УДК 004.81:004.85 *Научная статья* DOI: 10.18287/2223-9537-2023-13-2-282-293

Метод извлечения знаний и навыков/компетенций из текстов требований вакансий

© 2023, И.Е. Николаев

Челябинский государственный университет, Челябинск, Россия

Аннотация

Анализ требований вакансий на рынке труда показывает, что они представляют собой многоуровневые языковые конструкции из нескольких слов со сложными семантическими связями. Целью исследования является разработка метода извлечения коротких текстов знаний и навыков/компетенций из текстов требований, имеющих сложную организационную структуру. Метод состоит в дополнении структуры сложных предложений новыми связями посредством дообученной на текстах онлайн-вакансий нейросетевой модели *BERT* и перехода от сложного текста к набору простых комбинаций слов. Показан процесс дообучения нейросетевых моделей *BERT* от лаборатории по искусственному интеллекту Сбербанка на текстах онлайн-вакансий. Реализованы два механизма добавления новых связей между словами требований с учётом знаний из предметной области: линейный и через дополнение дерева синтаксического разбора. Проведён сравнительный анализ для нескольких комбинаций инструментов. Наилучший результат показала комбинация: дообученная модель *BERT* плюс *deeppavlov_syntax_parser* плюс линейный способ дополнения связей. Применимость метода проверена на текстовом корпусе требований онлайн-вакансий. Предложенный метод показал более высокую эффективность, чем подход, основанный на правилах, который предполагает использование формальных правил и правил грамматики для анализа естественного языка. Использование метода позволяет оперативно определять ключевые изменения потребностей рынка труда на уровне текстов требований отдельных знаний и навыков/компетенций.

Ключевые слова: *нейросетевые модели, дообучение языковых моделей, синтаксический анализатор, синтаксические деревья, рынка труда, компетенции, вакансии.*

Цитирование: *Николаев И.Е.* Метод извлечения знаний и навыков/компетенций из текстов требований вакансий // Онтология проектирования. 2023. Т.13, №2(48). С.282-293. DOI:10.18287/2223-9537-2023-13-2-282-293.

Введение

Динамичное развитие многих отраслей экономики, изменение форм занятости и переход многих сфер деятельности на удалённый формат работы, замена живого общения на взаимодействие с цифровыми устройствами, а также изменения, происходящие в сфере образования и подготовки кадров, существенно изменяют структуру глобального рынка труда.

Эти изменения приводят к тому, что особую значимость приобретают исследования, направленные на создание инструментов оперативного мониторинга рынка труда на уровне отдельных *знаний, навыков и компетенций* на основе анализа открытых данных систем поиска и подбора работников с помощью Интернет-ресурсов, и в первую очередь текстовой информации из описаний вакансий.

Разработка и внедрение таких инструментов позволит оперативно определять ключевые изменения потребностей рынка труда, повысить эффективность управленческих решений по созданию образовательных программ подготовки, переподготовки и повышения квалификации кадров, помогут всем участникам рынка точнее оценивать существующие и зарождающиеся тенденции на рынке труда.

В контексте требований к вакансиям можно использовать термин «сущность» для обозначения того, что является наиболее важным аспектом требований к кандидатам. Как правило, требования вакансии представляют собой набор таких сущностей, которые можно разделить на *знания, навыки и компетенции*, необходимые для выполнения задач профессиональной деятельности.

Знание – это сущность, которая представляет собой наличие у человека информации об определённой области знаний. Знание может быть абстрактным и теоретическим, в отличие, например, от навыков. Примером знания может служить «знание программирования на языке *Java*», которое может не применяться на практике, но при этом является важным фактором при найме на должность разработчика программного обеспечения.

Навыки – это сущность, которая описывает уровень умения человека выполнять определённые действия в рамках определённой области знаний. Навыки могут быть связаны с конкретными технологиями или инструментами, их наличие свидетельствует о том, что кандидат способен выполнить задачу. Например, «навыки программирования на языке *Java*» являются конкретным умением разрабатывать программное обеспечение на этом языке.

Компетенции – это сущность, которая описывает способность человека применять знания и навыки в профессиональной деятельности. Компетенции включают в себя комплекс навыков, без которых невозможно эффективное выполнение работы. Компетенции могут включать коммуникационные навыки, умение работать в команде, аналитические навыки, способность к принятию взвешенных решений в сложных ситуациях и др. [1].

Разработка методов выделения сущностей *знаний*, *навыков* и *компетенций* из текстов требований вакансий играет ключевую роль при переходе от статистического анализа рынка труда к анализу на уровне отдельных сущностей. В статье отсутствует разделение между сущностями *навыки* и *компетенции*, т.к. на уровне коротких текстов данные категории практически неразличимы.

Для выделения и анализа сущностей, связанных с требованиями к кандидатам в вакансиях, можно использовать таксономии и онтологии. Таксономия — это иерархия понятий, которые логически связаны между собой. Онтология — это модель знаний, которая описывает концептуальные отношения между объектами в предметной области (ПрО).

Существующие методы извлечения сущностей способны её обнаружить, если искомая сущность представлена в исходном тексте в виде подряд идущих слов. Данное обстоятельство снижает эффективность использования таких методов на текстах требований рынка труда, т.к. в большинстве случаев такие тексты представляют собой сложно организованные конструкции [2].

