Открытое извлечение информации из текстов Часть I. Постановка задачи и обзор методов*

А.О. Шелманов, В.А. Исаков, М.А. Станкевич, И.В. Смирнов

ИСА ФИЦ ИУ РАН, Москва, Россия

Аннотация. В статье представлена постановка задачи открытого извлечения информации. Выполнен аналитический обзор работ в этой области, а также обзор смежных работ по извлечению из текстов сущностей и семантических отношений, в которых применяется машинное обучение с частичным привлечением учителя и без него. Рассмотрены предполагаемые направления исследований в области открытого извлечения информации из текстов, основанные на машинном обучении без учителя.

Ключевые слова: открытое извлечение информации, семантические отношения, извлечение терминов, машинное обучение без учителя, машинное обучение с частичным привлечением учителя.

DOI 10.14357/20718594180204

Введение

Проблема извлечения информации из текстов на естественном языке остается на сегодняшний день весьма актуальной. Извлечение информации востребовано во многих типах прикладных программных систем, среди которых информационно-поисковые и вопросно-ответные системы, программы автоматического реферирования и аналитической обработки коллекций документов, системы автоматического построения и пополнения баз знаний (БЗ) и др.

Существует большое количество традиционных методов извлечения информации, основанных на правилах, тезаурусах, методах машинного обучения с учителем. Для таких методов необходимо достаточно большое количество априорных знаний о рассматриваемой предметной области, а также объемные лингвистические ресурсы: размеченные корпуса, словари, тезаурусы, системы правил и грамматики.

Для ручного построения правил и грамматик разработано множество инструментов и технологий, которые в настоящей момент активно применяются в коммерческих продуктах: CPSL [1], Jape [2], LSPL [3], UIMA Ruta [4], ABBYY Compreno [5], Tomita parser[1] и др. Главные недостатки этих технологий и такого подхода в целом заключаются в больших затратах на разработку правил и грамматик, отсутствии гибкости при их переносе на другие предметные области (для новых предметных областей и новых типов извлекаемых объектов необходимо обновлять или создавать правила и грамматики заново), жестко заданном и, как правило, небольшом проблемно-ориентированном наборе извлекаемых отношений.

*Исследование выполнено за счет гранта Российского научного фонда (проект №14-11-00692).

[1]https://tech.yandex.ru/tomita/

✉ *Шелманов Артем Олегович E-mail: shelmanov@isa.ru*

Методы извлечения информации на основе машинного обучения с учителем позволяют несколько снизить трудозатраты на создание систем анализа текстов. Вместо ручного построения правил, которое требует привлечения квалифицированных экспертов, для обучения моделей машинного обучения необходимы размеченные текстовые корпуса, которые могут создаваться, например, за счет краудсорсинга. Среди множества работ, посвященных методам и системам извлечения информации на основе обучения с учителем, можно выделить [6-10].

Хотя эти методы позволяют упростить разработку систем извлечения информации из текстов, создание размеченных корпусов для их обучения остается весьма затратным, что также делает трудоемким адаптацию систем извлечения информации к новым предметным областям, поскольку для этого требуется создание новых размеченных корпусов текстов. Машинное обучение с учителем также не решает проблему, заключающуюся в том, что набор извлекаемых из текстов отношений изначально жестко ограничен схемой разметки корпуса.

Таким образом, подходы на основе правил, грамматик или обучении с учителем обычно не позволяют быстро настроиться на извлечение из текстов новых типов информации и знаний, а также, как правило, оказываются малополезными, когда необходимо решить задачи извлечения информации в новой предметной области, для которой лингвистические ресурсы отсутствуют, либо имеют малую репрезентативность.

Эти проблемы могут быть решены с помощью подходов, которые в англоязычной литературе принято объединять в область исследований под названием «открытое извлечение информации» (open information extraction) [11]. Этот термин может применяться в узком смысле, обозначая подходы к извлечению из текстов «поверхностных» связей или триплетов: пар сущностей и фрагмента текста, указывающего на наличие какого-либо семантического отношения между ними. Однако в настоящей работе этот термин рассматривается в широком смысле, подразумевая, что в рамках открытого извлечения информации объединяются разнообразные подходы, ориентированные на извлечение разнородной информации из неструктурированных текстов без привязки к конкретной предметной области, а также с минимальным привлечением ручного труда. Многие из таких подходов основаны на методах машинного обучения с частичным привлечением учителя и без учителя. Стоит отметить, что качество работы подобных методов существенно меньше по сравнению с методами на основе машинного обучения с учителем. Однако они востребованы в тех информационно-аналитических системах, в которых принципиально отсутствуют ограничения на предметную область, а также ограничения на типы извлекаемой информации (например, в системах автоматического пополнения баз знаний).

В статье проведен аналитический обзор методов открытого извлечения информации на основе машинного обучения с частичным привлечением учителя и без учителя. Статья структурирована следующим образом. В первом разделе представлены методы и подходы для извлечения сущностей и поверхностных связей, во втором – методы и подходы для выявления семантических отношений. В третьем разделе рассматриваются направления наших дальнейших исследований по теме извлечения знаний из текстов. В заключении резюмируются результаты работы.

1. Извлечение сущностей и поверхностных связей

Одной из подзадач в области открытого извлечения информации является извлечение сущностей и поверхностных связей из текстов (триплетов). Триплетом или поверхностной связью будем называть пару сущностей и фрагмент текста, указывающий на наличие между этими сущностями некоторой семантической связи. Сущности этого триплета будем называть аргументами связи, например:

– исходный текст: «*Sugar beets grow in temperate zone*»;

– выделенный триплет: («*Sugar beets*»; «*grow in*»; «*temperate zone*»).

Основополагающей для области открытого извлечения информации считается работа [12]. В ней представлена одна из первых систем такого рода – TextRunner. Система использует машинное обучение по автоматически сгенерированной разметке. Обучающая выборка генерируется на основе синтаксически размеченного корпуса Penn Treebank [13] с помощью эвристик. Для этого выделяются наименьшие

именные группы, и если текст между ними удовлетворяет ряду условий (среди них, например, ограничение на длину пути в синтаксическом дереве между аргументами; проверка того, что оба аргумента находятся в одном предложении; проверка того, что аргументы состоят не только лишь из одного местоимения), то эти именные группы и текст между ними образуют обучающий пример в виде триплета. В качестве признаков для обучения используются: последовательность частей речи в синтаксическом пути между аргументами, длина пути, количество стоп-слов между аргументами, части речи аргументов и др. В качестве метода классификации в оригинальной статье использовался наивный байесовский классификатор. Дальнейшие работы в данном направлении показали, что использование условных случайных полей [14] или марковских логических сетей [15] вместо наивного байесовского классификатора повышает качество извлечения триплетов. Стоит также отметить систему Woe [16], в которой была предложена модификация подхода, реализованного в TextRunner, значительно повышающая точность и полноту извлечения триплетов. В Woe для выделения аргументов используются метаданные из информационных таблиц статей Wikipedia. Улучшения, реализованные в наиболее эффективном алгоритме Woe, дают от 18 до 34% повышения F_1-меры по сравнению с результатами TextRunner. В работе [17] также используется Wikipedia для генерации обучающих примеров. Для генерации авторы предложили использовать предложения, содержащие концепт, которому посвящена статья, а также гиперссылку на другой концепт. Соответствующий ссылке фрагмент текста, а также сущность, которой посвящена статья, формируют аргументы пары, между которыми установлено отношение. Подобный подход позволяет выделить пары сущностей, между которыми потенциально имеется некоторое семантическое отношение.

