюме студента в информационных системах вуза. С помощью профиля также может быть решена задача рекомендации активностей, рассмотрение которой выходит за рамки данной работы.

II. Возможные подходы к решению

Для выявления основных навыков необходимо извлечь подходящие слова или фразы из текста. При решении поставленной задачи может быть применен подход распознавания именованных сущностей (Named entity recognition, NER). В этом случае из текста извлекаются объекты определенного класса, такие как имена людей, названия мест и организаций и другие.

Также может быть применен подход извлечения ключевых слов и фраз. Как и многие другие методы NLP, данный подход имеет несколько подвидов: классические методы и нейросетевые. Классические могут быть статистическими, то есть использовать такие эвристики, как позиция слова в предложении, капитализация букв, позиция предложения в тексте и другие. Кроме того, к классическим относятся графовые методы, которые представляют текст в виде графа, где слова становятся вершинами, а ребра представляют связи между ними. После чего вычисляются веса вершин, исходя из которых выбираются ключевые слова. Представители указанных методов будут приведены далее.

Особенность нейросетевых методов заключается в использовании глубокого обучения для определения ключевых слов. Одной из наиболее распространенных архитектур стала «трансформер» [1]. Трансформеры позволяют использовать предобученные модели, представленные в HuggingFace Hub [2]. Эти методы способны решить некоторые задачи, которые невозможно решить при помощи более простых классических методов, например, учитывать семантику слов. Нейросетевые методы также зачастую показывают результаты на бенчмарках лучше, чем статистические методы [3]. Поэтому было принято решение использовать их.

В отличие от задачи NER, ключевые слова могут быть извлечены разными способами. К тому же ключевые слова могут быть извлечены из более широкого набора текстов, тогда как в случае NER извлекаемые классы могут быть нерепрезентативны или вовсе не присутствовать в тексте. Кроме того, при обучении модели NER извлечению новых классов требуется большой набор размеченных данных. Поэтому можно комбинировать NER и извлечение ключевых слов для достижения желаемого результата.

В процессе поиска метода решения были рассмотрены несколько моделей. Среди них были как классические, так и нейросетевые:

- PositionRank [4] – графовая модель извлечения ключевых фраз. Учитывает не только частоту появления слова в документе, но и его позицию – больший вес получают слова, которые встречаются в начале документа.

- YAKE! [5] – статистическая модель извлечения ключевых слов, которая является State-of-the-Art из классических методов. Оценка одного термина вычисляется по формуле:

$$S(t) = \frac{T_{Rel} \cdot T_{Position}}{T_{Case} + \frac{TF_{Norm}}{T_{Rel}} + \frac{T_{Sentence}}{T_{Rel}}}$$

где:

T_{Case} учитывает регистр слова,

$T_{Position}$ учитывает положение слова в документе, TF_{Norm} является нормализованной частотой термина,

T_{Rel} учитывает связь термина с контекстом,

$T_{Sentence}$ учитывает частота появления термина в разных предложениях.

- KeyBERT [6] – нейросетевая модель извлечения ключевых слов и фраз. В реализации по умолчанию используется модель на основе BERT, но возможно использование других предобученных моделей с репозитория HuggingFace, в т. ч. многоязыковых. Для выявления кандидатов используется косинусное подобие и оптимизация максимальной предельной релевантности.
- WikiNEuRal [7] – нейросетевая модель для распознавания именованных сущностей – находит в тексте слова, называющие людей (PER), мест (LOC), организации (ORG) и другие сущности (MISC). Модель обучена на текстах Wikipedia и поддерживает 9 языков, включая русский.
- KBIR-inspec [3] – нейросетевая модель извлечения ключевых слов, основанная на KBIR и настроенная на датасете Inspec [8]. Поддерживает английский язык.

III. Описание метода решения

Рассматриваемый набор данных состоит из рабочих программ, т. е. текстовых описаний дисциплин, читающихся в университете. Решение задачи включает в себя несколько этапов:

- Извлечение ключевых фраз и слов.
- Приведение полученных фраз на русском языке к «нормальной форме».
- Поиск схожих по смыслу фраз для уменьшения дублирования в итоговом резюме.

Схема этапов представлена на рис. 1.

A. Извлечение именованных сущностей

Для извлечения именованных сущностей (NER) была использована модель WikiNEuRal [7].

