

УДК 004.85

АКТИВНОЕ ОБУЧЕНИЕ ДЛЯ ИЗВЛЕЧЕНИЯ ЗНАНИЙ ИЗ ОПИСАНИЙ ОБРАЗОВАТЕЛЬНЫХ КУРСОВ В УСЛОВИЯХ МАЛЫХ ОБЪЁМОВ ДАННЫХ

Ю.Д. Кленин

*Институт информационных технологий, Челябинский государственный университет, Челябинск, Россия
Югорский научно-исследовательский институт информационных технологий, Ханты-Мансийск, Россия
jklen@yandex.ru*

Аннотация

В условиях постоянного роста объёмов учебных данных их «ручная» обработка не представляется возможной, уступая место различным моделям и методам машинного обучения. В то же время именно наличие обучающих выборок достаточного объёма позволяет современным алгоритмам машинного обучения хорошо справляться с базовыми прикладными задачами. Однако многие современные задачи сложны и узкоспециализированы. Это ограничивает количество данных, доступных для качественного обучения, снижая эффективность полностью автоматических систем. В работе рассматривается подход к задаче автоматизированного извлечения фактов из коллекций неразмеченных текстовых документов, в условиях малых объёмов учебных данных. Освещаются вопросы интеграции экспертных правил для конкретных предметных областей с обобщёнными, предметно-независимыми моделями машинного обучения, предварительно обученными на больших объёмах данных. Предложенный подход, опираясь на методику активного обучения, позволяет сократить трудозатраты эксперта, необходимые для эффективной генерации шаблонов извлекаемых фактов, сохраняя при этом высокое качество результатов работы системы. Применение предлагаемого метода поиска фактов по шаблону показано на примере задачи поиска информации о целевой аудитории в неструктурированном описании онлайн курсов.

Ключевые слова: граф знаний, онтология, извлечение знаний, активное обучение, экспертные правила, машинное обучение, малые данные.

Цитирование: Кленин, Ю.Д. Активное обучение для извлечения знаний из описаний образовательных курсов в условиях малых объёмов данных / Ю.Д. Кленин // Онтология проектирования. – 2019. – Т. 9, №4(34). – С.522-535. – DOI: 10.18287/2223-9537-2019-9-4-522-535.

Введение

В постоянно растущем объёме доступной информации, значительную часть составляют различные коллекции слабо обработанных, «сырых» текстовых документов, как доступных публично, так и частных, проприетарных.

Подобная информация находит всё более широкое применение в самых различных задачах: от повышения продаж [1] до раннего выявления и предотвращения террористических угроз [2]. Таким образом, можно говорить о постоянном возрастании потребности в средствах автоматической и автоматизированной обработки текстовых данных.

Подходы к обработке данных можно разделить на два класса: использующие алгоритмы и модели машинного обучения и основывающиеся на системах экспертных правил [3, 4]. Машинное обучение представлено двумя группами алгоритмов: «с учителем» и «без учителя». Для обучения первых требуется наличие значительных объёмов качественно размеченных данных, которые, как правило, готовятся «вручную» для каждой конкретной задачи. Вторые способны обучаться на неразмеченных данных, однако, для качественного обучения нуждаются в значительно больших объёмах обучающей информации.

Обе группы хорошо справляются с общими, низкоуровневыми задачами, для которых существуют либо очень большие коллекции документов, либо качественные, созданные вручную разметки. Для задач более высокого уровня, а также для специальных задач, модели машинного обучения зачастую показывают себя хуже их аналогов на основе правил. Показательны в этом смысле итоги соревнований, прошедших в рамках конференции Диалог-2016. В задаче анализа тональности текстов на русском языке *SentiRuEval* [5] лучшие результаты были достигнуты участниками, применявшими метод опорных векторов и рекуррентные нейронные сети с использованием дополнительных коллекций для обучения моделей. Однако в соревновании по извлечению именованных сущностей и фактов из текстов на русском языке *FactRuEval* [6] системы на основе правил показали более высокие результаты в сравнении с алгоритмами машинного обучения. Организаторы соревнования отметили, что в задаче извлечения фактов малые размеры корпуса документов лишали возможности применить системы на основе машинного обучения. Тем не менее, значительная часть реально существующих корпусов документов и связанных с ними задач являются узкоспециализированными, ограниченными в объемах доступной информации, что может вызывать трудности в работе моделей машинного обучения.

В статье внимание сосредоточено на задаче выделения знаний. Именно её решение позволит применять неразмеченные текстовые коллекции в качестве ресурсов для более сложных систем, таких как вопросно-ответные системы [7], системы автоматического резюмирования [8], системы поддержки принятия решений [9], системы борьбы с плагиатом [10] и др. Для этого предпринята попытка интеграции моделей машинного обучения, опирающихся на анализ больших коллекций данных, с системами на основе правил, стремящаяся совместить лучшие качества обоих подходов в целях качественного решения задачи автоматизированного извлечения фактов из ограниченных по размеру коллекций слабоструктурированных текстовых документов.

1 Моделирование знания

Первоочередным вопросом в задаче извлечения знаний, является выбор модели, используемой для работы с полученными фактами и их отображения. Наиболее популярными в инженерии знаний, на сегодняшний день, являются такие модели, как таксономия, семантическая сеть и онтология.

1.1 Таксономия

Таксономии, как правило, отражены как иерархии концептов, связанных отношением подтип-супертип, в котором каждый последующий уровень таксономии состоит из подтипов предыдущего. Математически таксономию можно описать следующим образом:

$$T = (C, H),$$

где C обозначает множество концептов, H - иерархию отношений между концептами.

1.2 Семантическая сеть

В семантических сетях выделяются такие компоненты как концепты и отношения, что позволяет применить графы для визуализации и использования модели, обозначая концепты вершинами и соединяя их ребрами-отношениями различных типов:

$$N = (C, R),$$

где C обозначает множество концептов, R - множество отношений между концептами.