Вышеописанные системы имеют два значительных недостатка: они могут выделять триплеты, для которых нет адекватной семантической интерпретации, а также – неинформативные триплеты, в которых упущена ключевая информация. Система ReVerb [18] нацелена на устранение данных недостатков. В ней, в отличие от подходов, предложенных ранее, выделение связей происходит перед выделением аргументов, что значительно улучшает качество извлечения триплетов. Например, более ранние системы не способны выделить связь «*claimed responsibility for*», потому что слово «*responsibility*» выделяется как аргумент.

В ReVerb сначала в качестве связей извлекаются последовательности слов, у которых части речи соответствуют заданным шаблонам (*глагол*; *глагол + предлог/частица* и др.). Если подстроки, удовлетворяющие шаблонам, пересекаются, то выбирается наибольшая подстрока, а если несколько соседних последовательностей удовлетворяют шаблонам, то они объединяются в одну связь. Подобный подход, однако, приводит к тому, что в результате могут получаться длинные и малополезные связи, например, такие как «*is offering only modest greenhouse gas reduction targets at*». Для фильтрации подобных связей в работе применяется классификатор на основе логистической регрессии. В качестве аргументов в предложениях просто извлекаются именные группы слева и справа от выделенной связи.

Создателями ReVerb была также разработана система ArgLearner [19], которая была ориентирована на повышение точности выделения аргументов. В ней используется три классификатора, определяющих границы аргументов. Второй аргумент практически всегда следует за связью, поэтому левая граница второго аргумента выделяется без классификации. Для идентификации правой границы первого аргумента используется дерево принятия решений, а для остальных границ – классификаторы на основе условных случайных полей. Для генерации обучающей выборки применяется набор вручную созданных шаблонов. В качестве признаков используются длина предложения, части речи, пунктуация и т.д. Отметим, что ReVerb и ArgLearner были объединены в новую систему под названием R2A2.

Ollie – наиболее поздняя из систем, созданных коллективом исследователей, разработавших ReVerb и ArgLearner [20]. Она имеет два преимущества по сравнению со своими предшественниками. Эта система позволяет выделять также триплеты, в которых связь между двумя аргументами осуществляется за счет существительных или прилагательных, например, в предложении «*Microsoft co-founder Bill Gates spoke at ...*» выделяется триплет («*Bill Gates*»; «*be co-*

founder of»; «*Microsoft*»). Другой особенностью Ollie стала возможность учитывать контекст в выделяемых триплетах и добавлять к ним специальные модификаторы, которые свидетельствуют об условии, при котором выделенная связь имеет место быть. Например, в предложении «*If he wins five key states, Romney will be elected President*» данная система выделит триплет ((«*Romney*»; «*will be elected*»; «*President*») *ClausalModifier if*; «*he wins five key states*»).

Сам процесс можно разделить на два основных этапа: сначала подготавливается множество разнообразных шаблонов, которые используются для выделения триплетов из предложений, а затем добавляются модификаторы, учитывающие контекст. На первом этапе используется предыдущая система ReVerb. На втором этапе в результате анализа синтаксических деревьев собранных предложений [21] создаются шаблоны для открытого извлечения. Шаблоны выделяются путем обобщения используемых связей, аргументов и предлогов в синтаксической структуре предложения. В проведенных экспериментах с помощью ReVerb было извлечено 110 000 триплетов из набора данных ClueWeb[2]. Затем для каждого триплета из веб корпуса были извлечены все предложения, в которых встречались оба аргумента и предикат, что дало 18 млн новых предложений. На основе набора эвристик исследователи отфильтровали этот набор, в результате было построено 4 млн новых обучающих примеров для генерации шаблонов.

После генерации шаблонов с их помощью можно извлекать простые триплеты из предложений. На этом этапе добавляются модификаторы контекста, которые составляются за счет различных дополнительных эвристик, примененных к синтаксическому дереву предложения. Например, для выявления модификатора условия слова типа («*if*», «*when*», «*although*», «*because*» и др.) ищутся в определенном узле синтаксического дерева.

Описанная в работе [22] система открытого извлечения информации немного отходит от концепции выделения триплетов из текста с помощью обширного набора сгенерированных шаблонов, как, например, это реализовано в системе Ollie. Авторы этой системы сначала разделяют предложения на небольшие клаузы, из которых в дальнейшем легко выделяются триплеты при помощи небольшого количества простых шаблонов. Для решения задачи разбиения цельного предложения на клаузы авторы системы обучали классификатор на заранее подготовленном наборе данных, в который входили предложения, полученные при помощи методов опосредованного обучения из [23, 24], а также предложения из вручную размеченного корпуса [25].

Для выделения клауз авторы системы обращаются к синтаксической структуре предложений и рассматривают различные варианты возможных действий при проходе по дереву. В частности, анализируются переходы между узлами (связями/глаголами) в синтаксическом дереве. Так, например, в предложении «*Born in a small town, she took the midnight train going anywhere*» присутствует дуга «*took*» → «*Born*» (где «*took*» будет вершиной дерева, а «*Born*» потомком), а целью системы является выделение следующих клауз: «*she took midnight train*» и «*she Born in town*». Задача классификатора состоит в том, чтобы на каждом шаге при обходе синтаксического дерева выбирать одно из действий: выделить клаузу на дочернем узле и перейти к этому узлу (*Split*), перейти по дуге к дочернему узлу и не выделять клаузу (*Recurse*), не переходить по дуге (*Stop*). На основе начального набора данных создается множество различных последовательностей обхода синтаксического дерева, которые состоят из действий *Split*, *Recurse* и *Stop*. Если последовательность действий приводит к извлечению известного триплета (содержащегося в исходном наборе данных), все элементы этой последовательности маркируются действием *Recurse*, кроме последнего, который маркируется как *Split*. Последний переход, не приводящий к извлечению триплета, маркируется действием *Stop*. В качестве признаков используется информация о дуге синтаксического дерева: вид связи между узлами [26], последнее ребро, по которому был осуществлен переход, соседи родительского узла, количество исходящих ребер дочернего узла, информация о передаче субъекта или объекта из родительского узла дочернему, часть речи родительского и дочернего узла и др.

После завершения процесса выделения клауз при помощи классификатора, полученные клаузы фильтруются на основе эвристик для того, чтобы привести их в наиболее ком-

[2] http://lemurproject.org/clueweb09.php/

пактную форму, содержащую основную информацию. Например, в предложении «*all rabbits eat vegetables»* будет отброшено уточнение «*all»,* так как данный модификатор обобщает аргумент «*rabbits».* Извлеченный триплет будет иметь следующий вид:«*rabbits»; «eat»; «vegetables»*.