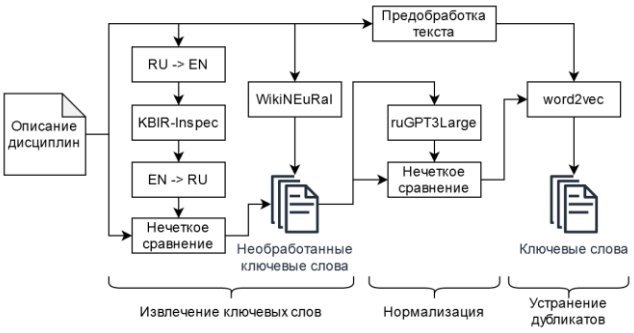

Рис. 1. Схема метода решения

На рассматриваемом наборе данных WikiNEuRal смогла распознать как класс MISC некоторые языки программирования, программы, библиотеки и прочие технологий. Например, из описания дисциплины «Web-технологии» были извлечены такие сущности, как «HTML», «CSS», «HTTP», «JavaScript» и другие технологии.

Также ею были извлечены названия некоторых методов, именованных в честь ученых (класс PER). Из дисциплины «Теория оптимального управления» были извлечены фразы «Нелдера-Мида» и «Коши».

Предобработка текстов не требовалась, так как рабочие программы содержат исключительно русский и английский языки, которые входят в список поддерживаемых моделью языков.

Однако, в силу специфики вуза из текстов стабильно извлекается фамилия «Ленин» и государство «Российская Федерация». В качестве временного решения этой проблемы выход этого метода проверяется по черному списку.

Основной недостаток данного подхода в том, что его можно успешно применить только к подмножеству рассматриваемого набора данных. Субъективно осмысленные результаты были получены на 69 из 158 случайно выбранных рабочих программ.

B. Извлечение ключевых фраз

Из-за описанной ограниченности результатов NER имеет смысл параллельно использовать более «широкие» методы, т.е. методы извлечения ключевых фраз. Модель KBIR-Inspec [3] показала хорошие результаты на датасете Inspec [8], однако она напрямую неприменима для текстов на русском языке. Поэтому предложена следующая схема адаптации модели на русский язык.

Перед использованием модели тексты переводятся на английский. Затем модель извлекает ключевые фразы из подготовленных текстов, и полученный набор фраз переводится обратно с английского на русский. Поскольку второй перевод может быть неточным из-за отсутствия контекста, производится фильтрация фраз при помощи нечеткого поиска (fuzzy search) в исходном тексте с использованием библиотеки thefuzz [9].

Таким образом, список навыков был дополнен фразами, не содержащими именованные сущности. Например, для дисциплины «Алгоритмы беспилотного транспорта» были найдены такие ключевые слова, как «автономные интеллектуальные системы», «машинное обучение», «нейросети», «алгоритм семантического сегментирования».

C. *Приведение полученных фраз к нормальной форме*

Из-за специфики русского языка извлекаемые ключевые фразы остаются в форме, определенном контекстом – могут иметь множественное число, падеж, отличный от именительного, и т. п. Например, из предложения «студент обучается программированию на языке JavaScript» извлекается ключевая фраза «программированию на языке JavaScript».

Поскольку результат планируется отображать пользователям, разумно попробовать привести фразы к нормальной форме, т. е. переформулировать данный пример в «программирование на языке JavaScript».

Для решения этой задачи была выбрана модель ruGPT3Large [10]. Модель предназначена для генерации текста на русском языке. Для ее использования в целях нормализации был создан шаблон следующего вида:

магазины приложений -> магазин приложений

человеко-машинного интерфейса -> человеко-машинный интерфейс

принципах взаимодействия нейронов -> принципы взаимодействия нейронов

отладки приложений -> отладка приложений

содержанием дисциплины -> содержание дисциплины

…

[слово для нормализации] ->

Примеры фраз, нормализованных таким образом:

- «объекта управления» → «объект управления»
- «классических задач оптимизации» → «классические задачи оптимизации»
- «алгоритмов теории управления» → «алгоритмы теории управления»

Однако, такой подход также иногда дает очевидно плохие результаты. Например, фраза «показатели управления БП» была преобразована во фразу «q-критерий Фишера».

Для решения этой проблемы применен метод нечеткого сравнения строк, аналогичный п. «извлечение ключевых фраз». Если расстояние между нормализованной и исходной ключевой фразой слишком велико, нормализованная фраза отбрасывается.

D. *Поиск схожих по смыслу фраз*

На данном этапе мы имеем набор частично нормализованных ключевых фраз. Следующая проблема заключается в том, что при написании текстов рабочих программ часто используются синонимы. Например, для дисциплины «Базы данных» выделяется как «реляционные системы управления базами данных», так и аббревиатура «рсубд». Наиболее «тяжелые» случаи связаны с фразами, полностью отличными по написанию: например, для дисциплины «проектирование человеко-машинного интерфейса» это «разработка интерфейса пользователя» и «проектирование экранных форм».