1.3 Онтология

Онтологии имеют более сложную структуру, выделяя большее число типов элементов и связей между ними. Конкретный набор типов компонентов зависит от реализации модели. Как правило, онтологии включают следующие типы элементов: экземпляры, классы, атрибуты, и отношения. Экземпляры и классы являются двумя разновидностями объектов-вершин графа. Классы являются коллекциями объектов, в то время как экземпляры определяют конкретные объекты нижнего уровня. Внутренняя организация объектов усложнена также наличием атрибутов-свойств, описывающих конкретные аспекты объекта. Математически простейшая онтология определяется как

$$O = (I \cup C, R, A),$$

где I, C, R, A - множества экземпляров, классов, отношений и атрибутов соответственно.

2 Выделение знаний

Первичным этапом формирования графа знаний является определение и извлечение концептов. Наиболее распространённым подвидом этой задачи является извлечение именованных сущностей (*named entity recognition, NER*). Так, по результатам соревнования *wnut16* [11] в задаче извлечения и классификации именованных сущностей из сообщений сети *Twitter*, первое место занял подход, использующий двунаправленную нейронную сеть с долгой краткосрочной памятью (*bidirectional long short-term memory, bi-LSTM*) [12], опирающуюся на символьные и словесные векторные представления.

В русскоязычном соревновании *FactRuEval-2016* [6] один из лучших результатов был достигнут системой, использующей тексты *Wikipedia* и базу знаний *Wikidata* для формирования словарей именованных сущностей [13]. Другая команда [14] добилась схожего результата, используя систему лаконичных правил извлечения фактов различных типов на основе синтактико-семантического дерева предложения. Основной проблемой данного подхода является привязанность к заранее определённым экспертами общим фактам конкретных типов, соответствующих именованным сущностям, что усложняет процесс расширения и дополнения системы.

Подходы, рассмотренные в [15-17] предлагают модели на основе *bi-LSTM*, обученных без использования синтаксической и грамматической информации, опираясь лишь на неразмеченный текст.

Более сложным этапом выделения знаний является выявление и классифицирование отношений между сущностями. В ряде работ [18, 19] используются свёрточные нейронные сети (*convolutional neural networks, CNN*), восполняющие необходимость в тренировочной разметке с помощью автоматической разметки на основе графа знаний.

Известны схожие исследования, которые опираются на нейронные сети для разметки типов отношений, использующие синтаксическую структуру обрабатываемого текста, включая синтаксические деревья предложений [20–22]. Так в [20] использована особая архитектура *graph-LSTM*, на вход которой подаётся информация как о порядке слов в предложении, так и о порядке участников синтаксического дерева фразы. В работах [21, 22] показана эффективность использования слоя внимания, который автоматически обучается определять важность той или иной составляющей предложения.

Примерами комплексного подхода к извлечению знаний можно считать алгоритмы инженерии онтологий. Один такой подход, опирающийся на набор экспертных правил, описан в [23], где рассмотрен процесс генерации онтологии на основе коллекции «*is-a*» фактов: предложений, содержащих формальные определения тех или иных понятий на английском

языке. Подход на основе глубокого обучения и использования рекуррентных сетей предложен в [24]. Здесь нейросетевая модель обучается автоматически переводить простые предложения с определениями из неразмеченного вида в размеченный в формате *OWL*.

Альтернативный вариант составления онтологии - конкретно, на основе тематического состава документов, - изложен в работе [25], где предлагается автоматически наполнять онтологию терминами исходя из результатов тематического моделирования коллекции.

Ограничением упомянутых методов является их ориентация на конкретную структуру предложений.

В статье [26] рассмотрена возможность использования машинного перевода и нейронных сетей в задаче преобразования неструктурированного текста в структурированную разметку. Такой подход требует большого объёма данных для обучения, что ограничивает возможности его использования.

3 Активное обучение в условиях нехватки данных

Одним из способов преодоления ограниченности коллекций является активное обучение, которое позволяет сократить затраты труда экспертов: моделям, использующим данный подход, предоставляется возможность интерактивно взаимодействовать с экспертом в процессе обучения, с целью выделения данных, «ручная» разметка которых принесла бы наибольшую пользу обучаемой модели.

Наиболее сложным в активном обучении является выбор принципа отбора данных для обработки экспертом. Основными алгоритмами отбора являются случайный отбор и отбор тех примеров, в которых система наименее уверена. Уверенность для каждого примера определяется вероятностью ответа. В [27] показано применение различных алгоритмов отбора на примере задачи *NER* в медицине. Авторы [28] математически показывают, что традиционная неуверенность имеет «предвзятость» к выбору более длинных предложений, предлагают алгоритмы *MNLP* (*Maximum Normalized Log-Probability*) и *BALD* (*Bayesian Active Learning by Disagreement*), снижающие требуемый объём данных до одной четвертой.

Другой подход на основе идеи разнообразия предложен в [29], где алгоритм группирует примеры по близости их структур (учитывая части речи и синтаксические роли элементов текста). Схожий метод применяется и в [30].

В работе [31] вместо статистической оценки каждого примера предложено использовать марковский процесс принятия решений в выборе новых примеров для экспертной обработки. Важно отметить, что определённые сложности могут возникать в обучении данных моделей в задачах, где невозможно составление «золотого стандарта» тестовой выборки, и результаты работы системы должны оцениваться экспертом.

Принципиально иной подход к активному обучению рассмотрен в [32]. Вместо того чтобы использовать каждый пример единожды, система может запросить повторную разметку экспертом, что может уменьшить влияние ошибок, допущенных экспертами при первичной разметке. Отмечено, что после определённого объёма размеченных данных прирост качества становится незначительным, а на малых объёмах данных, лидируют традиционные алгоритмы выбора примеров.

4 Предлагаемый подход

Поскольку разрабатываемый подход предназначен для структурирования текстовой информации в условиях ограниченности набора документов, то применяемые модели и алгоритмы должны опираться на общезыковые принципы, независимые от специфики Про.

4.1 Предлагаемая модель

Приведённые в разделе 1 модели являются абстрактными форматами структурирования знания. Основным недостатком, связанным с их применением, является оторванность от источников содержащихся в них знаний. В рамках реальных прикладных задач любое знание, отражаемое в модели, происходит из конкретного документа. Эта информация выражается в документе с помощью простого, неструктурированного текста. Её можно представить в качестве множества взаимосвязанных элементарных фрагментов знания, а документ - в виде набора упоминаний этих фрагментов, отражающих содержащееся в них знание в виде текста.