В настоящее время перспективным направлением анализа текстов на естественном языке являются нейросетевые модели, построенные на трансформерах с использованием механизма внимания, за счёт которого такие модели получили способность выявлять взаимосвязи между частями текста [5]. Первой такой моделью стала языковая модель *BERT* (англ. *Bidirectional Encoder Representations from Transformers*) [6]. *BERT* предназначена для предобучения языковых представлений с целью их последующего применения в различных задачах обработки естественного языка. Особенность *BERT* заключается в том, что эта модель учитывает контекст для всех слов и способна успешно работать с долгосрочными зависимостями в тексте и долго удерживать информацию о контексте для каждого слова [7].

В настоящее время *BERT* и её производные (*RoBERTa* [8], *GPT* [9], *T5* [10], *XLM* [11]) показывают наилучший результат на большинстве задач обработки и анализа текстов на естественных языках и превосходит нейросетевые модели предыдущих поколений (*word2vec, LSTM* и др.). Модели на трансформерах универсальны и способны извлекать из текста признаки, полезные для решения множества задач текстового анализа. К недостаткам таких мо-

делей можно отнести требовательность к ресурсам и длительный процесс обучения. Их эффективность может быть улучшена путём дообучения существующих моделей на текстах определённой ПрО. В работах [12] показано, что дообучение моделей существенно повышает качество анализа текстов для соответствующих ПрО.

Помимо нейросетевых моделей для извлечения структурированной информации из текстов на естественном языке также используются синтаксические анализаторы и анализаторы на основе контекстно-свободных грамматик.

На конференции по компьютерной лингвистике были проведены соревнования по полной грамматической разметке текстов на русском языке [15]. По результатам анализа соревнования можно сделать вывод, что наибольший интерес для практического применения представляет синтаксический парсер от исследовательской группы *DeepPavlov* [16]. Данная модель не чувствительна к пунктуации, производит выходные данные в формате *CONLL*-U и обучалась на корпусах *Universal Dependency* [17].

Инструментами, позволяющими извлекать структурированную информацию из текстов на естественном языке, являются анализаторы контекстно-свободных грамматик, среди которых для русского языка можно выделить [18, 19].

Перечисленные инструментальные средства зарекомендовали себя при решении задач обработки и анализа текстов на естественных языках, но имеют ряд ограничений. Например, при построении синтаксических деревьев не всегда способны находить все смысловые связи между токенами предложения, а разработка и постоянная актуализация правил на контекстно-свободных грамматиках требует больших человеческих ресурсов.

В данной работе предлагается метод извлечения *знаний* и *навыков*/*компетенций* из текстов требований онлайн-вакансий путём объединения потенциала перечисленных подходов и технологий.

1 Алгоритм извлечения знаний и навыков/компетенций из текстов требований вакансий

Метод состоит в переходе от сложных предложений к набору простых комбинаций слов путём добавления дополнительных связей между словами. Рассматриваются два подхода: на основе линейной комбинации слов; путём расширения дерева синтаксического разбора сложного предложения новыми связями. Схема алгоритма предлагаемого метода приведена на рисунке 1.

Этап 1. Блоки 1.1-1.3. Подготовка корпуса текстов примеров *знаний* и *навыков/компетенций* и правил для извлечения этих категорий сущностей на основе контекстно свободных грамматик.

Этап 2. Блок 2.1. Сбор текстов требований к вакансиям с сайта *headhunter.ru* из категории «Информационные технологии, Интернет, телеком», на основе *html* разметки.

Этап 3. Блоки 3.1-3.2. Предполагается, что распространение информации в тексте, как правило, происходит слева направо, поэтому получение новых комбинаций токенов в предложении можно представить в виде комбинации токенов, где каждый следующий токен должен иметь порядковый номер строго больше предыдущего (см. рисунок 2). Каждый текст требования разбивается на отдельные токены, которые нумеруются в порядке следования, а далее в блоке 3.2 происходит получение линейных комбинаций токенов.

Этап 4. Блоки 4.1-4.5. Дополнение дерева синтаксического разбора рёбрами между токенами, которые согласно дообученной на текстах вакансий модели могут иметь высокую вероятность совместного использования в ПрО.

В блоке 4.1 происходит дообучение языковой модели на текстах вакансий; в этом блоке модель учится понимать особенности текстов ПрО.

В блоке 4.2 происходит построение дерева синтаксического разбора для текстов требований вакансий. Два примера дополненных деревьев синтаксического разбора представлены на рисунке 3 для текста требования «Описание, моделирование (желательно *Bizagi*[1]) и/или оптимизация бизнес-процессов» (см. только прямые рёбра; а) - *spacy_syntax_parser*, б) - *deeppavlov_syntax_parser*).

В блоке 4.3 на основе дообученной модели *BERT* и инструмента заполнения маски происходит добавление новых рёбер в дерево синтаксического разбора, полученное в блоке 4.2. Заполнение маски представляет собой инструмент автозаполнения пропущенных токенов в предложении. В этой задаче токен в предложении заменяется специальной маской «[MASK]», и модель должна предсказать наиболее вероятный токен, которым можно заменить маску «[MASK]», основываясь на контексте предложения, заданном другими словами в предложении. Инструмент заполнения маски используется в качестве основного в процессе дообучения модели, а также при эмпирической оценке качества полученных моделей.

