

АВТОМАТИЧЕСКИЕ МЕТОДЫ ИЗВЛЕЧЕНИЯ ТАКСОНОМИЧЕСКИХ ОТНОШЕНИЙ ИЗ ТЕКСТОВ

Н.В. Лукашевич (N.V. Loukachevitch)^{a,*}

^a МГУ имени М.В.Ломоносова, 119991, Российская Федерация, Москва,

Ленинские горы, д. 1

* e-mail: louk_nat@mail.ru

Аннотация. В статье представлен обзор подходов к автоматическому извлечению таксономических отношений (IS_A отношений) из текстов, включая классические методы на основе лексико-семантических шаблонов, векторных представлений слов и их современное развитие. Также описаны новые подходы, основанные на языковых моделях типа BERT и использующие новые варианты работы с лексико-семантическими шаблонами, в которых модель должна заполнить нужную позицию в шаблоне. В статье также представлены различные способы тестирования качества извлечения таксономических отношений, приведены результаты этих тестирований. По достигаемым в настоящее время результатам можно сделать вывод, что, несмотря на прогресс в развитии методов извлечения таксономических отношений, качество извлечения является недостаточным для полностью автоматического построения или пополнения таксономий.

Ключевые слова: таксономии, отношения класс-подкласс, гипоним-гипероним, извлечение знаний из текстов

Введение

Тексты на естественном языке содержат большие объемы знаний о мире, о различных предметных областях, использовании разных слов и выражений в языке, что дает

возможность автоматического или автоматизированного построения онтологий из текстов [1, 16].

Таксономические отношения являются наиболее известным типом отношений, востребованным при описании многих предметных областей. В разных сферах знаний эти отношения могут иметь разные названия такие, как класс-подкласс (разработка онтологий), родовидовые отношения (информационно-поисковые тезаурусы), гипоним-гипероним (лексическая семантика, тезаурус WordNet [26]), is_a - отношения (искусственный интеллект).

Таксономические отношения образуют структурное ядро описания предметной области. Поэтому важной задачей является автоматизация построения таксономий, а также возможность автоматического приписывания этого отношения новым сущностям, появляющимся в текстах предметной области, так называемое пополнение таксономий (онтологий) на текстах предметной области.

В данном обзоре будут рассмотрены подходы к извлечению таксономических отношений из текстов, существующие наборы данных и подходы к оценке качества извлечения отношений, а также достигаемые на текущий момент результаты.

1. Методы, применяемые для извлечения таксономических отношений из текстов

Автоматическое построение таксономий по текстам в течение уже почти 30 лет привлекает пристальное внимание исследователей [18]. При построении семантических отношений по текстовым корпусам выделяют два основных подхода: лингвистический подход на основе лексико-семантических шаблонов, и подход, основанный на дистрибутивной семантике. В настоящее время также большое развитие получили комбинированные подходы. Кроме того, в связи с появлением предобученных на больших корпусах нейросетевых моделей типа BERT [13] развиваются подходы на основе

автоматического заполнения шаблонов нужного типа. Далее в тексте мы будем использовать для обозначения слова с широким значением термин гипероним, а для нижестоящего слов с более узким значением – термин гипоним.

1.1 Методы на основе лингвистических подходов

Лингвистический подход представляет собой извлечение отношений между словами на основе их совместных вхождений в некоторый набор шаблонов. Методы, использующие шаблоны, основаны на том, что если обнаруживается некоторый заданный контекст между сущностями X и Y , то, значит, между этими сущностями существует заданное отношение. Для таксономического отношения таким контекстом может быть контекст « X – это вид Y ». Такой подход обладает высокой точностью и простотой, может быть применен к различным типам отношений. Работы по извлечению таксономических отношений по шаблонам начались с работы [18], в которой было предложено несколько наиболее частотных шаблонов.

Применение шаблонов для извлечения отношений обычно имеет высокую точность, но также связано с несколькими проблемами, снижающими полноту извлечения отношений.

Во-первых, в используемой текстовой коллекции может не встретиться ни одного явного упоминания контекстов нужного типа, несмотря на то, что X и Y находятся между собой в нужном отношении. Из-за этого подход обладает очень низкой полнотой. Во-вторых, для качественного извлечения отношений нужно задать большое количество шаблонов.