Для учёта всей информации и минимизации потерь, связанных со структурированием знания, содержащегося в документе, предлагается использовать комбинированную модель, учитывающую как абстрактные знания, так и их упоминания в документе. Математически данную модель можно описать как

$$MKG = (M_C, M_R, C, R, F),$$

где M_C описывает множество упоминаний концептов, M_R - множество упоминаний отношений, C - множество концептов предметной области (ПрО), R - множество отношений между концептами, F - набор фактов, существующих внутри модели.

В предлагаемой модели каждый концепт ПрО содержит информацию о конкретных упоминаниях этого концепта, использованных в обработанной коллекции. То же верно и для отношений. Кроме того, предлагается использовать такой дополнительный структурный элемент, как «факт» - традиционный триплет, состоящий из двух концептов и отношения между ними.

4.2 Синтаксический анализатор

Для извлечения базовой структуры обрабатываемой фразы применена модель *UDPipe* [33], производящая синтаксический и грамматический анализ обрабатываемых документов, выделяя в них так называемые универсальные зависимости (*universal dependencies, UD*). Универсальные зависимости - это открытый проект, предназначенный для создания устойчивой кросс-языковой структуры синтаксических и семантических отношений [34]. Проект содержит более 100 размеченных текстовых коллекций на более чем 70 языках, все с общей синтактико-семантической разметкой.

UDPipe - это тренируемая модель для токенизации, разметки, лемматизации, и разбора синтаксических зависимостей. В основе *UDPipe* лежат *bi-LSTM*-сети, обучаемые производить полный разбор текстов на универсальные зависимости. Такой подход позволяет говорить о практической языковой независимости *UDPipe*: её применение возможно для любого из языков, поддерживаемых проектом. *UDPipe* является одним из победителей соревнования *CoNLL-2018* [35], задачей которого было определить наиболее качественные анализаторы *UD*-разметки.



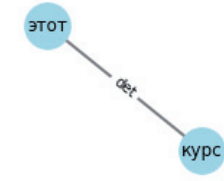

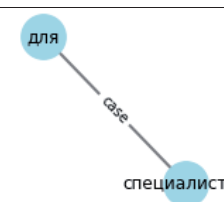





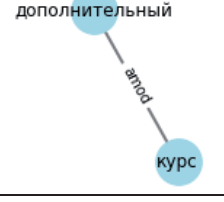



4.3 Переход от синтаксической к семантической схеме предложения

Результатом работы *UDPipe* является дерево, отражающее синтаксическую структуру фразы и содержащее базовую семантическую информацию о составляющих. Однако для структурирования текстовой информации требуются дополнительные действия по преобразованию этого дерева к полному семантическому графу.

С этой целью предлагается использовать набор правил преобразования дерева, независимых от специфики коллекции. Важно отметить, что в ходе применения этих правил, не теряется никакой информации об оригинальной структуре фразы: все оригинальные вершины

и их отношения сохраняются, а все изменения определяются только дополнительно вводимыми вершинами. Эти новые вершины представляют собой объединения уже существующих вершин в соответствии с их принадлежностью той или иной логической составляющей факта, описанного во фразе. Предлагаемые преобразования описаны в таблице 1.

Таблица 1 – Основные преобразования дерева универсальных зависимостей для выделения фактов

Правило	Пример	Результат применения
Устоявшиеся словосочетания и целостные комбинации слов, выраженные универсальными отношениями <i>«flat»</i> , <i>«fixed»</i> , и <i>«compound»</i> , объединяются в одну вершину, с которой они связываются элементарным семантическим отношением часть-целое (<i>part-of</i>).		
Определители, выраженные отношением <i>«det»</i> , объединяются с вершиной, которой они подчинены, в общую вершину, связывающуюся с определителем отношением вспомогательного элемента (<i>helper</i>), а с оригинальной вершиной отношением подтипа (<i>is-a</i>).		
Предлоги, выраженные отношением <i>«case»</i> , объединяются с вершиной, которой они подчинены, в общую вершину, связывающуюся с предлогом отношением вспомогательного элемента (<i>helper</i>), а с оригинальной вершиной отношением подтипа (<i>is-a</i>).		
Вспомогательные глаголы-связки (копулы), выраженные отношением <i>«cop»</i> , объединяются с вершиной, которой они подчинены, в общую вершину, связывающуюся с предлогом отношением вспомогательного элемента (<i>helper</i>), а с оригинальной вершиной отношением подтипа (<i>is-a</i>).		
Если существуют несколько конъюнктивных элементов, что выражается отношением <i>«conj»</i> , то все отношения, ведущие к ним от вышестоящих элементов иерархии, дублируются на все элементы, что позволяет восстановить связи с каждым из равноправных вариантов.		
Различные модификаторы, такие как прилагательные (<i>«amod»</i>), наречия (<i>«advmod»</i>), существительные (<i>«nmod»</i>), числительные (<i>«nummod»</i>), объединяются с вершиной, которой они подчинены в новую вершину. Эта объединённая вершина связана с модифицируемым словом отношением подтипа (<i>is-a</i>).		
Отношения, определяющие связи между глаголами и существительными предложения, наделены семантическими смыслами. Выделены: субъект (универсальное отношение <i>«nsubj»</i>), объект (<i>«obj»</i> для прямого объекта, <i>«iobj»</i> для косвенного), и аргумент (<i>«obl»</i> для обстоятельств действия).		

После применения этих преобразований к дереву универсальных зависимостей становится возможным выделить основные компоненты возможных фактов: триплеты «субъект-отношение-объект» и «субъект-отношение-аргумент», выраженные, пользуясь терминологией русской грамматики, подлежащим, сказуемым, и дополнением.

Пример построения семантического графа для фразы показан на рисунке 1. Здесь упоминания, состоящие из нескольких вершин, объединены в одну, при этом старые вершины сохранены с уточнением семантических связей с объединённой вершиной. В частности, «студенты магистратуры» - это подвид «студентов», а «магистратура» является в данном случае его модификатором.