В качестве основного метода для решения этой проблемы была выбрана модель word2vec [11]. Данная модель отображает слова и фразы в векторы, причем схожие по смыслу элементы находятся рядом в пространстве признаков.

Модель была обучена на тексте, содержащем все описания дисциплин. Чтобы обучение дало лучший результат, текст был предобработан – из него были удалены все запятые, тире, точки в конце предложения и другие символы, не изменяющие смысла слов.

После на множестве векторных представлений ключевых слов применяется алгоритм кластеризации DBSCAN.

Отладка данного шага решения является частью дальнейшей работы: сейчас в одном кластере могут оказаться слова, которые нужно различать: например, «pytorch» и «tensorflow», «java» и «c++».

IV. Заключение

В ходе работы были рассмотрены методы обработки естественного языка для формирования набора навыков для резюме студента. Для решения задачи были выбраны нейросетевые методы.

По сравнению с предыдущей работой [12], извлекающей ключевые слова с помощью классических методов NLP, использование NER (WikiNEuRal) позволяет дать исчерпывающее перечисление определенных классов сущностей, в частности названий библиотек и языков программирования. Предположительно, это повысит качество автоматического составления резюме для студентов IT-специальностей.

Дальнейшая работа на шаге извлечения ключевых слов будет включать в себя разработку более объективного способа оценки результатов вместо приведенных субъективных оценок. Возможно, на этапе поиска дубликатов имеет смысл ввести механизм «доверия» источникам ключевых слов: более «лояльно» относится к выводу методов NER, меньше доверять нейросетевым методам извлечения ключевых слов, ещё меньше графовым или статистическим.

Русскоязычная версия модели GPT-3 (ruGPT3Large) применена для приведения ключевых фраз в «нормальную форму». Дальнейшая работа на этом шаге также включает в себя оценку эффективности и испытание более современных генеративных моделей, например LLaMA [13].

Итоговое решение будет внедрено в ИС «Личный кабинет партнера» для применения с методами, обеспечивающими сопоставление извлеченных ключевых слов с текстами активностей – вакансий, мероприятий, проектов и т. п. [12].

Список литературы

[1] Tunstall L., Von Werra L., Wolf T. Natural language processing with transformers. O'Reilly Media, Inc., 2022, 571 с.

[2] Wolf T. и др. HuggingFace's Transformers: State-of-the-art Natural Language Processing // arXiv preprint arXiv:1910.03771. 2020.

[3] Kulkarni M., Mahata D., Arora R., Bhowmik R. Learning Rich Representation of Keyphrases from Text // arXiv preprint arXiv:2112.08547. 2021.

[4] Florescu C., Caragea C. Positionrank: An unsupervised approach to keyphrase extraction from scholarly documents //Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: long papers). 2017. С. 1105-1115.

[5] Campos R. et al. YAKE! Keyword extraction from single documents using multiple local features // Information Sciences. 2020. Т. 509. С. 257-289.

[6] Grootendorst M. KeyBERT: Minimal keyword extraction with BERT. URL: https://github.com/MaartenGr/KeyBERT (дата обр. 11.03.2023)

[7] Tedeschi S., Maiorca V., Campolungo N., Cecconi F. Navigli R. WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER // Findings of the Association for Computational Linguistics: EMNLP 2021. 2021. С. 2521-2533.

[8] Hulth A. Improved automatic keyword extraction given more linguistic knowledge // Proceedings of the 2003 conference on Empirical methods in natural language processing. Association for Computational Linguistics, 2003. С. 216–223.

[9] seatgeek/thefuzz: Fuzzy String Matching in Python. — URL: https://github.com/seatgeek/thefuzz (дата обр. 11.03.2023).

[10] ruGPT-3 Large модель генерации текста на русском языке с 760 миллионами параметров. URL: https://sbercloud.ru/ru/datahub/rugpt3family/rugpt-3-large (дата обр. 11.03.2023)

[11] Mikolov T., Chen K., Corrado G., Jeffrey D. Efficient estimation of word representations in vector space // arXiv preprint arXiv:1301.3781. 2013.

[12] Korytov P. V., Kholod I. I. Application of Text Analysis Methods to Recommend Student Choices // 2022 XXV International Conference on Soft Computing and Measurements (SCM). СПб.: IEEE, 2022. С. 107–110.

[13] Touvron H. et al. LLaMA: Open and Efficient Foundation Language Models // arXiv preprint arXiv:2302.13971. 2023.