Для увеличения полноты лексико-семантических шаблонов было предложено несколько методов. В ряде работ предлагалось использовать расширенный набор лексико-семантических шаблонов. Например, авторы статьи [37] используют набор из 59 шаблонов, собранных из разных работ.

В работе [31] для улучшения качества предсказания для низкочастотных слов, создается матрица взаимной встречаемости слов в заданных лексико-семантических шаблонах,

оцененной с помощью формулы взаимной информации. К получившейся матрице затем применяется SVD-разложение и снижение размерности, так называемый метод латентного семантического анализа [12]. В результате для слов появившихся в схожих шаблонах с похожим множеством слов предсказываются одинаковые гиперонимы, что ведет к улучшению качества предсказания гиперонимов для низкочастотных слов или слов, редко упоминавшихся в используемых шаблонах.

Еще одним подходом к увеличению полноты покрытия шаблонами является так называемый бутстраппинг (частичное обучение с учителем), то есть несколько шаблонов извлечения заданного отношения задаются вручную, а извлекаемые на основе этих шаблонов примеры отношения используются для сбора новых видов шаблонов [41]. Однако при нарастании числа шаблонов точность извлечения отношений обычно снижается.

При создании шаблонов можно заметить, что некоторые из них имеют схожий вид. В частности, система Patty [27] увеличивает полноту извлечения отношений шаблонами за счет обобщения шаблонов. Например, шаблоны «X inc. is a Y», «X group is a Y», «X organization is a Y» можно обобщить до «X * is a Y», что однако может привести к некоторому снижению точности извлечения отношения.

Наконец, в лексико-семантические шаблоны могут вклиниваться лишние слова, поэтому в ряде работ предлагается использовать для извлечения шаблонов синтаксические структуры предложений, а сам шаблон записывать на языке синтаксических отношений [2, 39], которые могут соединять и удаленно расположенные слова синтаксического шаблона.

Для русского языка лексико-семантические шаблоны для извлечения родовидовых отношений, отношений гипоним-гипероним рассматриваются в работах [8, 20]. Авторы работы [33] применяли лексико-семантические шаблоны для извлечения гиперонимов слов из их толкований в толковых словарях.

1.2. Методы на основе векторных представлений слов

Второе основное направление извлечения таксономических отношений основано на методах дистрибутивной семантики и на так называемой дистрибутивной гипотезе, которая заключается в том, что семантическое сходство между словами (или другими языковыми единицами) может моделироваться на основе сравнения (сходства) контекстов этих слов [17, 22], т.е. предполагается, что сходство контекстов слов коррелирует с семантическим сходством слов.

Классический подход дистрибутивной семантики заключается в том, что для каждого целевого слова w вычисляется матрица, в которой каждая строка соответствует целевому слову, а каждый столбец представляет собой слова, которые встречались в контексте исходного слова. Таким образом, каждое слово представляется вектором слов, которые встречаются в его контексте, т.е. геометрически слово представляется точкой в n -мерном пространстве, размерность которого равна количеству разных слов, найденных в текстовой коллекции [14, 43].

Задача нахождения сходства между словами сводится в таком случае к определению близости расположения соответствующих точек. Близость расположения точек может вычисляться на основе подсчета Евклидова расстояния между точками. Однако в этом случае расстояние будет сильно зависеть от частоты употребления слова в корпусе. Поэтому предполагается, что небольшой угол между векторами, исходящими из точки начала координат важнее, чем близость по Евклидову расстоянию. Для вычисления близости по углу можно считать косинусную меру близости, в которой предполагается, что чем меньше угол, тем ближе косинус этого угла к нулю. В таком случае сходство между словами можно искать с помощью скалярного произведения между векторами-контекстами [14, 43]. Чем ближе скалярное произведение к 1, тем выше семантическое сходство между словами.

В настоящее время большую популярность получил подход к построению векторных представлений слов на основе нейронных сетей. Считается, что во многих задачах нейросетевые подходы превосходят традиционные подходы дистрибутивной семантики, кроме того, созданы более эффективные инструменты расчета этих представлений.