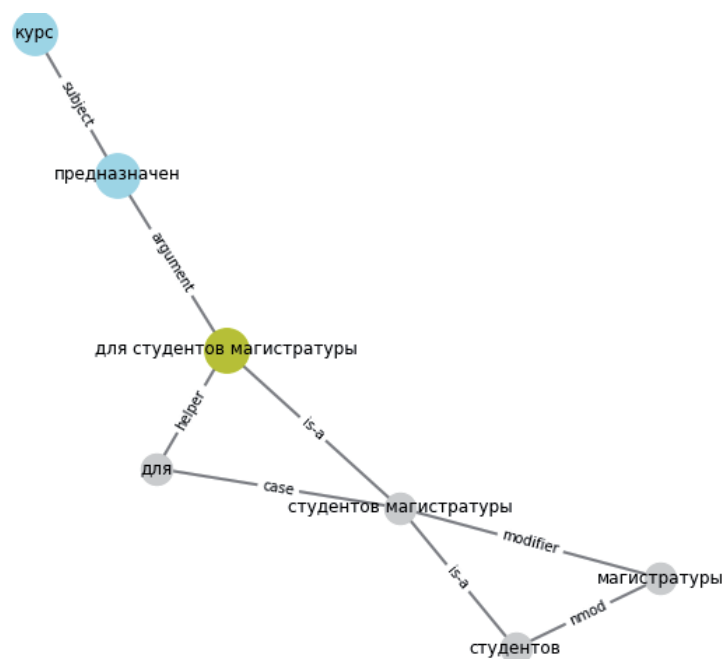


Рисунок 1 – Пример графа, построенного для фразы «курс предназначен для студентов магистратуры»

4.4 Выделение шаблона факта

После преобразования графа универсальных отношений к графу семантических фактов для каждой конкретной фразы документа в наборе этих фактов необходимо выделить те, которые интересуют эксперта в конкретной прикладной задаче. Эти факты должны быть внесены в предложенную в подразделе 4.1 модель графа знаний с упоминаниями *MKG*. Согласно данной модели триплеты «субъект-отношение-объект» и «субъект-отношение-аргумент» могут быть отображены с помощью триплета элементов «концепт–отношение–концепт», или apb , где $a, b \in C$, $a \rho \in R$. При этом все три элемента в триплетах имеют конкретное текстовое представление в виде упоминаний, составляющих ту или иную фразу, а также чётко определённые типы универсальных зависимостей и семантических связей.

Например, приведённый на рисунке 1 факт состоит из:

- концепта «курс», выраженного упоминанием «курс»;
- концепта «студенты магистратуры», выраженного упоминанием «для студентов магистратуры»;
- отношения «предназначен», выраженного упоминанием «предназначен»;
- связи между упоминаниями «курс» и «предназначен», имеющей синтаксический тип *nsubj* с семантическим отношением субъект (*subject*);
- связи между упоминаниями «для студентов магистратуры» и «предназначен», имеющей синтаксический тип *obl* с семантическим отношением аргумент (*argument*);

Вместе эти составляющие образуют факт «аудитория курса», описывающий людей, которым может быть интересно прохождение некоторого курса. Для внесения данного факта в модель эксперту достаточно задать эти составные части.

Очевидно, что множество различных фраз могут описывать аудитории тех или иных курсов, следуя, при этом, одному и тому же шаблону, представленному на рисунке 2а. Одни и те же концепты и отношения могут быть выражены разными упоминаниями, например, оба

графа на рисунке 2 следуют одному и тому же шаблону, используя различные формулировки. Таким образом, один и тот же факт должен включать в себя все известные упоминания, отражающие входящие в него концепты и отношение между ними, а также все известные варианты связей. Так, на рисунке 2а показано упоминание «аудитория» в качестве аргумента упоминания «предназначен»; на рисунке 2б представлен вариант, когда упоминание «аудитория» является объектом для упоминания «будет интересен», ввиду различной синтаксической структуры двух фраз.



Рисунок 2 - Варианты шаблонов факта «Аудитория курса»

4.5 Применение семантической близости для поиска близких упоминаний

Для сопоставления различных фрагментов текста необходимо применение методов определения их семантической близости - вычислительных методов расчёта соответствующей количественной оценки. Одной из моделей определения смысловой близости является *word2vec* [36].

Word2vec основывается на неглубокой нейронной сети, обучающейся сопоставлять слова и их контексты таким образом, чтобы быть способной предсказать слово по контексту или наиболее вероятный контекст по слову. Побочным эффектом является формирование векторных представлений изученных слов, которые тем ближе друг к другу, чем более схожи контексты этих слов. Данный подход основывается на дистрибутивной гипотезе: языковые единицы, встречающиеся в похожих контекстах, похожи, что позволяет говорить о семантической близости слов, чьи векторные отображения взаимно близки, и о семантической независимости слов, чьи векторы далеки друг от друга.

Для измерения близости векторов традиционно применяется косинусная мера близости, измеряющая близость как косинус угла между многомерными направлениями векторов: совпадающие по направлению векторы имеют близость равную 1, а перпендикулярные, полностью независимые друг от друга 0.

4.6 Извлечение факта по шаблону

Удобство шаблонов фактов состоит в том, что их можно с лёгкостью применять для поиска схожих фактов в неструктурированном тексте. Для этого необходимо:

- проверить наличие в предложении упоминания первого концепта факта, используя известные его упоминания и оценивая их семантическую близость с упоминаниями в обрабатываемом предложении;
- проверить наличие в предложении упоминания отношения факта, используя известные его упоминания и оценивая их семантическую близость с упоминаниями в обрабатываемом предложении;
- проверить наличие между этими упоминаниями связи, соответствующей по типу известным шаблонам связей между упоминаниями первого концепта и упоминаниями отношения;

- проверить наличие исходящей из упоминания отношения связи, соответствующей по типу известным шаблонам связей между упоминаниями отношения и упоминаниями второго концепта;
- извлечь второй концепт этой связи как упоминание второй сущности.