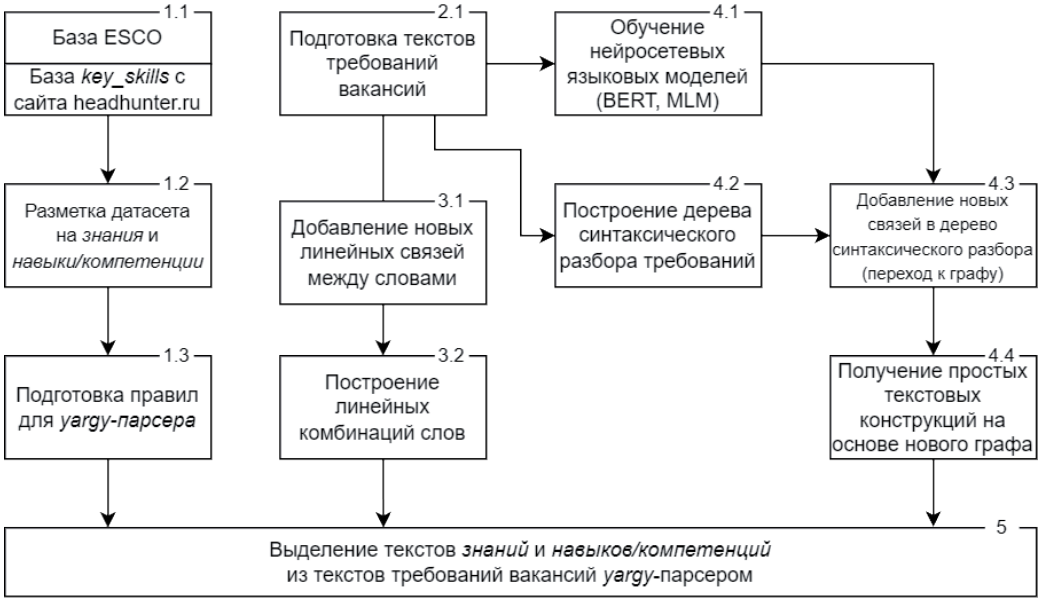


Рисунок 1 - Схема алгоритма извлечения *знаний* и *навыков/компетенций* из текстов требований вакансий

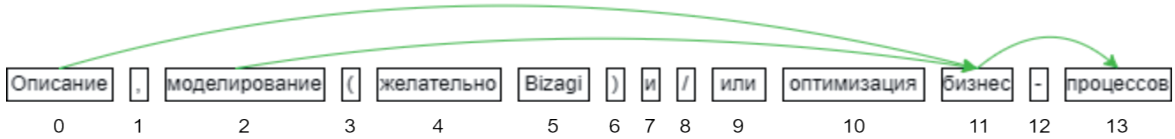
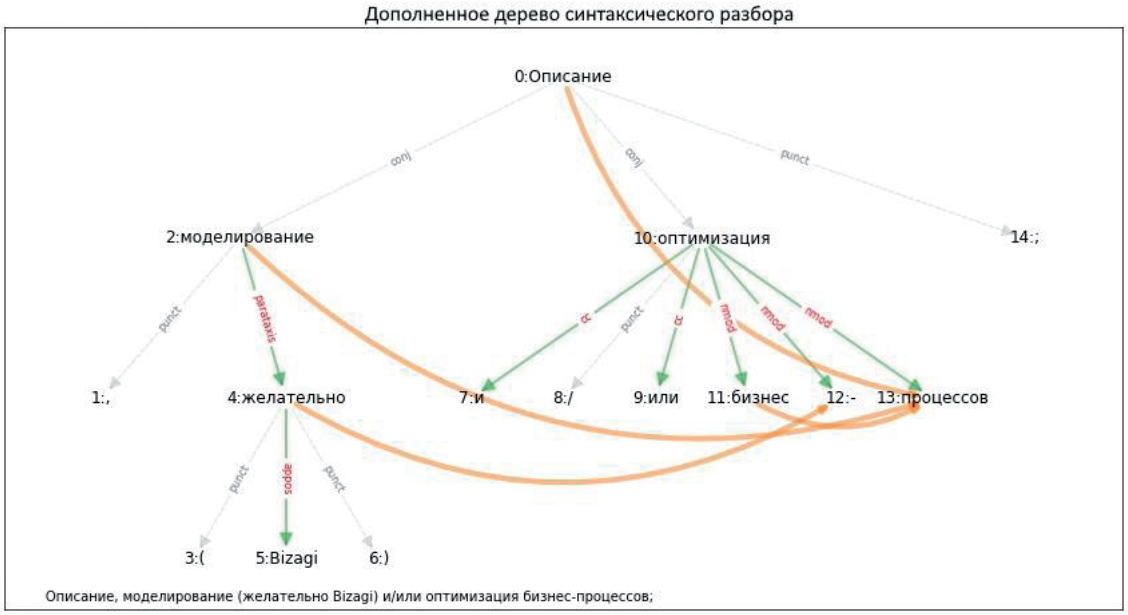