В 2013 году в работе [25] была предложена эффективная реализация метода создания векторных представлений слов относительно небольшого числа размерностей (обычно от 100 до 1000) на основе нейронных сетей – пакет word2vec. Основная идея подхода состоит в том, чтобы обучать такие векторные представления (называемые word embeddings - эмбединги) на текстовом корпусе в процессе предсказания слов в контексте для одного или нескольких слов. В настоящее время, опубликованы уже насчитанные векторные представления на корпусах большого объема, которые позволяют использовать их в различных задачах автоматической обработки текстов без дополнительных вычислительных затрат [5, 21].

При извлечении таксономических отношений на основе векторных представлений предполагается, что гипонимы и гиперонимы, как близкие по смыслу слова, используются в похожих контекстах, что должно приводить к формированию похожих векторных представлений. Однако нужно отметить, что векторные представления слов сами по себе не позволяют определить тип отношения, которым связаны два слова. Близкими по векторным представлениям к исходному слову могут оказаться как синонимы, гипонимы или гиперонимы целевого слова, так и антонимы (дорогой – дешевый), а также тематически связанные слова (например, собака – ошейник) и др. Поэтому исходные векторные представления слов используются как основа для дальнейшего применения двух базовых подходов для извлечения гиперонимов: подходов без учителя (unsupervised) и подход с учителем (supervised).

В первой группе методов без учителя подразумевается, что существует некоторое преобразование исходных векторов слов, которое лучше соответствует отношению

гипоним-гипероним. В частности, может предполагаться, что контексты слова-гиперонима включают контексты слова гипонима – так называемая гипотеза дистрибутивного включения (Distributional Inclusion Hypothesis) [32].

Вторая группа методов (с учителем) использует исходные векторные представления как признаки для последующего применения специализированных классификаторов, обученных на наборах данных гипоним-гипероним [4]. Первые подходы с учителем использовали арифметические операции над векторами слов (сложение, вычитание, конкатенация), к которым затем применялся классификатор (логистическая регрессия, метод опорных векторов и т.п.) для предсказания гиперонима для слова [4, 32, 34].

В рамках дальнейшего развития методов с учителем на основе векторного пространства эмбедингов было сделано предположение, что можно попытаться найти линейное преобразование, которое преобразует гипоним в гипероним – данный подход называется *projection learning* [15]. Однако в дальнейших экспериментах было показано, что невозможно найти единое преобразование гипонима в гипероним, поскольку оптимальное преобразование отличается для разных семантических классов слов. Поэтому были предложены методы одновременной семантической кластеризации слов и поиска подходящего преобразования гипоним-гипероним, соответствующего каждому семантическому классу [6, 44, 47].

Для вышеупомянутых операций в векторном пространстве слов важен исходный корпус, в частности его размер и жанр текстов (интернет-страницы, Википедия, тексты предметной области и др.). Для улучшения качества векторных представлений был предложен подход на основе так называемых мета-эмбедингов, т.е. объединений различных векторных представлений, полученных на разных текстовых корпусах [7, 42]. В качестве мета-эмбедингов могут выступать такие комбинации векторов как конкатенация или усреднение. Более гибким методом является подход на основе так называемых автокодировщиков, задача которых, стартуя с нескольких различных векторных

представлений, найти наилучшее комбинированное представление (мета-эмбединг), из которого можно наилучшим образом восстановить исходные представления, а также добиться наилучшего соответствия мета-эмбедингов некоторому дополнительному критерию. В задаче предсказания гиперонимов таким критерием может быть использование так называемого триплет-лосса – мета эмбединг должен быть похож на вектора близких по смыслу слов и не похож на вектора случайных слов [42].

1.3. Комбинированные методы извлечения таксономических отношений

Лингвистические методы извлечения отношений могут использовать лексико-семантические шаблоны или синтаксические отношения слов [40], совместную встречаемость слов в предложении [48] и др. Методы извлечения отношений на основе дистрибутивных подходов базируются на извлечении контекстов употребления слов и вычислении их сходства как векторов, т.е. эти подходы позволяют извлекать отношения, которые не указываются в текстах явным образом.

