О проблемах идентификации именованных объектов и извлечения связей при анализе научных и технических русскоязычных текстов

А. П. Преображенский, email: app@vivt.ru [1]
Д. В. Меняйлов, email: dmitriy.menyaylov111@yandex.ru [1]

[1] АНОО ВО – Воронежский институт высоких технологий, Воронеж, Российская Федерация

Аннотация. *Работа посвящена рассмотрению методов извлечения информации (распознавание объектов и классификация отношений) из научных текстов по информационным технологиям. Научные публикации предоставляют ценную информацию для передовых научных достижений, но эффективная обработка растущих объемов данных требует много времени. В данной работе рассматривается несколько модификаций методов для русского языка. Представлены результаты экспериментов по сравнению метода извлечения ключевых слов, словарного метода и некоторых методов, основанных на нейронных сетях. Рассмотрен корпус научных текстов на русском языке RuSERRC, который состоит из 1600 неразмеченных документов и 80 размеченных объектов и семантических отношений, рассмотрено 6 типов отношений. Набор данных и модели полезны для исследовательских целей и разработки систем извлечения информации.*

Ключевые слова: *распознавание объектов, классификация отношений, модели нейронных сетей, построение набора данных, извлечение информации.*

Введение

Количество информации в интернете, в том числе текстовой, растет чрезвычайно быстро. По данным журнала «Nature», мировое научное сообщество ежегодно публикует более миллиона статей только на медико-биологическую тематику [1]. Научные публикации содержат ценную информацию о ведущих научных достижениях, но эффективная обработка такого огромного объема данных требует много времени.

Одной из задач извлечения информации является распознавание объектов. Целью этой задачи является обнаружение и классификация объектов по предопределенным категориям, таким как имена, организации, местоположения, временные выражения, деньги и т.д. Эта задача часто решается вместе с задачей извлечения связей, суть которой

заключается в том, чтобы найти пары объектов, которые могут быть связаны семантическим отношением. Если набор отношений предопределен, то речь идет о задаче классификации отношений – сопоставлении каждой пары объектов с определенным семантическим отношением. Часто делается следующее допущение: объекты должны находиться в одном предложении.

В настоящее время методы, основанные на глубоком обучении, достаточно хорошо решают эти задачи. В таких методах используются языковые модели, построенные на больших неразмеченных корпусах. Чтобы добиться хорошего качества данных из определенных доменов, необходимо точно настроить модели на специальных корпусах. Набор данных RuSERRC (Russian Scientific Entity Recognition and Relation Extraction) содержит тексты в области информационных технологий. Проведено несколько экспериментов для исследования и сравнения различных методов на этом корпусе.

Распознавание именованных объектов и классификация отношений являются важными шагами для извлечения информации из текстов. Наиболее известные наборы данных на английском языке для этих задач – TACRED [2], SemEval-2010 Task 8 [3], CONLL04 [4]. Для русского языка наиболее популярен набор данных FactRuEval-16 [5], состоящий из новостных текстов. В настоящее время наиболее перспективными считаются методы, основанные на архитектуре Transformer, которые обычно проходят полуконтролируемое обучение, включающее неконтролируемую предварительную подготовку, за которой следует контролируемая точная настройка для рассматриваемой задачи. Пример тонкой настройки BERT представлен в [6]. Для достижения представления отношений путем точной настройки BERT с крупномасштабным «сопоставлением пробелов» используются тексты, связанные с объектами перед обучением. Этот метод хорошо работает с набором данных SemEval-2010 Task 8 (оценка F1 88,6%) и превосходит предыдущие методы на TACRED (оценка F1 70,4%).

Чтобы добиться хорошего качества данных в конкретной области, необходимо точно настроить модель на соответствующем наборе данных. Предварительное обучение обычно выполняется на гораздо большем наборе данных, чем точная настройка для конкретной области знаний, из-за ограниченной доступности помеченных обучающих данных. В качестве полезного полигона доступен корпус российских документов стратегического планирования RuREBus [7]. Модель на основе BERT получает наилучший показатель F1: 0,563 для NER и 0,443 для извлечения отношений. Еще один набор данных на русском языке – RURED [8]. Он состоит из текстов экономических новостей.

Многоязычная модель BERT дает 0,86 для NER, а модель SpanBERT [9] дает наилучшие результаты 0,79 для извлечения отношений F1-показателя в этом корпусе.