После выполнения этих операций фразу можно считать обработанной. Если удалось выполнить все операции успешно, то исходная фраза содержит описание факта, соответствующего шаблону, по которому ведётся поиск.

4.7 Применение активного обучения для снижения затрат труда экспертов

Недостатком предлагаемого подхода является его сильная зависимость от наличия экспертной разметки извлекаемых фактов. Для снижения объёма требуемой разметки предлагается применять активное обучение.

Активное обучение основывается на итеративном процессе обработки данных. При этом на каждой итерации происходит обработка некоторой части корпуса с последующей консультацией у эксперта. Ввиду малого объёма данных в коллекции, имеет смысл применять поиск сразу ко всему набору фраз, что позволяет отслеживать эффективность алгоритма в зависимости от числа итераций и объёма размеченных «вручную» данных. Каждая итерация, таким образом, состоит из двух фаз: разметка и поиск фактов по шаблону.

В первой фазе эксперт осуществляет разметку очередного примера. Выбор примера для разметки производится случайным образом среди фраз, которые содержат описание аудитории курса, но прежде не были размечены и не поддавались автоматической обработке во время второй фазы на предыдущей итерации.

Во время второй фазы алгоритм производит поиск всех фактов, соответствующих обновлённому шаблону.

Результатом выполнения некоторого числа итераций является набор фраз, распознанных как описания аудитории курса, и соответствующих им графов фактов.

5 Проведение эксперимента

Рассматривается задача извлечения из текстового описания онлайн-курсов информации об их целевой аудитории. Исследуется возможность минимизации экспертных затрат с помощью применения активного обучения для автоматизированного формирования факта «аудитория курса».

В качестве коллекции документов был взят набор из 143 курсов платформы *Coursera*, на русском языке, содержащих информацию об их целевой аудитории. Суммарно, эта информация составляет 357 предложений. Малые размеры этого корпуса не позволяют обучать на нём сложные модели, и он хорошо подходит для проведения данного эксперимента.

Использованный корпус текстов не подвергался никакой предварительной обработке, перед тем как быть поданным на вход системе. Был произведён экспертный анализ всех входящих в него предложений, на основании которого было определено, что общее число фраз, содержащих информацию об аудитории курса, составило 137.

В рамках данной работы используется *UDPipe* модель для русского языка, предобученная на банке деревьев *SynTagRus*, разрабатываемом лабораторией компьютерной лингвистики Института проблем передачи информации РАН. Для отражения в векторное пространство применяется предварительно обученная на Национальном корпусе русского языка модель *word2vec CBOW*, публично предоставленная проектом *RusVectōrēs*. В качестве порога семантической близости было выбрано значение 0,7.

Первоначальный шаблон факта создаётся на основе единственного предложения, размеченного экспертом, после чего производится поиск фактов по данному шаблону. Разложение на требуемые составляющие первого экспертного примера - «Курс предназначен для старшеклассников, абитуриентов и студентов начальных курсов» - имеет вид:

- упоминание первого концепта: «Курс»;
- упоминание отношения: «предназначен»;
- упоминания второго концепта: «старшеклассников», «абитуриентов», «студентов начальных курсов».

На основании этой разметки система запоминает пять упоминаний, один тип связи между упоминанием отношения и упоминанием первого концепта, и один тип связи между упоминанием отношения и упоминанием второго концепта.

Результаты четырёх раундов поиска по шаблону отображены в таблице 2. По итогам четвёртого раунда система выявила 90 фактов из 137.

Таблица 2 – Результаты по 4 раундам поиска

Раунд	1	2	3	4
Найдено фактов	47	62	86	90

Пользуясь четырьмя размеченными примерами из 137, система смогла найти все эти примеры и ещё 86 из тех, которые ей раньше не встречались.

Для оценки качества работы системы был произведён экспертный анализ ответов системы на каждую фразу коллекции в четвёртом раунде, результаты представлены в таблице 3.

Таблица 3 – Оценка качества работы системы после четырёх раундов поиска

Метрика	Значение
<i>Accuracy</i>	0.866
<i>Precision</i>	1.000
<i>Recall</i>	0.652
<i>F1</i>	0.789

По результатам оценки можно сделать вывод, что в условиях нехватки обучающих данных предложенный подход способен адекватно справиться с задачей уменьшения нагрузки на эксперта, добившись 0.789 баллов *F*-меры (*F-score*). Кроме того, следует отметить, что система имеет «предвзятость» по отношению к точности (*precision*), нежели чем к возврату (*recall*) - она находит только корректные факты ценой их ограниченного количества. И хотя система не даёт ложноположительных результатов, 47 ненайденных ей фактов приводят к снижению оценки правильности (*accuracy*), в виду своей ложной отрицательности.

Заключение

В статье показан подход к структурированию «сырых» текстовых документов, позволяющий эффективно снижать трудозатраты экспертов при сохранении достаточного качества работы. Рассмотренный алгоритм, опираясь на активное обучение, успешно интегрирует известные подходы на основе экспертных правил Про и на больших предметно-независимых объёмах данных, используемых в моделях машинного обучения.

Благодарности

Исследование выполнено при поддержке Российского фонда фундаментальных исследований в рамках проекта №18-47-860013 р_а «Интеллектуальная система формирования обра-

зовательных программ на основе нейросетевых моделей естественного языка с учётом потребностей цифровой экономики Ханты-Мансийского автономного округа - Югры» (договор №18-47-860013\18).