Достоинства обоих методов сочетают методы инкрементального пополнения таксономии [40, 47], при которых все вышеперечисленные характеристики (встречаемость в шаблонах, наличие сходства в разного вида контекстах (линейных и синтаксических, локальных и глобальных), совместная встречаемость) служат как признаки, на основе которых принимается решение о присоединении очередного слова к создаваемой таксономии. При принятии решений о присоединении к таксономии в работе [40] используется принцип максимизации условной вероятности отношения на основе имеющихся признаков, в работе [48] используется принцип оптимизации таксономической структуры и моделировании уровня абстрактности слова в таксономической структуре.

В модели НуреNET [35] для представления предложения используются три вектора: два вектора исходных слов и третий вектор, который представляет собой векторное представление для синтаксического пути (=пути в синтаксическом дереве) между двумя

словами. Для создания вектора синтаксического пути используется рекуррентная нейронная сеть LSTM. Результаты данной работы значительно превосходили предыдущие результаты в задаче извлечении гиперонимов как задачи классификации, по сравнению с лингвистическими и дистрибутивными подходами, поскольку модель еще обучается еще и оптимальному кодированию синтаксических путей между словами вместо заранее заданных лексико-семантических шаблонов.

Предсказание гиперонимов посредством порождения ответов на вопрос

В связи с появлением больших предобученных нейросетевых моделей типа BERT [13] произошло улучшение в решении многих задач автоматической обработки текстов. В этой модели было предложено использовать нейронную сеть архитектуры трансформер с механизмом самовнимания [45] для предобучения на больших объемах неразмеченных текстовых данных для решения двух задач: 1) предсказания маскированного слова в контексте, 2) предсказания, следует ли второе предложение из первого. В результате модель BERT создает контекстуализированные (зависящие от контекста) представления слов, которые позволяют лучше учесть контекст и решать многие задачи автоматической обработки текстов.

Специфичный подход к обучению модели BERT дал дополнительные возможности решения задач автоматической обработки текстов на основе специально заданного вопроса (подсказки или prompt) [34]. Такой подход стал применяться и в задаче предсказания гиперонимов [46]. В частности, в методе TaxoPrompt [46] в качестве подсказки используется следующий шаблон: “What is parent-of [X]? It is [MASK] (Что есть родитель [X]? Это [MASK]).

В слот X вставляется новое слово, для которого нужно определить гипероним. Далее данный шаблон подается на вход языковой модели BERT, особенность обучения которой является то, что модель обучается предсказывать слово, скрытое маской [13]. В качестве

альтернативных подсказок могут использоваться также следующие: 1) [X], [MASK]. 2) [X] is a [MASK]. 3) [MASK], such as [X].

Для придания дополнительного контекста на вход языковой модели описанный лексико-семантический шаблон подается вместе со вторым предложением, отделенным служебным токеном [SEP]. Вторым предложением может быть толкование целевого слова из словаря, или пути их базовой онтологии, записанные в виде текстовых последовательностей. Для обучения модели строится обучающая коллекция с правильными и неправильными ответами из какой-либо таксономии. Таким образом, модель, предсказывая гипероним, использует и громадный объем данных, на котором происходило предобучение, и толкования из словаря, и векторное представление таксономии.

2. Тестирование методов извлечения таксономических отношений

Для оценки качества извлечения таксономических отношений используются различные подходы, которые могут быть разделены на четыре основных группы.

1. Предсказание отношения гиперонимии как задача классификации на заданном датасете. При такой постановке тестирование подходов осуществляется на основе специального набора данных. Набор содержит позитивные примеры, между которыми имеется нужное отношение, и негативные примеры, между которыми отношение отсутствует. Позитивные примеры обычно берутся из некоторых известных онтологических ресурсов, отрицательные примеры порождаются в результате некоторой процедуры. Таким образом, задача тестирования ставится как задача классификации пары слов из датасета на два класса: есть нужное отношение между словами или нет. Оценка качества в этой задаче производится традиционными мерами для классификации: точность, полнота, F-мера.