В итоге из построенных таким образом клауз при помощи 14 простых шаблонов выделяются сами триплеты. Следует отметить, что данная система сравнивалась с другими системами открытого извлечения на соревновании TAC-KBP 2013 Slot Filling task [24] и показала лучший результат, существенно превзойдя Ollie (наилучший сравнимый результат – $F_1 = 22{,}7\%$ против $F_1 = 19{,}6\%$, при этом максимальный результат рассматриваемой системы составил $F_1 = 28{,}3\%$).

Стоит также отметить систему PATTY [27]. В ней для выделения бинарных отношений берется кратчайший путь в синтаксическом дереве между двумя именованными сущностями, которые присутствуют в наборе данных. Авторы этой системы обращают внимание на то, что подход с обработкой синтаксического дерева дополняет подход с использованием автоматической морфологической разметки и увеличивает точность извлечения отношений.

2. Выявление семантических отношений между сущностями

Другой подзадачей в области открытого извлечения информации является выявление семантических отношений между сущностями. Из текстов обычно извлекается очень много поверхностных связей. При этом многие из них выражают одно и то же семантическое отношение. Например, из предложений «*Тим Кук руководит Apple*» и «*Тим Кук является главой Apple*» можно выделить две поверхностные связи «*руководит*» и «*является главой*», которые имеют одинаковый смысл. Для решения многих задач, в том числе для пополнения баз знаний, необходимо группировать семантически близкие связи в абстрактные семантические отношения. Основные предъявляемые к ним требования заключаются в том, что эти абстракции должны либо приближенно соответствовать схеме семантических отношений в какой-либо базе знаний, либо соответствовать некоторым требованиям прикладной задачи.

Подходы к решению этой задачи можно разделить на три группы: опосредованное обучение (distant supervision), итерационное расширение обучающей выборки (bootstrapping), а также кластеризация поверхностных связей. Первые два подхода являются подвидами машинного обучения с частичным привлечением учителя, а последний – подвид машинного обучения без учителя.

Опосредованное обучение заключается в том, что обучающая выборка для моделей машинного обучения создается не вручную, а автоматически на основе другого ресурса. В работе [28] предлагается метод для автоматического построения обучающего корпуса для задачи извлечения семантических отношений на основе базы знаний Freebase [29]. В большом неразмеченном корпусе текстов автоматически выявляются именованные сущности, которые сопоставляются с концептами в базе знаний. Если две сущности встречаются в одном предложении, и между ними в базе знаний существует отношение, то формируется положительный обучающий пример, в котором контекст пары сущностей становится совокупностью признаков этого отношения. Несмотря на то, что такой подход генерирует большое количество шумовых примеров, этот недостаток в значительной степени нивелируется большим объемом построенной таким образом выборки. Авторы использовали два вида признаков: лексические (морфологические теги и леммы слов контекста, в котором встречаются именованные сущности) и синтаксические (путь в синтаксическом дереве от одной сущности до другой). Для обучения системы извлечения отношений использовалась логистическая регрессия. Средняя точность (усредненная по типам отношений) на уровне 100 и 1000 при применении такого подхода составляет – 69 и 68% соответственно. Полученные результаты свидетельствуют о том, что представленный подход позволяет с минимальным использованием ручного труда создать систему с удовлетворительным качеством извлечения многих семантических отношений.

Этот подход был развит в работах [30-32], где было предложено рассматривать проблему извлечения отношений не как классическое обучение по прецедентам, а как задачу «многоэкземплярного» обучения (multiple-instance learning) [33]. Постановка этой задачи предпо-

лагает, что все множество обучающих примеров разбито на подмножества, для каждого из которых известно лишь, что оно либо содержит хотя бы один положительный пример, либо не содержит ни одного положительного. При этом важно обучить модель определять класс не каждого обучающего примера в отдельности, а целой группы. Такая постановка соотносится с задачей определения наличия семантического отношения между заданными сущностями на основе множества текстовых примеров (некоторые из которых могут быть ошибочными в том смысле, что в них нет информации о наличии подобного отношения). Таким образом, предположение, сделанное в работе [28], о том, что каждое предложение, содержащее пару сущностей, выражает отношение, заданное между ними в БЗ, заменяется на более мягкое предположение о том, что если в БЗ имеется отношение, то мы выбрали достаточно репрезентативный корпус такой, что найдется хотя бы одно предложение, в котором выражено это семантическое отношение. В подобных подходах создается модель, которая описывает не каждый экземпляр, состоящий из пары сущностей и их общего контекста, а целые группы экземпляров для одних и тех же пар сущностей. Так, в [30] предложена байесовская модель с двумя типами скрытых переменных, которые связаны между собой: переменные-отношения между сущностями (в базе знаний), переменные-отношения между сущностями в конкретном предложении (в экземпляре). Однако эта модель не позволяет извлекать несколько разных типов отношений между парой сущностей. В [31] предложена модификация, которая это ограничение снимает. Схожая модель с некоторыми улучшениями предложена также в работе [32].

Другой подход, основанный на обучении векторных представлений, предложен в [34]. В этой работе исследователи используют Freebase не только для построения набора обучающих примеров, но также и для обучения векторных представлений именованных сущностей и связей, которые отражают взаимосвязи между сущностями в базе знаний. Метод предполагает обучение двух независимых моделей: для векторных представлений связей в тексте и для векторных представлений сущностей и связей во Freebase. Первая модель позволяет по набору признаков пары сущностей в тексте (часть речи, путь в синтаксическом дереве, тип именованных сущностей) определить вектор потенциального отношения, который затем сравнивается по косинусной мере с векторными представлениями типов отношений во Freebase. Вторая модель, опираясь на связность сущностей в базе знаний, позволяет по векторному представлению одной сущности e_1 и отношению r получить векторное представление второй сущности e_2 путем простого линейного преобразования представлений $(e_1 + r = e_2)$. При определении типа отношения для заданной пары сущностей сначала выбирается наиболее похожий тип отношения в векторном пространстве первой модели (включая тип «*NA*», который кодирует отсутствие связи), затем оценка близости для связей, отличных от «*NA*», пересчитывается с учетом оценки сходства по второй модели. В экспериментах исследователи показали, что использование второй модели, опирающейся на связность концептов в базе знаний, дает существенный прирост (до 7% по точности при фиксированном уровне полноты в диапазоне от 0 до 10%) по сравнению с использованием только первой модели векторных представлений семантических связей.

В области опосредованного обучения стоит также отметить работу [35]. В ней исследователи предложили совместить поверхностные связи, извлекаемые с помощью систем типа TextRunner, а также схемы различных БЗ в единой БЗ с «универсальной» схемой. Кроме того, был предложен метод для пополнения такой БЗ: на основе наблюдаемых фактов он позволяет предсказывать наличие новых семантических отношений между сущностями. В его основе лежат методы факторизации матриц и коллаборативной фильтрации. БЗ в этом подходе представляется в виде матрицы, строки которой обозначают пары сущностей, а столбцы – различные связи (как поверхностные связи, так и семантические отношения в схеме некоторой БЗ). Задача представленного метода состоит в заполнении элементов матрицы вероятностями наличия семантического отношения в соответствующем столбце. Для оценки вероятности авторы предложили комплексную модель, учитывающую схожесть сущностей, а также контекстов, в которых они встречаются.