СПИСОК ИСТОЧНИКОВ

- [1] **He, W.** Gaining competitive intelligence from social media data: Evidence from two largest retail chains in the world / W. He, J. Shen, X. Tian, Y. Li, V. Akula, G. Yan, R. Tao // *Industrial Management & Data Systems*. – 2015. – Vol. 115, No. 9. – P.1622-1636.
- [2] **Ahmad, U.** Counter Terrorism on Online Social Networks Using Web Mining Techniques / U. Ahmad // *Intelligent Technologies and Applications: First International Conference, INTAP 2018, Bahawalpur, Pakistan, October 23-25, 2018, Revised Selected Papers*. – Springer, 2019. – Vol. 932. – P.240.
- [3] **Toulouse, T.** Automatic fire pixel detection using image processing: a comparative analysis of rule-based and machine learning-based methods / T. Toulouse, L. Rossi, T. Celik, M. Akhloufi // *Signal, Image and Video Processing*. – 2016. – Vol. 10, No. 4. – P.647-654.
- [4] **Ginneken, B.** Fifty years of computer analysis in chest imaging: rule-based, machine learning, deep learning / B. Ginneken // *Radiological Physics and Technology*. – 2017. – Vol. 1, No. 10. – P.23-32.
- [5] **Lukashevich, N.V.** SentiRuEval-2016: overcoming time gap and data sparsity in tweet sentiment analysis / N.V. Lukashevich, Y.V. Rubtsova // *Proceedings of International Conference Dialog-2016* – 2016. – P.416-426.
- [6] **Starostin, A.S.** FactRuEval 2016: evaluation of named entity recognition and fact extraction systems for Russian / A.S. Starostin, V.V. Bocharov, S.V. Alexeeva, A.A. Bodrova, A.S. Chuchunkov, S.S. Dzhumaev, I.V. Efimenko, D.V. Granovsk, V.F. Khoroshevsky, I.V. Krylova, M.A. Nikolaeva, I.M. Smurov, S.Y. Toldova. // *Proceedings of International Conference Dialog-2016* – 2016. – P.702-720.
- [7] **Abdi, A.** QAPD: an ontology-based question answering system in the physics domain / A. Abdi, N. Idris, Z. Ahmad // *Soft Computing*. – 2018. – Vol. 22, No. 1. – P.213-230.
- [8] **Mohan, M.J.** A study on ontology based abstractive summarization / M.J. Mohan, C. Sunitha, A. Ganesh, A. Jaya // *Procedia Computer Science*. – 2016. – Vol. 87. – P.32-37.
- [9] **Kontopoulos, E.** An ontology-based decision support tool for optimizing domestic solar hot water system selection / E. Kontopoulos, G. Martinopoulos, D. Lazarou, N. Bassiliades // *Journal of Cleaner Production*. – 2016. – Vol. 112. – P.4636-4646.
- [10] **Vrbanec, T.** The struggle with academic plagiarism: Approaches based on semantic similarity / T. Vrbanec, A. Meštrović // *40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. – IEEE, 2017. – P.870-875.
- [11] **Strauss, B.** Results of the wnut16 named entity recognition shared task / B. Strauss, B. Toma, A. Ritter, M.C. De Marneffe, W. Xu // *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*. – 2016. – P.138-144.
- [12] **Limsopatham, N.** Bidirectional LSTM for named entity recognition in Twitter messages / N. Limsopatham, N.H. Collier // *Proceedings of the 2nd Workshop on Noisy User-generated Text* – 2016. – P.145-152.
- [13] **Sysoev, A.A.** Named entity recognition in Russian: the power of wiki-based approach / A.A. Sysoev, I.A. Andrianov // *Proceedings of International Conference "Dialogue"*. – 2016. – P.746-755.
- [14] **Stepanova, M.E.** Information Extraction Based on Deep Syntactic-Semantic Analysis / M.E. Stepanova, E.A. Budnikov, A.N. Chelombeeva, P.V. Matavina, D.A. Skorinkin. // *Proceedings of International Conference "Dialogue"*. – 2016. – P.721-733.
- [15] **Lample, G.** Neural architectures for named entity recognition / G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer // *arXiv preprint arXiv: 1603.01360*. – 2016.
- [16] **Dernoncourt, F.** NeuroNER: an easy-to-use program for named-entity recognition based on neural networks / F. Dernoncourt, J.Y. Lee, P. Szolovits // *arXiv preprint arXiv: 1705.05487*. – 2017.
- [17] **Corbett, P.** Chemlistem: chemical named entity recognition using recurrent neural networks / P. Corbett, J. Boyle // *Journal of Cheminformatics*. – 2018. – Vol. 10, No. 1. – P.59.
- [18] **Zeng, X.** Large scaled relation extraction with reinforcement learning / X. Zeng, S. He, K. Liu, J. Zhao // *Thirty-Second AAAI Conference on Artificial Intelligence*. – 2018.
- [19] **Bai, F.** Structured Minimally Supervised Learning for Neural Relation Extraction / F. Bai, A. Ritter // *arXiv preprint arXiv: 1904.00118*. – 2019.
- [20] **Peng, N.** Cross-sentence n-ary relation extraction with graph lstms / N. Peng, H. Poon, C. Quirk, K. Toutanova, W.-t. Yih // *Transactions of the Association for Computational Linguistics*. – 2017. – Vol. 5. – P.101-115.
- [21] **Ji, G.** Distant supervision for relation extraction with sentence-level attention and entity descriptions / G. Ji, L. Kang, H. Shizhu, Z. Jun // *Thirty-First AAAI Conference on Artificial Intelligence*. – 2017.