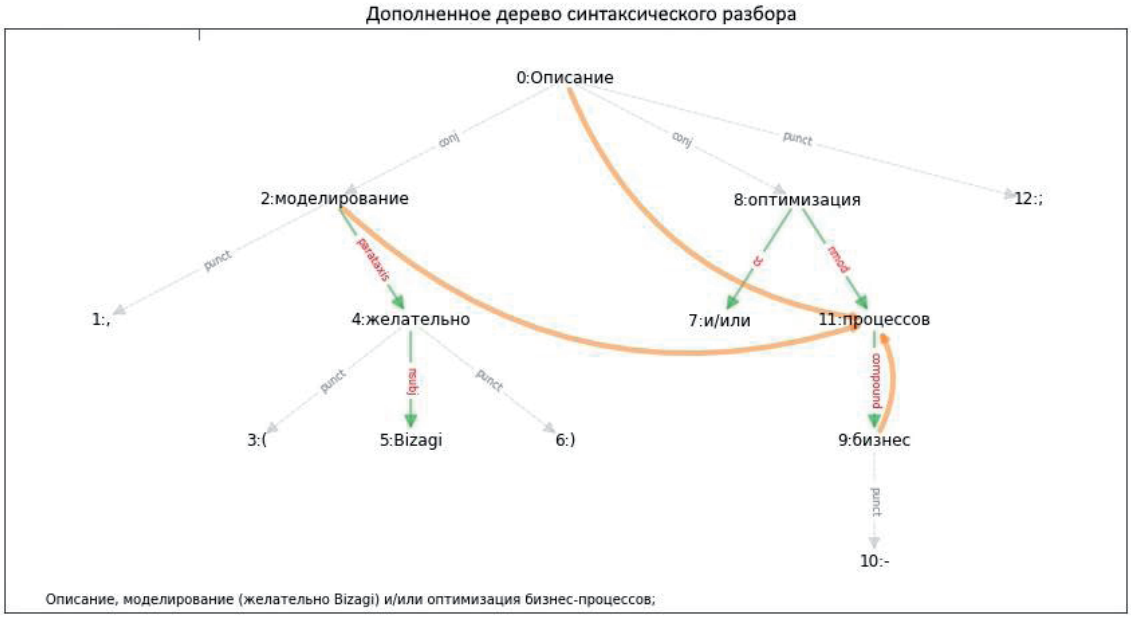
Рисунок 2 – Схема получения линейных комбинаций токенов

Для добавления в дерево новых рёбер попарно рассматриваются узлы дерева, не соединённые ребром. Формируются пары таких узлов. Далее поочерёдно маской закрываются узел 1 и узел 2 в паре, и если оба варианта считаются дообученной моделью с ненулевой вероятностью, то можно считать связь между двумя этими узлами *устойчивой* и добавить новое ребро между парой соответствующих узлов. Иллюстрация работы блока представлена на рисунке 3 (см. изогнутые рёбра). На рисунках 3а и 3б видно, как дообученная модель добавляет новые рёбра в деревья синтаксического разбора.

[1] *Bizagi* – одна из наиболее популярных автоматизированных систем управления бизнес-процессами. https://www.bizagi.com.



а)



б)

Рисунок 3 – Примеры дополненных деревьев синтаксического разбора новыми связями (рёбрами) на основе дообученной модели *BERT* на текстах вакансий: а) *spacy_syntax_parser*, б) *deeppavlov_syntax_parser*

Блок 4.4. Случайным блужданием отбираются пути на графе длиной от 1 до 5 токенов, и на их основе формируются комбинации токенов.

Этап 5. Блок 5. Комбинации токенов от блоков 3 и 4.1 проходят через подготовленный парсер на основе контекстно свободных грамматик, который на основе правил извлекает тексты *знаний* и *навыков/компетенций*. В дальнейшем вместо парсера на основе контекстно свободных грамматик может использоваться модель *BERT,* дообученная на задаче извлечения именованных сущностей *знаний* и *навыков/компетенций*.

2 Подготовка правил для *yargy*-парсера

2.1 Подготовка корпуса текстов *знаний* и *навыков/компетенций*

В процессе подготовки размеченного корпуса текстов *знаний* и *навыков/компетенций* были и отобраны два основных ресурса: база *ESCO* (разметка на *знания* и *навыки/компетенции* уже присутствует по умолчанию, см. рисунок 1); база ключевых знаний и навыков из параметра *key_skills* вакансий *headhunter.ru* (таблица 1). Во втором корпусе текстов отсутствовала разметка текстов на *знания* и *навыки/компетенции*, что потребовало дополнительной разметки этих данных экспертом (в таблице помечены *). Примеры текстов *знаний* и *навыков/компетенций* из двух корпусов текстов представлены в таблице 2.

Таблица 1 – Структура корпуса текстов *знаний* и *навыков/компетенций*

Ресурс	Количество уникальных текстов	Количество текстов знаний	Количество текстов навыков/компетенций
ESCO	9936	2411	7525
key_skills hh.ru	9635	4401 *	5234 *
Итого	**19531**	**6812**	**12759**

Таблица 2 – Примеры текстов знаний и навыков/компетенций

Тип	Текст	Ресурс
знание	*ASP.NET 2.0*	*ESCO*
знание	*Adobe Illustrator CC*	*ESCO*
знание	яндекс.Директ	*hh.ru key_skills*
знание	*numPy*	*hh.ru key_skills*
знание	1С-Битрикс	*hh.ru key_skills*
навык/компетенция	анализировать бизнес-требования	*ESCO*
навык/компетенция	диагностика потребностей клиента	*ESCO*
навык/компетенция	документировать разработку	*hh.ru key_skills*
навык/компетенция	разработка распределенных систем	*hh.ru key_skills*
навык/компетенция	написание кода	*hh.ru key_skills*

2.2 Разработка правил для *yargy*-парсера

Были подготовлены правила извлечения текстов *знаний* и *навыков/компетенций*. Правило для определения *знаний* может быть самостоятельных правилом и опциональным элементом, встраиваемым в правило определения *навыков/компетенций*. Структура правил для извлечения *навыков/компетенций* представлена на рисунке 4.

3 Дообучение модели *BERT* на текстах онлайн-вакансий

Для дообучения модели BERT собраны тексты вакансий с сайта *headhunter.ru* из отрасли «Информационные технологии» за последние 10 лет (1,8 миллиона). Тексты вакансий очищены от дублей, от английских текстов и от *html*-разметки. Для процесса дообучения языковых моделей на текстах онлайн-вакансий случайным образом отобраны 300 тысяч текстов. Были отобраны две языковые модели от проектной группы *Sberbank-AI: sberbank-ai/ruBert-base* (178 миллионов параметров), *sberbank-ai/ruBert-large* (427 миллионов параметров).