Большой интерес исследователи в области обработки текстов на естественном языке проявляют к глубокому машинному обучению на низкоуровневых признаках, таких как вектор-

ные представления слов. Многие исследователи отмечают, что такой подход, с одной стороны, избавляет от трудоемкого процесса разработки признакового пространства, который обычно проводится при применении других методов машинного обучения, а с другой стороны, позволяет нивелировать ошибки, возникающие на разных этапах лингвистического анализа (морфологического, синтаксического, семантического и др.), поскольку глубокий многоэтапный анализ при таком подходе не требуется. В недавних работах глубокие нейросетевые архитектуры были применены к задаче извлечения семантических отношений совместно с опосредованным обучением.

В работе [36] представлена архитектура сверточной нейронной сети для извлечения отношений на основе опосредованного обучения по базе знаний Freebase. Модель не использует лингвистические признаки, а опирается на векторные представления слов, заранее построенные по большому неразмеченному корпусу, а также на расстояния слов до связываемых сущностей. Архитектура сети состоит из сверточного слоя, «кусочного» слоя выбора максимума, а также слоя экспоненциальной нормализации (softmax). Главная особенность архитектуры заключается в том, что применяется не стандартный слой выбора максимума, который выбирает максимумы по всему входному вектору, а «кусочный» слой, выбирающий максимумы в рамках определенных отрезков текста. В частности, исследователи разбили все предложения, в которых встречается пара сущностей, на три отрезка, где выбирает максимумы «кусочный» слой: слева от левой сущности, между сущностями, справа от правой сущности. Экспериментальные исследования предложенного подхода показали его существенное преимущество перед методами, представленными в работах [28, 31, 32] (в проведенных экспериментах точность, усредненная по трем уровням (топ 100, 200, 500, составляет 78,3% против 67,6, 72,0, 73,7% соответственно).

В [37] также предлагается обучаться на корпусе, автоматически размеченном с использованием Freebase. В качестве признаков используются только векторные представления слов. Авторы предложили рекуррентную нейросетевую архитектуру с двух уровневым механизмом внимания (attention mechanism). Рекуррентный слой с ячейками долго-краткосрочной памяти (LSTM) [38] используется для первоначального кодирования слов контекста, в котором находится пара связываемых сущностей. Выходы рекуррентной сети на словах контекста затем пропускаются через первый механизм внимания, который определяет важность слов контекста для решения задачи. Построенные в результате применения механизма внимания векторные представления слов усредняются, формируя тем самым вектор контекста обучающего примера. Затем полученные таким образом векторные представления контекстов проходят через второй механизм внимания, который определяет важность примера в группе. Этот механизм частично позволяет решить возникающую в опосредованном обучении проблему, которая заключается в том, что часть сгенерированных обучающих примеров являются ложноположительными. Предложенный подход в экспериментах продемонстрировал более высокое качество извлечения семантических отношений по сравнению с другими методами, включая подход, использующий сверточную архитектуру, представленную в [36] (точность до 2% выше на одном и том же уровне полноты).

Принцип итерационного расширения обучающей выборки заключается в том, что по небольшому набору вручную размеченных стартовых примеров формируются новые обучающие примеры. Одной из первых работ, в которых этот принцип был применен к задаче извлечения семантических отношений, является [39]. В ней был предложен подход к автоматическому формированию шаблонов извлечения отношений в виде регулярных выражений. Как и в случае с опосредованным обучением набор заданных отношений между сущностями используется для разметки отношений в масштабной текстовой коллекции. Контексты пар сущностей в свою очередь используются для генерации шаблонов, которые в конечном итоге позволяют достаточно точно находить заданные стартовым набором отношения между произвольными сущностями. В работе [40] представлена модификация этого подхода, которая получила название Snowball. Авторы предложили новый метод для генерации шаблонов и извлечения экземпляров отношений, а также стратегию для оценки их достоверности. В Snowball набор стартовых примеров кластеризуется для выявления множества извлекае-

мых семантических отношений. Оценка близости экземпляров отношений рассчитывается как косинус угла между векторами TF-IDF слов из контекста пар связанных сущностей.

Под контекстом понимается наборы слов, стоящие слева и справа от связываемых сущностей, а также набор слов между этими сущностями. Центроиды получившихся таким образом кластеров представляют собой начальные признаковые описания отношений (шаблоны отношений). Шаблоны сопоставляются с текстом, в результате чего выделяется набор новых экземпляров отношений. В этих экземплярах могут присутствовать новые пары сущностей, которых не было в стартовом наборе. По извлеченным экземплярам оцениваются и выбираются наиболее точные шаблоны. Извлеченные экземпляры также получают оценку на основе близости их вектора признаков к вектору признаков шаблона и достоверности его самого. Алгоритм на каждой итерации отбирает наиболее достоверные шаблоны и экземпляры отношений, постепенно расширяя обучающую выборку и обобщая шаблоны. В результате алгоритм учится устанавливать семантические отношения, заданные небольшим стартовым набором примеров, между произвольными сущностями в разнообразных контекстах. Этот подход был развит в работе [41]. Авторы использовали тот же алгоритм, однако вместо TF-IDF в качестве признаков для оценки схожести шаблонов и экземпляров они использовали векторные представления слов контекста, вычисленные с помощью word2vec [42]. Авторы получили значительное улучшение по сравнению с оригинальной моделью (например, по отношению «*acquired*» F_1-мера была увеличена с 26 до 75%).

Опосредованное обучение и итерационное расширение обучающей выборки позволяют создавать достаточно устойчивые системы для извлечения отношений из текстов и не требуют большого количества ручного труда. Однако, как и в подходах, основанных на машинном обучении с учителем, они позволяют обучаться лишь заранее заданному набору отношений, которые уже присутствуют либо в базе знаний, либо в стартовом наборе примеров. Эту проблему призваны решить подходы к выделению семантических отношений без привлечения учителя. В них принципиально отсутствует какая-либо ручная разметка или исходная база знаний. Семантические отношения формируются кластеризацией контекстов связываемых сущностей.

В работе [17] авторы извлекают поверхностные связи с помощью разметки на основе Wikipedia, а затем кластеризуют их на основе синтаксических признаков, извлеченных из предложений, в которых встречаются пары сущностей, а также из поверхностных шаблонов, полученных с использованием поисковой машины в Интернете (в частности, исследователи использовали Google). Для получения данных, необходимых для построения поверхностных шаблонов, формируется поисковый запрос, используя глаголы и существительные из предложения, в котором встретилась соответствующая пара сущностей. Полученные из поисковой выдачи сниппеты используются для генерации поверхностных шаблонов вида: *<последовательность слов1> аргумент1 <последовательность слов2> аргумент2 <последовательность слов3>*. Для кластеризации был предложен метод, основанный на алгоритме k-средних, в котором при расчете расстояния используются и синтаксические признаки, и поверхностные шаблоны.

В работе [43] для кластеризации предлагается использовать подходы тематического моделирования, основанные на латентном размещении Дирихле (LDA) [44]. Авторы разработали несколько моделей с разным набором признаков. В них место документов занимают совокупности контекстов одинаковых пар именованных сущностей, а место тем занимают семантические отношения. В работе [45] представлено развитие этой идеи. Наряду с тематическими моделями в ней предлагается использовать принципы иерархической агломеративной кластеризации. Метод, представленный в этой работе, позволяет частично решить проблему полисемии связываемых сущностей.