Также представляет интерес решение задач распознавания объектов и классификации смысловых отношений в научных текстах. Хотя есть инструменты (например, natasha, spacy-ru) для извлечения традиционных типов объектов из общих текстов предметной области (лица, местоположения, организации и т.д.), распознавание объектов и отношений в научно-технических текстах на русском языке еще нуждается в исследовании. Такие сборники для английского языка [10], [11], [12] имеются и активно используются научным сообществом, однако в настоящее время сложно найти аналогичный набор данных на русском языке.

1. Набор данных

Данный корпус состоит из рефератов научных работ из области информационных технологий. Корпус состоит из 1600 неразмеченных текстов и 80 текстов, аннотированных вручную терминами и семантическими отношениями между ними.

В качестве объектов рассматриваются существительные или группы существительных, которые являются терминами в этой конкретной области. Под термином понимается словосочетание, являющееся названием определенного понятия области науки, техники, искусства и т.п. Объектами считаются термины, состоящие из одной лексемы или аббревиатуры («база данных» (БД), «программное обеспечение» (ПО), «интерфейс»), названия языков программирования («Python», «Java», «C++») и библиотек («Pytorch», «Keras», «pymorphy2»), дефисные понятия, содержащие латинские символы («n-грамма», «веб-сервис»). Термины с орфографическими ошибками или опечатками также были помечены как объекты. Объекты перечислены с ”,”, ”;” или соединенные союзом «и», по возможности, обозначались раздельно. Представлены некоторые абстрактные понятия (например, «метод», «явление», «свойство» и т.д.) как термины, чтобы связать объекты через эти понятия с базой знаний в будущем.

Основная трудность заключается в процессе разграничения терминов и нетерминов. Часто бывает трудно понять без контекста, является ли фраза термином или нет. Рассматривается многословный термин как цепочка токенов максимальной длины, из которой получается более общий термин, если убрать токены. Например, составной термин «модель структурной организации единого информационного пространства» считается корпусным термином, поскольку отдельное слово «модель» не отражает точный смысл в

данном контексте. Но в случае, когда слово «модель» встречается в тексте без дополнительных слов, то оно рассматривается как термин. Обычно такими объектами являются названия программных продуктов, методов, алгоритмов, задач, подходов («операционная система Android», «метод k ближайших соседей», «метод опорных векторов»).

В научных текстах на русском языке часто встречаются отглагольные существительные, обозначающие процессы. С точки зрения семантики процесс приводит к изменениям, влияет на результат. Вот почему такие существительные желательно включать в объект, это влияет на извлечение отношений. Примеры: «обработка изображений», «тестирование системы», «анализ текста» и т.д.

Объекты помечены в формате BIO: каждому токену присваивается тег B-TERM, если он является начальным тегом для объекта, I-TERM, если он находится внутри термина, или O, если он находится вне какого-либо объекта. Объекты не рекурсивны и не пересекаются. В результате проаннотировано 82 текста, содержащих 11159 токенов и 2029 терминов. Средняя длина термина составляет 2,45 токена. Самый длинный терм содержит 13 токенов.

Список отношений выбран в результате анализа работ [3], [12], [13] на основании следующих критериев. Во-первых, отношение должно быть однозначным (например, не рассматривается семантическое отношение <Объект – Назначение>, поскольку оно также имеет косвенное значение). Во-вторых, отношение должно связывать научные термины (например, в отношении <Коммуникация – Тема> (акт общения по теме) актанты не являются научными терминами). Таким образом, отобрано шесть семантических отношений.

1. Отношение причинности, например, [взаимодействие высокоэнергетических пучков : деформация].
2. Отношение сравнения, например, [реляционные базы данных: объектно-ориентированные базы данных].
3. Отношение таксономии, например, [Python : язык программирования].
4. Отношение мерономии, например, [модуль : система].
5. Отношения синонимии, например, [GPU : графический процессор].
6. Отношение использования, например, [метод статистической обработки : анализ текста].

Отношения между объектами аннотированы в рамках одного предложения. В результате размечено 632 связи между терминами: причинность – 27, сравнение – 23, таксономия – 92, мерономия – 79, синонимия – 24, использование – 387.

2. Методы

Могут проводиться базовые эксперименты с использованием как архитектуры преобразователей, так и традиционных методов решения задачи извлечения терминов, среди которых реализован метод на основе словаря, комбинированный метод и статистический метод.

Метод распознавания объектов на основе словаря предлагает использовать предопределенный набор (словарь) терминов. Может быть сформирован полуавтоматически двумя способами:

1. Извлечь 2-, 3- и 4-граммы из научных статей и отсортировать по значению TF-IDF, затем вручную отфильтровать фразы, которые потенциально могут быть терминами.