В качестве основного метода обучения выбран метод моделирование маскированного языка (*MLM, Masked-Language Modeling*). В рамках этого метода маскировалось 15% токенов исходного текста, которые модель училась предсказывать.

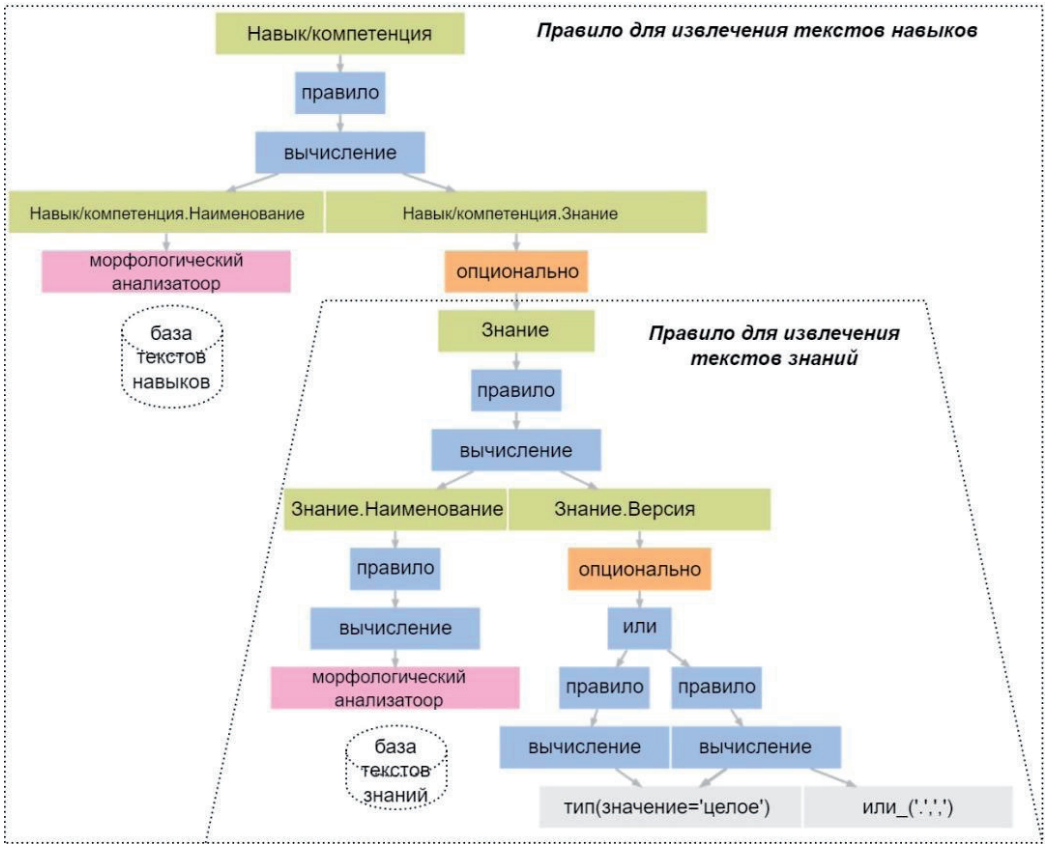


Рисунок 4 – Структура правила *yargy*-парсера для извлечения текстов *навыков/компетенций* (опционально включает в себя правило для извлечения *знаний*)[2]

В качестве основной метрики оценки качества языковых моделей использовалась перплексия [20]. Эта метрика - одна из наиболее распространённых метрик для оценки языковых моделей, определяется как экспоненциальное среднее отрицательное логарифмическое правдоподобие последовательности. Если есть токенизированная последовательность $(x_0, x_1, .., x_t)$, то перплексия для X определяется по формуле

$$PPL(X) = exp\left\{-\frac{1}{t}\sum_i^t \log p_\theta(x_i|x_{<i})\right\}, \quad (1)$$

где $log\ p_\theta(x_i|x_{<i})$ - логарифмическая вероятность i-го токена, обусловленная предыдущими токенами $x_{<i}$. Интуитивно это можно рассматривать как оценку способности модели равномерно прогнозировать среди набора заданных токенов в корпусе.

Дообучение моделей производилось по методу *MLM* в течение пяти эпох (1 эпоха – 30000 шагов, см. рисунок 5). Далее процесс дообучения моделей был расширен до 7 эпох (шаги 150000-210000 на рисунке 5), но это не привело к существенному улучшению показателя перплексии.

Для оценки качества обучения модели использовался инструмент заполнения маски. В таблице 3 приведены результаты заполнения маски базовой модели и модели, которая прошла процесс дообучения на текстах вакансий. Из таблицы 3 видно, что модель, не прошедшая процесс дообучения, справляется с задачей заполнения маски с использованием специфической для данной ПрО лексики хуже, чем дообученная модель. Дообученная модель даёт более релевантные ответы с учётом контекста текстов вакансий. В дальнейших экспериментах использовалась только базовая дообученная модель *sberbank-ai-base*.

[2] Рисунок подготовлен автоматическими средствами библиотеки *natasha.yargy* [19].