Еще одна модификация подхода, основанная на тематическом моделировании с помощью LDA, представлена в работе [46]. В ней исследователи совместили метод тематического моделирования для выделения семантических отношений с «внешними» ограничениями, заданными предикатами первого порядка. Был предложен алгоритм, позволяющий построить тематическую модель, которая нежестко удовлетворяет заданным ограничениям, содержащим некоторые «внешние» знания о предметной области. Для этого исследователи модифицировали модель First-Order Logic La-

tent Dirichlet Allocation, представленную в [47], а для настройки параметров применили представленный в этой же работе один из вариантов EM-алгоритма. В работе был также предложен способ для автоматического построения ограничений без учителя, основанный на анализе статистической совместной встречаемости признаков друг с другом. В нем генерируется два вида логических правил: правила, указывающие на необходимость отнесения обучающих примеров-связей в виде пар аргументов с заданными признаками (включая, например, путь в синтаксическом дереве от одного аргумента до другого) к единому кластеру, и правила, указывающие на запрет создания такого кластера. Первый тип правил строится на основе статистической оценки важности совместной встречаемости признаков в едином предложении. Наиболее часто встречающиеся пары признаков формируют логические правила, указывающие на необходимость формирования кластера связей. Второй тип правил строится путем поиска пар признаков, которые никогда друг с другом не встречаются в одном предложении. Такие правила запрещают относить связи, признаки которых никогда друг с другом не встречаются, в один кластер. Исследователи показали, что применение логических ограничений первого типа в тематических моделях повышает качество извлечения отношений по сравнению с моделью LDA на связях.

В [48] представлен метод для кластеризации поверхностных связей (шаблонов), основанный на дистрибутивной гипотезе, в которой связи, описывающие семантически близкие отношения, имеют схожее распределение связываемых сущностей. Исследователи указывают на две проблемы применения такого подхода к большим данным: трудоемкость расчета совместной встречаемости связей и пар сущностей, а также высокое потребление памяти. Для преодоления этих трудностей в работе предлагаются методы аппроксимации подсчета частот и снижения размерности.

В статье [49] для извлечения семантических отношений без учителя предлагается подход, основанный на минимизации ошибки восстановления входных данных (reconstruction error minimization). Схожий подход, например, используется в нейросетевых автокодировщиках: скрытые слои нейронных сетей обучаются так, чтобы из них можно было восстанавливать входные данные с наименьшей ошибкой. При этом размерность скрытых слоев обычно меньше, чем размерность входных данных. Авторы строят модель порождения семантического отношения при заданных аргументах и их контекстах (кодировщик) и модель порождения аргумента при заданном отношении и заданном другом аргументе (декодировщик). Параметры моделей обучаются совместно с помощью градиентного метода с адаптивным шагом. Обученный кодировщик используется для предсказания отношений между аргументами, а декодировщик является побочным результатом работы метода. Отметим, что в декодировщике авторы используют модель тензорной факторизации, генерирующую в процессе обучения векторные представления именованных сущностей, которые можно использовать в других задачах, например, для кластеризации именованных сущностей.

3. Предлагаемые методы и подходы

Извлечение семантических отношений из текстов предлагается разделить на три этапа: 1) извлечение из текстов сущностей и терминов; 2) извлечение поверхностных связей (триплетов) – пар сущностей, между которыми потенциально присутствует отношение, и фрагментов текстов, указывающих на наличие этого отношения; 3) группировка выделенных поверхностных связей с общей семантикой, из которых можно сформировать семантическое отношение.

На первом этапе предлагается генерировать из синтаксических деревьев предложений большое количество вариантов сущностей – фрагментов текста, которые соответствуют предметам, процессам или явлениям из реального мира. При этом нас интересуют не все факты, присутствующие в тексте, а только те связи и отношения, в которых участвуют термины предметной области – важные сущности для предметной области анализируемых текстов. Для оценки важности сущностей предлагается опираться на известные статистические подходы – различные модификации TF*IDF, и на небольшой набор эвристик, которые отфильтровывают заранее неперспективные варианты. Кроме того, при построении БЗ после извлечения сущностей необходима их нормализация, чтобы фрагментам с разной поверхностной

лексико-синтаксической структурой сопоставить одну абстрактную сущность. Например, сущности «*практикующий юрист*» и «*практический деятель в области права*» необходимо сгруппировать. Для этого предлагается применить подходы агломеративной иерархической кластеризации именных групп.

Для второго этапа извлечения поверхностных связей между выделенными сущностями предлагается адаптировать подходы систем типа ReVerb [18] к русскому языку. Планируется модифицировать эвристики, заложенные в этих системах с учетом флективности и свободного порядка слов в русском языке.

На третьем этапе группировку поверхностных связей предлагается осуществлять методами машинного обучения без учителя путем кластеризации признаковых описаний пар сущностей и их контекстов. Поверхностные связи, попавшие в один кластер, можно считать экземплярами некоторого неименованного отношения, выраженного в тексте. Если поверхностной связи не соответствует ни один из кластеров, то это означает, что отношение между парами сущностей устанавливать не следует.

Во многих методах извлечения отношений из текстов на основе машинного обучения без учителя для подобной кластеризации используются вероятностные модели со скрытыми переменными. Экспериментальные оценки качества работы таких моделей пока что значительно уступают оценкам методов на основе машинного обучения с учителем. Одним из возможных путей повышения качества моделей без учителя является расширение количества используемых признаков. Однако вместе с этим происходит усложнение структуры модели, что приводит к повышению трудоемкости ее обучения. Альтернативой усложнению структуры вероятностной модели является использование аддитивной регуляризации. Этот подход состоит в том, что помимо правдоподобия оптимизируются также дополнительные критерии-регуляризаторы, не обязательно вероятностной природы. Как было показано ранее в [50] на примере регуляризации тематических моделей, такой подход позволяет строить многоцелевые модели, удовлетворяющие большому количеству дополнительных ограничений, либо композитные модели, объединяющие в себе несколько более простых моделей без значительного увеличения сложности алгоритмов обучения и вывода.

В ходе работы над проектом предлагается разработать новые методы извлечения семантических отношений, основанные на моделях со скрытыми переменными, в которых в отличие от традиционных вероятностных моделей, применяется аддитивная регуляризация. Предполагается, что с помощью такого подхода удастся интегрировать в модели признаки различных модальностей, за счет чего повысить качество извлечения отношений. В признаковые описания предлагается включить векторные представления слов контекста. Как показано в последних работах, подобные представления хорошо описывают смысловую близость слов [51], что предположительно является важной информацией при определении семантической похожести поверхностных связей. В данной работе предлагается исследовать новый подход к построению векторных представлений: использовать более широкий набор характеристик слов контекста. В частности, при построении векторных представлений можно опираться не только на лексический контекст, но и, например, на синтаксические структуры предложения. Глубокие рекуррентные нейронные сети планируется также рассмотреть для свертки длинных последовательностей слов (клауз или предложений) в векторы признаков контекста.