2. Извлечь все названия статей, которые входят в подграф категории «Наука», а затем вручную отобрать фразы, которые потенциально могут быть терминами.

Таким образом, возможно собрать более 17260 терминов.

Комбинированный метод распознавания объектов. Основная трудность при проведении экспериментов с использованием различных алгоритмов машинного обучения – отсутствие размеченных данных. Для решения этой проблемы автоматически аннотировано 1120 научных статей (которые очищены от формул, таблиц, рисунков и т.д.) терминами из словаря, описанного выше. Таким образом, получен аннотированный набор данных, который состоит из 2 миллионов токенов и 179 тысяч терминов. Входная последовательность закодирована на уровне символов. Модель содержит один двунаправленный слой LSTM и слой CRF для формирования выходной последовательности тегов. Проанализировано, сколько терминов смогла извлечь модель, которых нет в словаре терминов, и обнаружено, что около 26,7% всех уникальных терминов – это фразы, которые модель ранее не видела.

Алгоритм RAKE (быстрое автоматическое извлечение ключевых слов) хорошо применим к динамическим корпусам и совершенно новым предметным областям, он не зависит от языка документа и его особенностей [14]. Первым шагом алгоритма является применение стоп-слов и списка разделителей для выделения терминов, состоящих из нескольких слов. После этого вычисляется некоторая статистическая информация. Для каждого слова оценивается частота, с которой оно встречается. Второй параметр – количество отношений между текущим словом и другими словами в тексте. На основе этих значений оценивается вес каждой ключевой фразы. Все фразы отсортированы по их весу, поэтому наиболее вероятные фразы получают максимальное значение.

При использовании реализации RAKE, которая поддерживает русский язык и автоматическое извлечение стоп-слов из текста, замечено, что алгоритм часто добавляет ключевые фразы, содержащие глагольные формы. Поскольку рассматриваются только существительные или группы существительных как объекты, необходимо предварительно обработать тексты и удалить все глаголы и их формы перед применением RAKE. Глагольные формы возможно извлечь с помощью Mystem.

Распознавание объектов и классификация отношений с использованием моделей на основе BERT. RuBERT [15] – модель BERT, предварительно обученная на русскоязычных текстах. Используя веса этой модели для инициализации, предварительно обучена BERT на коллекции научных текстов.

1. BertLinearER. Самый простой способ разметить последовательности с помощью BERT – использовать линейный слой поверх векторных представлений токенов, сгенерированных BERT. Функция потерь в этом случае основана на перекрестной энтропии (CrossEntropyLoss).

2. BertLstmLinearER и BertCnnLinearER. Разумное решение – усложнить классификатор, добавив слои, которые учатся находить зависимости в последовательности. Сверточные слои больше подходят для выделения краткосрочных зависимостей (для токенов на небольшом расстоянии друг от друга), а слои LSTM (Long Short-term memory) хороши для выделения долговременных зависимостей (для токенов на большом расстоянии друг от друга). Сверточные слои объединяются в блоки (блоки CNN), структура которых основана на проекте Кераса-Бертнера. Функция потерь также основана на перекрестной энтропии.

3. BertLstmCrfER и BertCnnCrfER. Следующим шагом для улучшения классификатора является добавление слоя CRF. Другая функция потерь, логарифм правдоподобия, используется вместе с CRF.

4. BertRC. Базовая линия для классификации отношений основана на архитектуре R-BERT [16]. Входные данные для классификации отношений отличаются от распознавания именованных объектов – помимо последовательности токенов на вход предоставляются битовые маски. Эти маски показывают, что токены принадлежат объектам. Слой softmax расположен на выходе модели. Потери MSE используются в качестве функции потерь.

3. Результаты

Для оценки качества алгоритмов распознавания объектов рассмотрено сопоставление как с точным, так и с нечетким токеном.

Метрики рассчитаны для всего набора данных RuSERRC, поскольку он не использовался во время алгоритмов подбора.

Низкие показатели словарного метода связаны с тем, что словарь состоит только из полных названий терминов, тогда как в реальных текстах последовательность терминологических токенов может быть нарушена другими токенами, содержать синонимы, аббревиатуры или даже быть неполным. Низкие метрики комбинированного метода обусловлены той же причиной, так как тексты для обучения автоматически размечались с помощью этого словаря, поэтому в обучающей выборке не было примеров, в которых термин каким-либо образом был изменен.