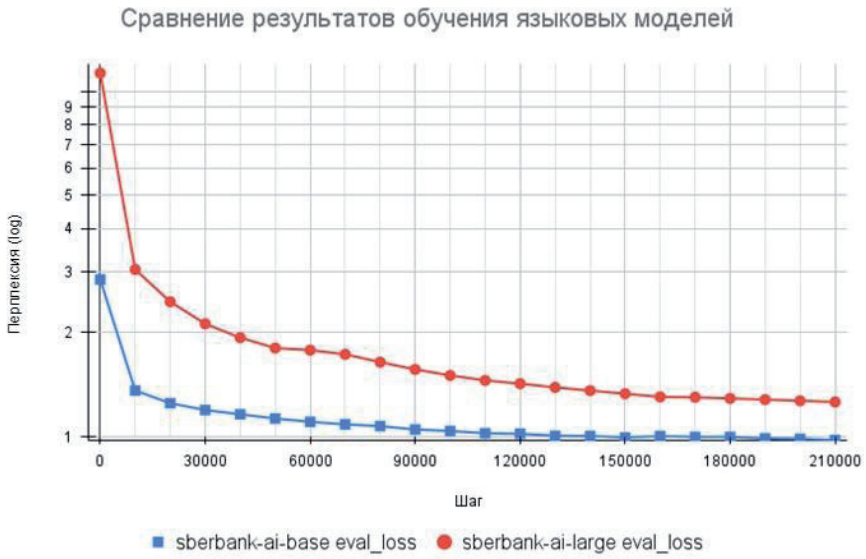


Рисунок 5 – Изменение перплексии в процессе обучения моделей на текстах вакансий

Таблица 3 – Результаты оценки моделей с помощью инструмента заполнения маски на примерах

Пример 1: «сетевой [MASK]»		Пример 2: «выявлять [MASK]»	
Базовая модель без дообучения	Модель после дообучения	Базовая модель без дообучения	Модель после дообучения
сетевой.	сетевой трафик	выявлять.	выявлять неисправность
сетевой!	сетевой мониторинг	выявлять!	выявлять ошибки
сетевой "	сетевой Интернет	выявлять?	выявлять неисправности
сетевой)	сетевой контроль	выявлять ...	выявлять неполадки
сетевой »	сетевой безопасности	выявляться	выявлять закономерности
сетевой интернет	сетевой пакет	выявлять :	выявлять уязвимости
сетевой?	сетевой интернет	выявлять и	выявлять информацию
сетевой суд	сетевой опыт	выявлять »	выявлять их
сетевой фронт	сетевой рост	выявлять "	выявлять проблемы
сетевой анализ	сетевой инженер	выявлять,	выявлять ответственных

4 Описание эксперимента

Этапы проведения эксперимента соответствуют этапам алгоритма из раздела 1.

Этап 1. Для проведения эксперимента из текстов вакансий были отобраны 2000 текстов требований длиной от 8 до 15 токенов, т.к. в текстах с большим количеством токенов, как правило, присутствует большое количество скрытых связей между токенами, что затрудняет извлечение коротких сущностей *знаний* и *навыков/компетенций* из таких текстов.

Этап 2. Тексты требований с помощью синтаксических анализаторов представляются в виде набора токенов и связей между ними (как показано на рисунках 2 и 3). В этом же блоке для каждого текста требований строятся деревья синтаксического разбора.

- *spacy_syntax_parser* - токенизатор и синтаксический анализатор *spacy* на модели для русского языка *ru_core_news_lg.*
- *deeppavlov_syntax_parser* - токенизатор и синтаксический анализатор *deeppavlov* на модели для русского языка *syntax_ru_syntagrus_bert.*

Этап 3. Добавление новых связей между токенами.

В блоке 3.1 на рисунке 1 добавляются все возможные связи между парами токенов, где номер первого токена в связи строго меньше номера второго токена.

В блоке 4.3 на рисунке 1 для каждой пары токенов, несоединённых в дереве синтаксического разбора, с помощь языковой модели и инструмента заполнения маски проверяется необходимость добавления связи между этими токенами. Для соединённых пар токенов проверяется устойчивость связи и её направление.

Проверка происходит следующим образом. Выбирается пара токенов, сначала маской закрывается первый из них и проверяется вероятность использования данного токена в данной паре с помощью языковой модели. На выходе получается вероятность встречаемости данного токена в этой паре на этом месте; если она не нулевая, то добавляется связь от первого токена ко второму. Далее то же повторяется со вторым токеном в паре.

Этап 4. С учётом всех добавленных связей и их направлений между токенами получаются комбинации от 1 до 5 токенов - простые предложения.

Этап 5. С помощью правил, описанных для *yargy-парсера*, в получившихся комбинациях токенов отыскиваются полные совпадения сущностей *знаний* и *навыков/компетенций*. *Yargy-парсер* способен, имея в базе всего один вариант некоторой сущности, находить все её возможные словоформы, в отличие, например, от регулярных выражений.

5 Оценка метода и результаты

Для оценки метода для каждого текста требований экспертом были размечены сущности *знаний* и *навыков/компетенций*, которые содержатся в этом требовании. Для каждого текста требований экспертом размечался полный набор содержащихся в этом тексте сущностей *знаний* и *навыков/компетенций*. Примеры такой разметки приводится в таблице 4.