Заключение

В статье рассмотрены методы открытого извлечения информации из текстов на естественном языке. Это относительно новое направление, возникшее вследствие появления необходимости масштабировать системы извлечения информации из текстов на очень большие данные, которые при этом являются смесью огромного количества различных тематик. Тексты на естественном языке все еще остаются одним из основных способов хранения и передачи информации. В связи с быстрым ростом объемов текстовой информации ужесточаются требования к срокам создания систем извлечения информации в новых предметных областях. Методы открытого извлечения информации на основе машинного обучения с частичным привлечением учителя и без учителя, рассмотренные в статье, могут помочь

частично решить эти проблемы, поскольку позволяют обходиться либо без, либо с минимальными затратами ручного труда.

Представлены также основные направления наших дальнейших исследований, связанных с разработкой систем пополнения баз знаний с использованием методов открытого извлечения информации. Во второй части настоящей работы планируется подробно представить методы извлечения и группировки сущностей, а также методы выделения терминов и поверхностных связей. Будут приведены результаты экспериментальных исследований методов на коллекциях текстов на английском и русском языках.

Литература

1. Appelt D. E. The common pattern specification language // Technical report / SRI International, Artificial Intelligence Center. — 1998.
2. A framework and graphical development environment for robust NLP tools and applications / Hamish Cunningham, Diana Maynard, Kalina Bontcheva, Valentin Tablan // Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics. — 2002. — P. 168–175.
3. Большакова Е. И. Язык лексико-синтаксических шаблонов LSPL: опыт использования и пути развития // Программные системы и инструменты: тематический сборник. — 2014. — 15. — С. 15–26.
4. UIMA Ruta: Rapid development of rule-based information extraction applications / Peter Kluegl, Martin Toepfer, Philip-Daniel Beck et al. // Natural Language Engineering. — 2016. — Vol. 22, no. 1. — P. 1–40.
5. Starostin A. S., Smurov I. M., Stepanova M. E. A production system for information extraction based on complete syntactic semantic analysis // Papers from the Annual International Conference "Dialogue" (2014). — 2014. — P. 659–667.
6. Culotta A., Sorensen J. Dependency tree kernels for relation extraction // Proceedings of the 42nd Meeting of the Association for Computational Linguistics. — 2004. — P. 423–429.
7. Bunescu R., Mooney R. A shortest path dependency kernel for relation extraction // Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing. — 2005. — P. 724–731.
8. Ebrahimi J., Dou D. Chain based RNN for relation classification // Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. — 2015. — P. 1244–1249.
9. Semantic relation classification via convolutional neural networks with simple negative sampling / Kun Xu, Yansong Feng, Songfang Huang, Dongyan Zhao // Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. — 2015. — P. 536–540.
10. Nguyen T. H., Grishman R. Relation extraction: Perspective from convolutional neural networks // Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing. — 2015. — P. 39–48.
11. TextRunner: open information extraction on the web / Alexander Yates, Michael Cafarella, Michele Banko et al. // Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations. — 2007. — P. 25–26.
12. Open information extraction from the web / Michele Banko, Michael J. Cafarella, Stephen Soderland et al. // Proceedings of the 20th International Joint Conference on Artificial Intelligence. — 2007. — P. 2670–2676.
13. Marcus M. P., Marcinkiewicz M. A., Santorini B. Building a large annotated corpus of English: The Penn Treebank // Computational linguistics. — 1993. — Vol. 19, no. 2. — P. 313–330.
14. Banko M., Etzioni O. The tradeoffs between open and traditional relation extraction // Proceedings of ACL-08: HLT. — 2008. — P. 28–36.
15. StatSnowball: a statistical approach to extracting entity relationships / Jun Zhu, Zaiqing Nie, Xiaojiang Liu et al. // Proceedings of the 18th international conference on World wide web. — 2009. — P. 101–110.
16. Wu F., Weld D. S. Open information extraction using Wikipedia // Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. — 2010. — P. 118–127.
17. Unsupervised relation extraction by mining Wikipedia texts using information from the web / Yulan Yan, Naoaki Okazaki, Yutaka Matsuo et al. // Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. — 2009. — P. 1021–1029.
18. Fader A., Soderland S., Etzioni O. Identifying relations for open information extraction // Proceedings of the Conference on Empirical Methods in Natural Language Processing. — 2011. — P. 1535–1545.
19. Open information extraction: The second generation / Oren Etzioni, Anthony Fader, Janara Christensen et al. // Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence. — 2011. — P. 3–10.
20. Open language learning for information extraction / Michael Schmitz, Robert Bart, Stephen Soderland et al. // Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. — 2012. — P. 523–534.
21. Nivre J., Nilsson J. Multiword units in syntactic parsing // Proceedings of Methodologies and Evaluation of Multiword Units in Real-World Applications (MEMURA). — 2004.
22. Angeli G., Johnson Premkumar M. J., Manning C. D. Leveraging linguistic structure for open domain information extraction // Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. — 2015. — P. 344–354.
23. Overview of the TAC 2010 knowledge base population track / Heng Ji, Ralph Grishman, Hoa Trang Dang et al. // Third Text Analysis Conference (TAC 2010). — Vol. 3. — 2010. — P. 3–3.
24. Surdeanu M. Overview of the TAC 2013 knowledge base population evaluation: English slot filling and temporal slot filling // Proceedings of the TAC-KBP 2013 Workshop. — 2013.