Метрики алгоритма RAKE немного лучше: алгоритм извлекает из текста больше терминов. Оптимизация с удалением глагольных форм снижает мощность извлекаемого набора терминов, тем самым повышая точность алгоритма.

Модели на основе BERT обучались в течение 30 эпох (полные проходы через обучающий набор данных). Для тестирования было выделено 10% набора данных (эти примеры не участвовали в обучении модели). Произведено сравнение F-оценки по корпусу RuSERRC для рассмотренных моделей распознавания объектов на основе BERT. Для модели на основе BERT для классификации отношений установлено 0,840 F-показателя.

Модель лучше всего распознает отношение USAGE, что неудивительно, поскольку оно имеет наибольшую частоту в обучающей выборке. Результаты можно улучшить, расширив корпус текстов для предобучения BERT и, увеличив количество эпох для предобучения и обучения.

В связи с тем, что сложно найти аналогичный корпус и модели для русского языка для решения поставленных задач, опубликованные результаты других исследователей на других аналогичных наборах данных имеют целью показать, каких результатов модель может достичь в принципе. Для задачи извлечения отношения F-оценка равна 0,563 на RuREBus [7], 0,705 – на SpERT [17], 0,852 – на RURED [8]. Для задачи классификации отношений F-оценка составляет 0,443 на RuREBus, 0,510 – на SpERT, 0,764 – на RURED. Конечно, при анализе результатов важно учитывать, что значения метрик на наборе данных во многом зависят от его свойств: размера, полноты, качества текстов, обучающих примеров и других характеристик.

Заключение

В данной работе представлены методы автоматического выделения объектов и классификации семантических отношений для построения

моделей, работающих с русским языком. Эти модели особенно актуальны, так как большинство существующих исследований ориентированы на данные на английском и китайском языках, а найти качественные модели для русского языка в открытом доступе достаточно сложно.

В дальнейшем необходимо провести серию экспериментов с моделью ERNIE [18], использующей дополнительную структурированную информацию о языке. Предварительно обученные модели ERNIE существуют для двух языков – английского и китайского. Согласно выводам, сделанным в [18], есть основания полагать, что такой подход улучшит результаты для русского языка.

Корпус RuSERRC находится в открытом доступе и может быть полезен для исследовательских целей и разработки систем извлечения информации. Набор данных научных текстов RuSERRC содержит разметку объектов и семантических отношений между ними.

Список литературы

1. Ландхус, Э. Научная литература: информационная перегрузка / Э. Ландхус // Nature. – 2016. – № 535. – С. 457-458.

2. Чжан, Ю., Чжун, В., Чен, Д., Анджели, Г., Мэннинг, К.Д. Внимание с учетом положения и контролируемые данные улучшают заполнение слотов / Ю. Чжан, В. Чжун, Д. Чен, Г. Анджели, К.Д. Мэннинг // Материалы конференции 2017 года по эмпирическим методам обработки естественного языка. Ассоциация компьютерной лингвистики. – Копенгаген, Дания, 2017. – С. 35-45.

3. Хендрикс, И., Ким, С.Н., Козарева, З., Наков, П., Падо, Д.О., Пеннаккиотти, М., Романо, Л., Шпакович, С. SemEval-2010, задача 8: многосторонняя классификация семантических отношений между парами имен / И. Хендрикс, С.Н. Ким, З. Козарева, П. Наков, Д.О. Падо, М. Пеннаккиотти, Л. Романо, С. Шпакович // Материалы 5-го Международного семинара по семантической оценке. Ассоциация компьютерной лингвистики. – Уппсала, Швеция, 2010. – С. 33-38.

4. Каррерас, К., Маркес, Л. Введение в общую задачу CoNLL-2005: маркировка семантических ролей / К. Каррерас, Л. Маркес // Материалы девятой конференции по компьютерному изучению естественного языка (CoNLL-2005). Ассоциация компьютерной лингвистики. – Анн-Арбор, Мичиган, 2005. – С. 152-164.

5. Старостин, А., Бочаров, В., Алексеева, С., Бодрова, А., Чучунков, А., Джумаев, С., Ефименко, И., Грановский, Д., Хорошевский, В., Крылова, И., Николаева, М., Смуров, И., Толдова, С. Factrueval 2016: Оценка систем распознавания именованных сущностей и извлечения фактов для русского языка / А. Старостин, В. Бочаров, С. Алексеева, А.

Бодрова, А. Чучунков, С. Джумаев, И. Ефименко, Д. Грановский, В. Хорошевский, И. Крылова, М. Николаева, И. Смуров, С. Толдова // FactRuEval-2016. – 2016. – С. 688-705.