Таблица 4 – Пример экспертной разметки текста

Текст требования	**Сущности**	**Категория**
Описание, моделирование (желательно *Bizagi*) и/или оптимизация бизнес-процессов;	Описание бизнес-процессов	навык/компетенция
	Описание процессов	навык/компетенция
	Моделирование бизнес-процессов	навык/компетенция
	Моделирование процессов	навык/компетенция
	Оптимизация бизнес-процессов	навык/компетенция
	Оптимизация процессов	навык/компетенция
	Bizagi	знания
Чувство вкуса и стиля	Чувство вкуса	навык/компетенция
	Чувство стиля	навык/компетенция
Разработка *web*-ориентированных, распределенных приложений на языке *Java* с применением технологии *Adobe Flex* на стороне клиента.	Разработка *web*-ориентированных приложений	навык/компетенция
	Разработка *web*-приложений	навык/компетенция
	Разработка распределенных приложений	навык/компетенция
	Adobe Flex	знания
	Разработка приложений на *Java*	навык/компетенция
	Java	знания

При проведении эксперимента были подготовлены восемь комбинаций моделей, инструментов и методов (см. таблицу 5 с результатами). На заключительном этапе по метрике f1 сравнивалось количество извлечённых сущностей *знаний* и *навыков/компетенций* по предложенному методу с экспертной разметкой. Из таблицы 5 видно, что все проанализированные инструменты превзошли базовый вариант извлечения с помощью правил *yargy-парсера*, а наилучший результат получен для комбинации: дообученная *BERT* модель + *deeppavlov_syntax_parser* + линейный способ дополнения связей, которая показала 83% извлечения по метрике f1 и превзошла базовый вариант на основе правил на 52%.

Таблица 5 – Результаты сравнения различных комбинаций инструментов в извлечении знаний и навыков /компетенций

Токенизатор и синтаксический парсер	Схема получения комбинаций токенов	Модель	f1
		yargy-napcep (baseline)	0.31
spacy_syntax_parser	линейная комбинация		0.65
spacy_syntax_parser	дополнение дерева	*sberbank-ai-base*	0.33
deeppavlov_syntax_parser	линейная комбинация		0.58
deeppavlov_syntax_parser	дополнение дерева	*sberbank-ai-base*	0.35
spacy_syntax_parser	линейная комбинация		**0.81**
spacy_syntax_parser	дополнение дерева	*sberbank-ai-base-finetune*	0.69
deeppavlov_syntax_parser	линейная комбинация		**0.83**
deeppavlov_syntax_parser	дополнение дерева	*sberbank-ai-base-finetune*	0.71

Заключение

Предложен и экспериментально проверен метод извлечения коротких текстов знаний и навыков из текстов требований онлайн-вакансий путём перехода от сложных предложений к набору простых комбинаций токенов через дополнение структуры сложных предложений новыми связями дообученной на текстах онлайн-вакансий нейросетевой моделью *BERT*. Применимость метода показана на текстовом корпусе требований вакансий.

Список источников

[1] ESCO — многоязычная классификация европейских навыков, компетенций и профессий. https://esco.ec.europa.eu/en.

[2] ***Burtsev M., Anh Le.*** A Deep Neural Network Model for the Task of Named Entity Recognition. International Journal of Machine Learning and Computing. 2019.

[3] ***Маслова М.А., Дмитриев А.С., Холкин Д.О.*** Методы распознавание именованных сущностей в русском языке: Инженерный вестник Дона, 2021. № 7(79).

[4] ***Хакимова Е.М.*** Сложные предложения в современном русском языке: ортологический аспект. Вестник ЮУрГУ. Серия: Лингвистика. 2013.

[5] ***Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A., Polosukhin I.*** Attention is all you need. In: Advances in neural information processing systems. 2017. 5998-6008.

[6] ***Devlin J., Chang M.W., Lee K, Toutanova K.*** Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[7] ***Ezen-Can A.*** A Comparison of LSTM and BERT for Small Corpus. 2020. arXiv preprint arXiv:2009.05451.

[8] ***Liu Y., Ott M., Goyal N., Du J., Joshi M., Chen D., Stoyanov V.*** Roberta: A robustly optimized bert pretraining approach. 2019. arXiv preprint arXiv:1907.11692.

[9] ***Brown T., Mann B., Ryder N., Subbiah M., Kaplan J.D., Dhariwal P., Amodei D.*** Language models are few-shot learners. Advances in neural information processing systems, 2020, 33, 1877-1901.

[10] ***Raffel C., Shazeer N., Roberts A., Lee K., Narang S., Matena M., Liu P.J.*** Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research, 2020, 21(1), 5485-5551.

[11] ***Lample G., Conneau A.*** Cross-lingual language model pretraining. 2019. arXiv preprint arXiv:1901.07291.

[12] ***Lee J., Yoon W., Kim S., Kim D., Kim S., So C.H., Kang J.*** BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 2020, 36(4), 1234-1240.

[13] ***Chalkidis I., Fergadiotis M., Malakasiotis P., Aletras N., Androutsopoulos I.*** LEGAL-BERT: The muppets straight out of law school. 2020. arXiv preprint arXiv:2010.02559.

[14] ***Beltagy I., Lo K., Cohan A.*** SciBERT: A pretrained language model for scientific text. 2019. arXiv preprint arXiv:1903.10676.

[15] ***Ляшевская О.Н., Шаврина Т.О., Трофимов И.В., Власова Н.А.*** GramEval 2020 Дорожка по автоматическому морфологическому и синтаксическому анализу русских текстов. Международная конференция Dialogue. 2020, 553-569.

[16] Синтаксический парсер *DeepPavlov*. http://docs.deeppavlov.ai/en/master/features/models/syntaxparser.html

[17] ***Zeman D.*** Universal Dependencies 2.5, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. 2019.
[18] Томита-парсер. Технологии Яндекса. https://yandex.ru/dev/tomita/.
[19] Natasha/yargy: Извлечение фактов на основе правил для русского языка. https://github.com/natasha/yargy.
[20] ***Meister C., Cotterell R.*** Language model evaluation beyond perplexity. 2021. arXiv preprint arXiv:2106.00085.