- [22] **Lin, Y.** Neural relation extraction with selective attention over instances / Y. Lin, S. Shen, Z. Liu, H. Luan, M. Sun // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers). – 2016. – P.2124-2133.
- [23] **Dasgupta, S.** Formal ontology learning from English is-a sentences / S. Dasgupta, A. Padia, G. Maheshwari, P. Trivedi, J. Lehmann // arXiv preprint arXiv: 1802.03701. – 2018.
- [24] **Petrucci, G.** Ontology learning in the deep / G. Petrucci, C. Ghidini, M. Rospocher // European Knowledge Acquisition Workshop. – Springer, Cham, 2016. – P.480-495.
- [25] **Rani, M.** Semi-automatic terminology ontology learning based on topic modeling / M. Rani, A.K. Dhar, O.P. Vyas // Engineering Applications of Artificial Intelligence. – 2017. – Vol. 63. – P. 108-125.
- [26] **Petrucci, G.** Expressive ontology learning as neural machine translation / G. Petrucci, M. Rospocher, C. Ghidini // Journal of Web Semantics. – 2018. – Vol. 52. – P.66-82.
- [27] **Chen, Y.** An active learning-enabled annotation system for clinical named entity recognition / Y. Chen, T.A. Lask, Q. Mei, Q. Chen, S. Moon, J. Wang, K. Nguyen // BMC medical informatics and decision making. – 2017. – Vol. 17, No. 2. – P.82.
- [28] **Shen, Y.** Deep active learning for named entity recognition / Y. Shen, H. Yun, Z.C. Lipton, Y. Kronrod, A. Anandkumar // arXiv preprint arXiv: 1707.05928. – 2017.
- [29] **Kim, S.** Mmr-based active machine learning for bio named entity recognition. / S. Kim, Y. Song, K. Kim, J.-W. Cha, G.G. Lee // In Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers, P.69–72.
- [30] **Kholghi, M.** Active learning: a step towards automating medical concept extraction / M. Kholghi, L. Sitbon, G. Zuccon, A. Nguyen // Journal of the American Medical Informatics Association. – 2015. – Vol. 23, No. 2. – P.289-296.
- [31] **Fang, M.** Learning how to active learn: A deep reinforcement learning approach / M. Fang, Y. Li, T. Cohn // arXiv preprint arXiv: 1708.02383. – 2017.
- [32] **Lin, C.H.** Re-active learning: Active learning with relabeling / C.H. Lin, M. Mausam, D.S. Weld // Thirtieth AAAI Conference on Artificial Intelligence. – 2016.
- [33] **Straka, M.** UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing / M. Straka, J. Hajic, J. Straková // Proceedings of the tenth international conference on language resources and evaluation (LREC 2016). – 2016. – P.4290-4297.
- [34] Universal Dependencies. - <https://universaldependencies.org/>.
- [35] **Zeman, D.** CoNLL 2018 shared task: multilingual parsing from raw text to universal dependencies / D. Zeman, J. Hajic, M. Popel, M. Potthast, M. Straka, F. Ginter, J. Nivre, S. Petrov // Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. – 2018. – P.1-21.
- [36] **Mikolov, T.** Efficient estimation of word representations in vector space / T. Mikolov, K. Chen, G. Corrado, J. Dean // arXiv preprint arXiv: 1301.3781. – 2013.

ACTIVE LEARNING APPROACH TO KNOWLEDGE EXTRACTION FROM DESCRIPTIONS OF EDUCATIONAL COURSES UNDER CONDITIONS OF SMALL DATA

J.D. Klenin

*Information Technologies Institute, Chelyabinsk State University, Chelyabinsk, Russia
Ugra Research Institute of Information Technologies, Khanty-Mansiysk, Russia
jklen@yandex.ru*

Abstract

With the constant growth of volumes of available data, their manual processing stops being possible, giving way to various machine learning models. Modern algorithms do a good job of basic tasks, provided that there is a sufficient amount of training data. However, many modern tasks are much more complicated and are highly specialized, which limits the amount of training data available for training, hindering the performance of fully automatic systems. In this paper, an approach to the task of automated fact extraction from the collections of raw text documents adapted for the lack of training data is presented. The integration of rule-based approaches for specific knowledge domains with generalized, domain-independent machine learning models pre-trained on large volumes of data is discussed. The proposed

approach based on the active learning methodology, seeks to reduce the expert's labor costs required for the efficient generation of extractable fact templates without compromising the system's performance. The paper also demonstrates the application of the proposed method of fact extraction based on the task of the target audience information search from the unstructured raw descriptions of online courses.

Key words: *knowledge graph, ontology, knowledge extraction, active learning, rule-based, machine learning, small data*

Citation: *Klenin J.D. Active learning approach to knowledge extraction from descriptions of educational courses under conditions of small data [In Russian]. *Ontology of designing*. 2019; 9(4): 522-535. – DOI: 10.18287/2223-9537-2019-9-4-522-535.*

Acknowledgments

The study is supported by the Russian Foundation for Basic Research in the framework of the project No. 18-47-860013 p_a "Intelligent system for the formation of educational programs based on neural network models of the natural language, taking into account the needs of the digital economy of the Khanty-Mansiysk Autonomous Okrug - Ugra" (contract No. 18-47- 860013 \ 18).

References

- [1] **He W, He W, Shen J, Tian X, Li Y, Akula V, Yan G, Tao R.** Gaining competitive intelligence from social media data: Evidence from two largest retail chains in the world. *Industrial Management & Data Systems*. 2015; 115(9): 1622-1636.
- [2] **Ahmad U.** Counter Terrorism on Online Social Networks Using Web Mining Techniques. *Intelligent Technologies and Applications: First International Conference, INTAP 2018, Bahawalpur, Pakistan, October 23-25, 2018, Revised Selected Papers*. – Springer, 2019; 932: 240.
- [3] **Toulouse T, Rossi L, Celik T, Akhloufi M.** Automatic fire pixel detection using image processing: a comparative analysis of rule-based and machine learning-based methods. *Signal, Image and Video Processing*. 2016; 10(4): 647-654.
- [4] **Ginneken B.** Fifty years of computer analysis in chest imaging: rule-based, machine learning, deep learning. *Radiological Physics and Technology*. 2017; 1(10): 23-32.
- [5] **Lukashevich, N.V.** SentiRuEval-2016: overcoming time gap and data sparsity in tweet sentiment analysis / N.V. Lukashevich, Y.V. Rubtsova // *Proceedings of International Conference Dialog-2016*. 2016; 416-426.
- [6] **Starostin, A.S.** FactRuEval 2016: evaluation of named entity recognition and fact extraction systems for Russian / A.S. Starostin, V.V. Bocharov, S.V. Alexeeva, A.A. Bodrova, A.S. Chuchunkov, S.S. Dzhumaev, I.V. Efimenko, D.V. Granovsk, V.F. Khoroshevsky, I.V. Krylova, M.A. Nikolaeva, I.M. Smurov, S.Y. Toldova. // *Proceedings of International Conference Dialog-2016*. 2016; 702-720.
- [7] **Abdi A, Idris N, Ahmad Z.** QAPD: an ontology-based question answering system in the physics domain. *Soft Computing*. 2018; 22(1): 213-230.
- [8] **Mohan MJ, Sunitha C, Ganesh A, Jaya A.** A study on ontology based abstractive summarization. *Procedia Computer Science*. 2016; 87: 32-37.
- [9] **Kontopoulos E, Martinopoulos G, Lazarou D, Bassiliades N.** An ontology-based decision support tool for optimizing domestic solar hot water system selection. *Journal of Cleaner Production*. 2016; 112: 4636-4646.
- [10] **Vrbanec T, Meštrović A.** The struggle with academic plagiarism: Approaches based on semantic similarity. 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). IEEE. 2017; 870-875.
- [11] **Strauss B, Toma B, Ritter A, De Marneffe MC, Xu W.** Results of the wnut16 named entity recognition shared task. *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*. 2016; 138-144.
- [12] **Limsopatham N, Collier NH.** Bidirectional LSTM for named entity recognition in Twitter messages. *Proceedings of the 2nd Workshop on Noisy User-generated Text*. 2016; 145-152.
- [13] **Sysoev AA, Andrianov IA.** Named entity recognition in Russian: the power of wiki-based approach. *Proceedings of International Conference "Dialogue"*. 2016; 746-755.
- [14] **Stepanova ME, Budnikov EA, Chelombeeva AN, Matavina PV, Skorinkin DA.** Information Extraction Based on Deep Syntactic-Semantic Analysis. *Proceedings of International Conference "Dialogue"*. 2016; 721-733.
- [15] **Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C.** Neural architectures for named entity recognition. *arXiv preprint arXiv: 1603.01360*. – 2016.