6. Соарес, Л.Б., Фитцджеральд, Н., Линг, Дж., Квятковски, Т. Сопоставление пробелов: дистрибутивное сходство для изучения отношений / Л.Б. Соарес, Н. Фитцджеральд, Дж. Линг, Т. Квятковски // ArXiv: 1906.03158. – 2019.

7. Артемова, Е., Батура, Т.В., Голенковская, А., Иванин, В., Иванов, В., Саркисян, В., Смуров, И., Тутубалина, Е. Документы по стратегическому планированию горных работ / Е. Артемова, Т.В. Батура, А. Голенковская, В. Иванин, В. Иванов, В. Саркисян, И. Смуров, Е. Тутубалина // ArXiv, vol. абс/2007.00257. – 2020.

8. Давлетов, А., Гордеев, Д., Алексей, Р., Арефьев, Н. Ренерсанс: извлечение отношений и распознавание именованных сущностей как аннотации последовательностей / А. Давлетов, Д. Гордеев, Р. Алексей, Н. Арефьев // Вычислительная лингвистика и интеллектуальные технологии: материалы международной конференции «Диалог». – Москва, 2020.

9. Джоши, М., Чен, Д., Лю, Ю., Велд, Д.С., Зеттлемойер, Л., Леви, О. Спанберт: улучшение предварительного обучения путем представления и прогнозирования интервалов / М. Джоши, Д. Чен, Ю. Лю, Д.С. Велд, Л. Зеттлемойер, О. Леви // Труды Ассоциации вычислительной лингвистики. – 2020. – С. 64-77.

10. Д'Суза, Дж., Хоппе, А., Брак, А., Джараде, М.Ю., Ауэр, С., Эверт, Р. Набор данных Stem-ecr: обоснование ссылок на научные объекты в основном научном содержании на авторитетные энциклопедические и лексикографические источники / Дж. Д'Суза, А. Хоппе, А. Брак, М.Ю. Джараде, С. Ауэр, Р. Эверт // ArXiv, т.1, с. абс/2003.01006. – 2020.

11. Смит, Х., Чжан, З., Калнан, Дж., Янсен, П. Scienceexamcer: высокоплотный мелкозернистый корпус в научной области для распознавания общих сущностей / Х. Смит, З. Чжан, Дж. Калнан, П. Янсен // ArXiv: 1911.10436. – 2019.

12. Габор, К., Бускальди, Д., Шуманн, А.-К., КасемиЗаде, Б., Заргаюна, Х., Шарнуа, Т. SemEval-2018, задача 7: извлечение и классификация семантических отношений в научных статьях / К. Габор, Д. Бускальди, А.-К. Шуман, Б. КасемиЗаде, Х. Заргаюна, Т. Шарнуа // Материалы 12-го Международного семинара по семантической оценке. Ассоциация компьютерной лингвистики. – Новый Орлеан, Луизиана, 2018. – С. 679-688.

13. Адитья, С., Синха, А. Раскрытие отношений для представления маркетинговых знаний / С. Адитья, А. Синха // ArXiv: 1912.08374. – 2019.

14. Роуз, С., Энгель, Д., Крамер, Н., Коули, В. Автоматическое извлечение ключевых слов из отдельных документов / С. Роуз, Д. Энгель, Н. Крамер, В. Коули // Интеллектуальный анализ текста: приложения и теория. – 2010. – № 1. – С. 1-20.

15. Куратов, Ю., Архипов, М. Адаптация глубоких двунаправленных многоязычных трансформеров для русского языка / Ю. Куратов, М. Архипов // Вычислительная лингвистика и интеллектуальные технологии: материалы международной конференции «Диалог». – Москва, 2019. – С. 333-339.

16. Ву, С., Хе, Ю. Обогащение предварительно обученной языковой модели информацией об объектах для классификации отношений / С. Ву, Ю. Хе // Материалы 28-й Международной конференции ACM по управлению информацией и знаниями. – 2019. – С. 2361-2364.

17. Эбертс, М., Ульгес, А. Извлечение объединенных объектов и отношений на основе Span с предварительным обучением трансформатора / М. Эбертс, А. Ульгес // ArXiv: 1909.07755. – 2019.

18. Чжан, З., Хань, С., Лю, З., Цзян, С., Сунь, М., Лю, К. Эрни: расширенное языковое представление с информативными объектами / З. Чжан, С. Хань, З. Лю, С. Цзян, М. Сунь, К. Лю // ArXiv: 1905.07129. – 2019.