25. Combining distant and partial supervision for relation extraction / Gabor Angeli, Julie Tibshirani, Jean Wu, Christopher D. Manning // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. — 2014. — P. 1556–1567.
26. Stanford typed dependencies manual : Rep. / Technical report, Stanford University ; Executor: Marie-Catherine De Marneffe, Christopher D Manning : 2008.
27. Nakashole N., Weikum G., Suchanek F. PATTY: A taxonomy of relational patterns with semantic types // Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. — 2012. — P. 1135–1145.
28. Distant supervision for relation extraction without labeled data / Mike Mintz, Steven Bills, Rion Snow, Dan Jurafsky // Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. — 2009. — P. 1003–1011.
29. Freebase: a collaboratively created graph database for structuring human knowledge / Kurt Bollacker, Colin Evans, Praveen Paritosh et al. // Proceedings of the 2008 ACM SIGMOD international conference on Management of data. — 2008. — P. 1247–1250.
30. Riedel S., Yao L., McCallum A. Modeling relations and their mentions without labeled text // Machine learning and knowledge discovery in databases. — 2010. — P. 148–163.
31. Knowledge-based weak supervision for information extraction of overlapping relations / Raphael Hoffmann, Congle Zhang, Xiao Ling et al. // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. — 2011. — P. 541–550.
32. Multi-instance multi-label learning for relation extraction / Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, Christopher D Manning // Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. — 2012. — P. 455–465.
33. Dietterich T. G., Lathrop R. H., Lozano-Pérez T. Solving the multiple instance problem with axis-parallel rectangles // Artificial intelligence. — 1997. — Vol. 89, no. 1. — P. 31–71.
34. Connecting language and knowledge bases with embedding models for relation extraction / Jason Weston, Antoine Bordes, Oksana Yakhnenko, Nicolas Usunier // Conference on Empirical Methods in Natural Language Processing. — 2013. — P. 1366–1371.
35. Relation extraction with matrix factorization and universal schemas / Sebastian Riedel, Limin Yao, Andrew McCallum, Benjamin M. Marlin // Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. — 2013. — P. 74–84.
36. Distant supervision for relation extraction via piecewise convolutional neural networks / Daojian Zeng, Kang Liu, Yubo Chen, Jun Zhao // Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. — 2015. — P. 1753–1762.
37. A customized attention-based long short-term memory network for distant supervised relation extraction / Dengchao He, Hongjun Zhang, Wenning Hao et al. // Neural Computation. — 2017.
38. Hochreiter S., Schmidhuber J. Long short-term memory // Neural computation. — 1997. — Vol. 9, no. 8. — P. 1735–1780.
39. Brin S. Extracting patterns and relations from the World wide web // International Workshop on The World Wide Web and Databases. — 1998. — P. 172–183.
40. Agichtein E., Gravano L. Snowball: Extracting relations from large plain-text collections // Proceedings of the fifth ACM conference on Digital libraries. — 2000. — P. 85–94.
41. Batista D. S., Martins B., Silva M. J. Semi-supervised bootstrapping of relationship extractors with distributional semantics // Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. — 2015. — P. 499–504.
42. Efficient estimation of word representations in vector space / Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean // arXiv preprint arXiv:1301.3781. — 2013.
43. Structured relation discovery using generative models / Limin Yao, Aria Haghighi, Sebastian Riedel, Andrew McCallum // Proceedings of the Conference on Empirical Methods in Natural Language Processing. — 2011. — P. 1456–1466.
44. Blei D. M., Ng A. Y., Jordan M. I. Latent dirichlet allocation // Journal of machine Learning research. — 2003. — Vol. 3. — P. 993–1022.
45. Yao L., Riedel S., McCallum A. Unsupervised relation discovery with sense disambiguation // Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. — 2012. — P. 712–720.
46. Lopez de Lacalle O., Lapata M. Unsupervised relation extraction with general domain knowledge // Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. — 2013. — P. 415–425.
47. A framework for incorporating general domain knowledge into latent dirichlet allocation using first-order logic / David Andrzejewski, Xiaojin Zhu, Mark Craven, Benjamin Recht // Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence. — 2011. — P. 1171–1177.
48. Takase S., Okazaki N., Inui K. Fast and large-scale unsupervised relation extraction. // Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation. — 2015.
49. Marcheggiani D., Titov I. Discrete-state variational autoencoders for joint discovery and factorization of relations // Transactions of the Association for Computational Linguistics. — 2016. — Vol. 4. — P. 231–244.
50. Vorontsov K., Potapenko A. Additive regularization of topic models // Machine Learning. — 2015. — Vol. 101, no. 1-3. — P. 303–323.
51. Mnih A., Kavukcuoglu K. Learning word embeddings efficiently with noise-contrastive estimation // Advances in neural information processing systems. — 2013. — P. 2265–2273.

Шелманов Артем Олегович. Кандидат технических наук, научный сотрудник ИСА ФИЦ ИУ РАН. Количество печатных работ: 30. E-mail: shelmanov@isa.ru

Исаков Вадим Алексеевич. Программист ИСА ФИЦ ИУ РАН. Количество печатных работ: 7. E-mail: visakov@isa.ru

Станкевич Максим Алексеевич. Сотрудник ИСА ФИЦ ИУ РАН. E-mail: maxastan95@gmail.com

Смирнов Иван Валентинович. Доцент, кандидат физико–математических наук, заведующий лабораторией ИСА ФИЦ ИУ РАН. Количество печатных работ: 52. E-mail: ivs@isa.ru

Open Information Extraction. Part I. The Task and the Review of the State of the Art

A.O. Shelmanov, V.A. Isakov, M.A. Stankevich, I.V. Smirnov

Systems Analysis of FRC CSC RAS, Moscow, Russia

The paper discusses the task of open information extraction from natural language texts. Open information extraction – is rather new approach to solving tasks of information extraction that do not specify structure and semantics of the information to be extracted. This approach is domain independent and does not require big annotated corpora. We present the formulation of the problem and review the state of the art related to extraction of entities and semantic relations from texts including methods of information extraction based on semi-supervised and unsupervised learning. We present the future directions of research of methods for relation extraction based on unsupervised learning.

Keywords: open information extraction, semantic relations, term extraction, unsupervised learning, semi-supervised learning.

DOI 10.14357/20718594180204

References

1. Appelt D. E. The common pattern specification language // Technical report / SRI International, Artificial Intelligence Center. — 1998.
2. A framework and graphical development environment for robust NLP tools and applications / Hamish Cunningham, Diana Maynard, Kalina Bontcheva, Valentin Tablan // Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics. — 2002. — P. 168–175.
3. Bolshakova E.I. Yazik leksiko-syntaksicheskih shablonov LSPL: opit ispolzovania i puti razvitija [The language of lexical-syntactic templates LSPL: the experience and development directions.] // Programnye systemi i instrumenty: tematicheskij sbornik [Program systems and instruments: the topical collection] — 2014 — P. 15–26.
4. UIMA Ruta: Rapid development of rule-based information extraction applications / Peter Kluegl, Martin Toepfer, Philip-Daniel Beck et al. // Natural Language Engineering. — 2016. — Vol. 22, no. 1. — P. 1–40.
5. Starostin A. S., Smurov I. M., Stepanova M. E. A production system for information extraction based on complete syntactic semantic analysis // Papers from the Annual International Conference "Dialogue" (2014). — 2014. — P. 659–667.
6. Culotta A., Sorensen J. Dependency tree kernels for relation extraction // Proceedings of the 42nd Meeting of the Association for Computational Linguistics. — 2004. — P. 423–429.
7. Bunescu R., Mooney R. A shortest path dependency kernel for relation extraction // Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing. — 2005. — P. 724–731.
8. Ebrahimi J., Dou D. Chain based RNN for relation classification // Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. — 2015. — P. 1244–1249.
9. Semantic relation classification via convolutional neural networks with simple negative sampling / Kun Xu, Yansong Feng, Songfang Huang, Dongyan Zhao // Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. — 2015. — P. 536–540.
10. Nguyen T. H., Grishman R. Relation extraction: Perspective from convolutional neural networks // Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing. — 2015. — P. 39–48.
11. TextRunner: open information extraction on the web / Alexander Yates, Michael Cafarella, Michele Banko et al. // Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations. — 2007. — P. 25–26.
12. Open information extraction from the web / Michele Banko, Michael J. Cafarella, Stephen Soderland et al. // Proceedings of the 20th International Joint Conference on Artifical Intelligence. — 2007. — P. 2670–2676.
13. Marcus M. P., Marcinkiewicz M. A., Santorini B. Building a large annotated corpus of English: The Penn Treebank // Computational linguistics. — 1993. — Vol. 19, no. 2. — P. 313–330.
14. Banko M., Etzioni O. The tradeoffs between open and traditional relation extraction // Proceedings of ACL-08: HLT. — 2008. — P. 28–36.
15. StatSnowball: a statistical approach to extracting entity relationships / Jun Zhu, Zaiqing Nie, Xiaojiang Liu et al. // Proceedings of the 18th international conference on World wide web. — 2009. — P. 101–110.