- [16] *Dernoncourt F, Lee JY, Szolovits P.* NeuroNER: an easy-to-use program for named-entity recognition based on neural networks. arXiv preprint arXiv: 1705.05487. – 2017.
- [17] *Corbett P, Boyle J.* Chemlistem: chemical named entity recognition using recurrent neural networks. Journal of Cheminformatics. 2018; 10(1): 59.
- [18] *Zeng X, He S, Liu K, Zhao J.* Large scaled relation extraction with reinforcement learning. Thirty-Second AAAI Conference on Artificial Intelligence. – 2018.
- [19] *Bai F, Ritter A.* //Structured Minimally Supervised Learning for Neural Relation Extraction. arXiv preprint arXiv: 1904.00118. – 2019.
- [20] *Peng N, Poon H, Quirk C., Toutanova K, Yih W-t.* Cross-sentence n-ary relation extraction with graph lstms. Transactions of the Association for Computational Linguistics. 2017; 5: 101-115.
- [21] *Ji G, Kang L, Shizh H, Jun Z.* Distant supervision for relation extraction with sentence-level attention and entity descriptions. Thirty-First AAAI Conference on Artificial Intelligence. – 2017.
- [22] *Lin Y, Shen S, Liu Z, Luan H, Sun M.* Neural relation extraction with selective attention over instances. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Long Papers. 2016; 2124-2133.
- [23] *Dasgupta S, Padia A, Maheshwari G, Trivedi P, Lehmann J.* Formal ontology learning from English is-a sentences. arXiv preprint arXiv: 1802.03701. – 2018.
- [24] *Petrucci G, Ghidini C, Rospoche M.* Ontology learning in the deep. European Knowledge Acquisition Workshop. Springer, Cham, 2016; 480-495.
- [25] *Rani M, Dhar AK, Vyas OP.* Semi-automatic terminology ontology learning based on topic modeling. Engineering Applications of Artificial Intelligence. 2017; 63: 108-125.
- [26] *Petrucci G, Rospoche M, Ghidini C.* Expressive ontology learning as neural machine translation. Journal of Web Semantics. 2018; 52: 66-82.
- [27] *Chen Y, Lask TA, Mei Q, Chen Q, Moon S, Wang J, Nguyen K.* An active learning-enabled annotation system for clinical named entity recognition. BMC medical informatics and decision making. 2017; 17(2): 82.
- [28] *Shen Y, Yun H, Lipton ZC, Kronrod Y, Anandkumar A.* Deep active learning for named entity recognition. arXiv preprint arXiv: 1707.05928. – 2017.
- [29] *Kim S, Song Y, Kim K, Cha J-W, Lee GG.* Mmr-based active machine learning for bio named entity recognition. In Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers. 69–72.
- [30] *Kholghi M, Sitbon L, Zuccon G, Nguyen A.* Active learning: a step towards automating medical concept extraction. Journal of the American Medical Informatics Association. 2015; 23(2): 289-296.
- [31] *Fang M, Li Y, Cohn T.* Learning how to active learn: A deep reinforcement learning approach. arXiv preprint arXiv: 1708.02383. – 2017.
- [32] *Lin C.H, Mausam M, Weld DS.* Re-active learning: Active learning with relabeling. Thirtieth AAAI Conference on Artificial Intelligence. – 2016.
- [33] *Straka M, Hajic J, Straková J.* UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing. Proceedings of the tenth international conference on language resources and evaluation (LREC 2016). 2016; 4290-4297.
- [34] Universal Dependencies. - <https://universaldependencies.org/>.
- [35] *Zeman D, Hajic J, Popel M, Potthast M, Straka M, Ginter F, Nivre J, Petrov S.* CoNLL 2018 shared task: multilingual parsing from raw text to universal dependencies. Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. 2018; 1-21.
- [36] *Mikolov T, Chen K, Corrado G, Dean J.* Efficient estimation of word representations in vector space. arXiv preprint arXiv: 1301.3781. – 2013.

Сведения об авторе



Кленин Юлий Дмитриевич, 1994 г. рождения. Окончил магистратуру Челябинского государственного университета по направлению «Фундаментальная информатика». Является аспирантом по специальности «Системный анализ и управление» (Институт информационных технологий Челябинского государственного университета). Автор более 15 статей по тематике интеллектуального анализа текстовой информации.

Klenin Julius Dmitrievich (b. 1994) master of “Fundamental Computer Science” (Chelyabinsk State University). Postgraduate student in “System analysis and control” (Information Technologies Institute at Chelyabinsk State University). He is an author of about 15 papers on intelligent analysis of text data.