В работе [36] представлен обзор различных операций над векторами слов, которые использовались в задачах идентификации гиперонимов и представлено тестирование предсказания гиперонимов на нескольких наборах данных в режимах обучения без учителя (unsupervised) и с учителем (supervised). Показано, что на нескольких наборах данных методы предсказания гиперонимов без учителя превышают качество 0.85, а с учителем на некоторых данных получают идеально верные ответы.

На основе результатов в таких тестированиях результатов можно предположить, что задача определения гиперонима для слова решена. Однако она «решена» только в рамках созданных датасетов. Проблемой такого подхода к тестированию извлечения гиперонимов является то, что данная постановка не является практической. Обычно в таких наборах данных количество позитивных и негативных примеров сравнимо между собой, а на практике количество таксономических отношений между словами на порядки меньше, чем потенциальные отношения между разными парами слов в текстовом корпусе. Например, одним из известных и больших наборов, применяемых для тестирования извлечения отношений гипоним-гипероним, является датасет BLESS [4]. BLESS содержит гиперонимы для 200 конкретных, в основном однозначных существительных. Отрицательные пары содержат ко-гипонимы, меронимы (слова, находящиеся в отношении часть-целое) и случайные пары слов, в итоге датасет содержит всего 14542 пары слов с 1337 положительными примерами, т.е. положительных пар лишь в 10 раз меньше, чем отрицательных, что не соответствует ситуации при извлечении из реальных текстовых коллекций

Кроме того, ряд исследований показали, что часто применяемые методы классификации отношений не обучаются классифицировать пару слов, а скорее запоминают типовую роль слова как гипонима или гиперонима. Например, есть слова, такие как животное, растение, здание, которые являются типичными гиперонимами и могут неоднократно в таком качестве использоваться в датасетах [23, 35]. Таким образом, результаты методов

могут завышаться — так называемая проблема лексического запоминания (lexical memorization) [23].

2. Извлечение гиперонимов из заданного корпуса. При такой постановке тестирования системы должны правильно извлечь гиперонимы из корпуса для заданного набора слов, т.е. системы должны сами найти подходящие кандидаты в гиперонимы. На тестировании SemEval-2018 [11] эта задача ставится как задача ранжирования, т.е. автоматическая система должна создать упорядоченный список кандидатов в гиперонимы для целевых слов, причем правильные кандидаты должны находиться в начале списка. Для оценки качества подходов в этой задаче применяются такие меры как MRR — средняя обратная величина первого правильного ответа, а также MAP (Mean Average Precision), которая представляет собой комбинированную меру точности и полноты и вычисляется как сумма точностей в момент поступления в список правильного ответа, усредненную на количество правильных ответов.

В качестве базовых методов для сравнения в тестировании SemEval-2018 применялись методы предсказания гиперонимов на основе дистрибутивных представлений, которые при тестировании на ограниченных датасетах показывали точность выше 0.7. В данной постановке задачи эти методы выдавали первый правильный гипероним слова в среднем на 30 позиции (MRR 0.03). Это связано с тем, что наиболее похожие на целевое слова по векторной модели слова часто оказываются не гиперонимами, а так называемыми ко-гипонимами (т.е. гипонимами одного и того же гиперонима).

Лучший результат в тестировании SemEval-2018 был достигнут системой CRIM [6], в которой был применен комбинированный подход, включающий лексико-семантические шаблоны для определения гиперонимов и ко-гипонимов, метод типа projection-learning с обучением преобразования пространства от гипонима к гиперониму с одновременным обучением семантических кластеров. Данная система достигла качества 0.33 MRR, что означает, что правильный гипероним предсказывается в среднем на 3 позиции.

В более позднем подходе [3] предлагается использовать не один переход в трансформации матрицы гипоним-гипероним, как в исходном подходе *projection learning*, но и использовать такую трансформацию несколько раз, что выявить вышестоящие гиперонимы как в цепочке: *Билл Клинтон — политик — человек*. Данный подход достигает MRR-0.39 на данных SemEval-2018.