Сведения об авторе

Николаев Иван Евгеньевич, 1986 г. рождения. Окончил Южно-Уральский государственный университет в 2008 г. Старший преподаватель кафедры информационных технологий и экономической информатики Института информационных технологий Челябинского государственного университета. ORCID: 0000-0002-9686-2435. *ivan_nilolaev@csu.ru.*

Поступила в редакцию 03.05.2023, после рецензирования 12.06.2023. Принята к публикации 19.06.2023.

Scientific article DOI: 10.18287/2223-9537-2023-13-2-282-293

Knowledge and skills extraction from the job requirements texts

© 2023, I.E. Nikolaev

Chelyabinsk State University, Chelyabinsk, Russia

Abstract

The analysis of the requirements for vacancies in the labor market shows that they are multi-level language constructions of several words with complex semantic relationships. The aim of the research is to develop a method for extracting short texts of knowledge and skills from texts of job requirements that have a complex organizational structure. The method consists in supplementing the structure of complex sentences with new relationships by means of a BERT neural network model trained on the texts of online vacancies and moving from a complex text to a set of simple word combinations. The process of additional training (finetunig) of BERT neural network models from the Sberbank AI laboratory on the texts of online vacancies is shown. Two mechanisms for adding new links between requirements words, taking into account knowledge from the subject area, are implemented: linear and through the addition of the parsing tree. In the course of the experiment, a comparative analysis was carried out for several combinations of the listed tools. The combination that showed the best result was 'the BERT + deeppavlov_syntax_parser model + a linear method of adding links'. The applicability of the method was demonstrated on the text corpus of online job requirements. The proposed method has shown higher efficiency than the rule-based approach, which involves the use of formal rules and grammar rules for natural language analysis. Using the method will allow you to quickly identify the key changes in the needs of the labor market at the level of requirements texts of individual knowledge and skills.

Key words: *neural network models, additional training of language models, parser, syntax trees, labor market, skills, vacancies.*

For citation: *Nikolaev IE.* Knowledge and skills extraction from the job requirements texts [In Russian]. *Ontology of designing*. 2023; 13(2): 282-293. DOI:10.18287/2223-9537-2023-13-2-282-293.

Conflict of interest: The author declare no conflict of interest.

List of figures and tables

References

[1] ESCO is a multilingual classification of European skills, competencies and occupations. https://esco.ec.europa.eu/en.
[2] ***Burtsev M, Ahn L.*** Deep neural network model for the problem of named objects recognition. International Journal of Machine Learning and Computing. 2019.
[3] ***Maslova MA, Dmitriev AS, Kholkin DO.*** Named Entity Recognition Methods in Russian language: Don Engineering Bulletin, 2021. No. 7(79).
[4] ***Khakimova EM.*** Compound sentences in modern Russian language: the orthological aspect. Bulletin of SUSU. Series: Linguistics. 2013.
[5] ***Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, Polosukhin I.*** Attention is all you need. In: Advances in neural information processing systems. 2017. 5998-6008.
[6] ***Devlin J, Chang MW, Lee K, Toutanova K.*** Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
[7] ***Ezen-Can A.*** A Comparison of LSTM and BERT for Small Corpus. 2020. arXiv preprint arXiv:2009.05451.
[8] ***Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Stoyanov V.*** Roberta: A robustly optimized bert pretraining approach. 2019. arXiv preprint arXiv:1907.11692.
[9] ***Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Amodei D.*** Language models are few-shot learners. Advances in neural information processing systems, 2020, 33, 1877-1901.
[10] ***Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Liu PJ.*** Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research, 2020, 21(1), 5485-5551.
[11] ***Lample G, Conneau A.*** Cross-lingual language model pretraining. 2019. arXiv preprint arXiv:1901.07291.
[12] ***Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J.*** BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 2020, 36(4), 1234-1240.
[13] ***Chalkidis I, Fergadiotis M, Malakasiotis P, Aletras N, Androutsopoulos I.*** LEGAL-BERT: The muppets straight out of law school. 2020. arXiv preprint arXiv:2010.02559.
[14] ***Beltagy I, Lo K, Cohan A.*** SciBERT: A pretrained language model for scientific text. 2019. arXiv preprint arXiv:1903.10676.
[15] ***Lyashevskaya ON, Shavrina TO, Trofimov IV, Vlasova NA.*** GramEval 2020 shared task: Russian full morphology and universal dependencies parsing. In Proc. of the International Conference Dialogue. 2020, 553-569.
[16] Syntactic parsing DeepPavlov documentation. http://docs.deeppavlov.ai/en/master/features/models/syntaxparser.html
[17] ***Zeman D.*** Universal Dependencies 2.5, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. 2019.
[18] Tomita-parser. Yandex Technologies. https://yandex.ru/dev/tomita/
[19] Natasha/yargy: Rule-based facts extraction for Russian language. https://github.com/natasha/yargy
[20] ***Meister C, Cotterell R.*** Language model evaluation beyond perplexity. 2021. arXiv preprint arXiv:2106.00085.

About the author

Ivan Evgenievich Nikolaev (b. 1986). Graduated from the South Ural State University in 2008, Senior Lecturer at the Department of Information Technology and Economic Informatics of the Institute of Information Technology Chelyabinsk State University). ORCID: 0000-0002-9686-2435. *ivan_nilolaev@csu.ru.*

Received May 3, 2023. Revised June 12, 2023. Accepted June 19, 2023.