16. Wu F., Weld D. S. Open information extraction using Wikipedia // Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. — 2010. — P. 118–127.
17. Unsupervised relation extraction by mining Wikipedia texts using information from the web / Yulan Yan, Naoaki Okazaki, Yutaka Matsuo et al. // Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. — 2009. — P. 1021–1029.
18. Fader A., Soderland S., Etzioni O. Identifying relations for open information extraction // Proceedings of the Conference on Empirical Methods in Natural Language Processing. — 2011. — P. 1535–1545.
19. Open information extraction: The second generation / Oren Etzioni, Anthony Fader, Janara Christensen et al. // Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence. — 2011. — P. 3–10.
20. Open language learning for information extraction / Michael Schmitz, Robert Bart, Stephen Soderland et al. // Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. — 2012. — P. 523–534.
21. Nivre J., Nilsson J. Multiword units in syntactic parsing // Proceedings of Methodologies and Evaluation of Multiword Units in Real-World Applications (MEMURA). — 2004.
22. Angeli G., Johnson Premkumar M. J., Manning C. D. Leveraging linguistic structure for open domain information extraction // Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. — 2015. — P. 344–354.
23. Overview of the TAC 2010 knowledge base population track / Heng Ji, Ralph Grishman, Hoa Trang Dang et al. // Third Text Analysis Conference (TAC 2010). — Vol. 3. — 2010. — P. 3–3.
24. Surdeanu M. Overview of the TAC 2013 knowledge base population evaluation: English slot filling and temporal slot filling // Proceedings of the TAC-KBP 2013 Workshop. — 2013.
25. Combining distant and partial supervision for relation extraction / Gabor Angeli, Julie Tibshirani, Jean Wu, Christopher D. Manning // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. — 2014. — P. 1556–1567.
26. Stanford typed dependencies manual : Rep. / Technical report, Stanford University ; Executor: Marie-Catherine De Marneffe, Christopher D Manning : 2008.
27. Nakashole N., Weikum G., Suchanek F. PATTY: A taxonomy of relational patterns with semantic types // Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. — 2012. — P. 1135–1145.
28. Distant supervision for relation extraction without labeled data / Mike Mintz, Steven Bills, Rion Snow, Dan Jurafsky // Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. — 2009. — P. 1003–1011.
29. Freebase: a collaboratively created graph database for structuring human knowledge / Kurt Bollacker, Colin Evans, Praveen Paritosh et al. // Proceedings of the 2008 ACM SIGMOD international conference on Management of data. — 2008. — P. 1247–1250.
30. Riedel S., Yao L., McCallum A. Modeling relations and their mentions without labeled text // Machine learning and knowledge discovery in databases. — 2010. — P. 148–163.
31. Knowledge-based weak supervision for information extraction of overlapping relations / Raphael Hoffmann, Congle Zhang, Xiao Ling et al. // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. — 2011. — P. 541–550.
32. Multi-instance multi-label learning for relation extraction / Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, Christopher D Manning // Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. — 2012. — P. 455–465.
33. Dietterich T. G., Lathrop R. H., Lozano-Pérez T. Solving the multiple instance problem with axis-parallel rectangles // Artificial intelligence. — 1997. — Vol. 89, no. 1. — P. 31–71.
34. Connecting language and knowledge bases with embedding models for relation extraction / Jason Weston, Antoine Bordes, Oksana Yakhnenko, Nicolas Usunier // Conference on Empirical Methods in Natural Language Processing. — 2013. — P. 1366–1371.
35. Relation extraction with matrix factorization and universal schemas / Sebastian Riedel, Limin Yao, Andrew McCallum, Benjamin M. Marlin // Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. — 2013. — P. 74–84.
36. Distant supervision for relation extraction via piecewise convolutional neural networks / Daojian Zeng, Kang Liu, Yubo Chen, Jun Zhao // Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. — 2015. — P. 1753–1762.
37. A customized attention-based long short-term memory network for distant supervised relation extraction / Dengchao He, Hongjun Zhang, Wenning Hao et al. // Neural Computation. — 2017.
38. Hochreiter S., Schmidhuber J. Long short-term memory // Neural computation. — 1997. — Vol. 9, no. 8. — P. 1735–1780.
39. Brin S. Extracting patterns and relations from the World wide web // International Workshop on The World Wide Web and Databases. — 1998. — P. 172–183.
40. Agichtein E., Gravano L. Snowball: Extracting relations from large plain-text collections // Proceedings of the fifth ACM conference on Digital libraries. — 2000. — P. 85–94.
41. Batista D. S., Martins B., Silva M. J. Semi-supervised bootstrapping of relationship extractors with distributional semantics // Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. — 2015. — P. 499–504.

42. Efficient estimation of word representations in vector space / Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean // arXiv preprint arXiv:1301.3781. — 2013.
43. Structured relation discovery using generative models / Limin Yao, Aria Haghighi, Sebastian Riedel, Andrew McCallum // Proceedings of the Conference on Empirical Methods in Natural Language Processing. — 2011. — P. 1456–1466.
44. Blei D. M., Ng A. Y., Jordan M. I. Latent dirichlet allocation // Journal of machine Learning research. — 2003. — Vol. 3. — P. 993–1022.
45. Yao L., Riedel S., McCallum A. Unsupervised relation discovery with sense disambiguation // Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. — 2012. — P. 712–720.
46. Lopez de Lacalle O., Lapata M. Unsupervised relation extraction with general domain knowledge // Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. — 2013. — P. 415–425.
47. A framework for incorporating general domain knowledge into latent dirichlet allocation using first-order logic / David Andrzejewski, Xiaojin Zhu, Mark Craven, Benjamin Recht // Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence. — 2011. — P. 1171–1177.
48. Takase S., Okazaki N., Inui K. Fast and large-scale unsupervised relation extraction. // Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation. — 2015.
49. Marcheggiani D., Titov I. Discrete-state variational autoencoders for joint discovery and factorization of relations // Transactions of the Association for Computational Linguistics. — 2016. — Vol. 4. — P. 231–244.
50. Vorontsov K., Potapenko A. Additive regularization of topic models // Machine Learning. — 2015. — Vol. 101, no. 1-3. — P. 303–323.
51. Mnih A., Kavukcuoglu K. Learning word embeddings efficiently with noise-contrastive estimation // Advances in neural information processing systems. — 2013. — P. 2265–2273.

Shelmanov Artem Olegovich. PhD, fellow of Institute for Systems Analysis of FRC CSC RAS. Number of published works: 20. E-mail: shelmanov@isa.ru

Isakov Vadim Alexeevich. Fellow of Institute for Systems Analysis of FRC CSC RAS. Number of published works: 7. E-mail: visakov@isa.ru

Stankevich Maxim Alexeevich. Fellow of Institute for Systems Analysis of FRC CSC RAS. E-mail: maxastan95@gmail.com

Smirnov Ivan Valentinovitch. PhD, associate professor, head of the laboratory in the Institute for Systems Analysis of the Federal Research Center “Computer Science and Control” (117312, prospekt 60-letiya Oktyabrya 9, Moscow). Authored 66 scientific papers. E-mail: ivs@isa.ru