3. Автоматическое порождение таксономии Третьим способом тестирования является задача автоматического порождения таксономии на основе текстового корпуса и сравнения порожденной таксономии с существующей. Такая задача ставилась на семинарах SemEval-2015, 2016 [9, 10]. Таким образом, нужно не только извлечь отношения гиперонимии для некоторого множества целевых слов, но и выстроить из них правильную структуру ациклического графа. В качестве тестовых данных были использованы существующие таксономии из разных областей: химические вещества, оборудование, еда, науки, объемом от 450 вершин (науки) до 17 тысяч вершин (химические вещества). Сами вершины заданы, для них нужно определить таксономические отношения. Слова могут иметь более одного гиперонима, тем самым таксономия может быть полииерархией.

Оценка осуществлялась как стандартная задача классификации с использованием мер точности, полноты и F-меры по сравнению с эталонными ресурсами. Также были выполнены оценки структурного качества полученной таксономии (например, отсутствие циклов). Большинство участников использовали Википедию как корпус и применяли комбинированные методы, включающие лексико-семантические шаблоны и дистрибутивное сходство термов по корпусу. Максимальный результат лучшей системы достиг f-меры 0.39 (точность 0.36). Ручная проверка точности фрагмента таксономии дала лучший результат около 0.6.

4. Автоматическое пополнение таксономии. При такой постановке предполагается, что некоторая общая таксономия имеется, и нужно ее достроить в некоторой предметной

области или на основе новых данных. Впервые решение такой задачи предлагалось и тестировалось на основе пополнения WordNet в работе [40].

Также такая задача, например, ставилась в тестировании SemEval-2016 task 14 [19]. Участникам тестирования нужно было найти место нового слова в иерархии WordNet [26] в качестве синонима или гипонима к существующему понятию. Особенность задачи состояла в том, чтобы достроить таксономию WordNet за счет существующих словарей, поэтому новые слова были снабжены толкованиями из этих словарей. Качество измерялось двумя мерами: Ассигасу — точность установления отношения с учетом расстояния (в количестве отношений) от правильного ответа и Recall — доля слов, для которых система предложила решение. Результаты лучшей системы были немного лучше, чем результаты базового решения, которое выбирало в качестве гиперонима нового слова наиболее частотное значение первого слова той же части речи, что и новое слово, в толковании. При этом точность базового метода была немногим более 0.5, ответы были даны для всех слов, что привело к значению F-меры – 0.68.

Лучшие на текущий момент результаты в тестировании SemEval-2016 task 14 получены системой ТахоЕхран [38], достигшая точности 0.566 и значения F-меры 0.723. Система представляет новое слово и слова-кандидаты в WordNet как эмбединги fastText, толкование представляется как усредненный вектор эмбедингов упомянутых слов. Представление слова-кандидата как усреднение векторов токена и толкования подается в систему подбора подходящего гиперонима, основанную на представлении понятий в таксономии WordNet с использованием графовых нейронных сетей (GNN).

В 2020 году в рамках конференции Диалог было организовано открытое тестирование по пополнению таксономии для русского языка RUSSE-2020 [28]. В тестировании использовались две версии русского варианта WordNet — RuWordNet [24]. Датасет для пополнения состоял из слов из неопубликованной новой версии RuWordNet (RuWordNet 2.0), участвующие системы должны были подобрать подходящий гипероним (точнее

синсет – понятие в RuWordNet) в опубликованной версии RuWordNet. В качестве правильных ответов засчитывались как указанные в RuWordNet 2.0 гиперонимы, так и их вышестоящие на 1 шаг гиперонимы (гиперонимы второго порядка), таким образом моделировалась работа «интеллектуального» помощника эксперта, в рамках которой должно быть найдено приблизительное место в семантической сети для привязывания нового слова. Качество предсказания оценивалось с использованием мер средней точности и MRR.

Базовым компонентом, использованным в работах большинства участников, было извлечение наиболее похожих слов из RuWordNet для целевого слова на основе векторных представлений; синсеты, гиперонимы и гиперонимы второго порядка использовались для формирования синсетов кандидатов в качестве гиперонимов для целевого слова. Этот же подход использовался при создании базового решения, предоставленного организаторами. Многие участники отмечали, что для достаточно конкретных слов, которые входили в набор данных для тестирования, наиболее похожими словами оказались не синонимы или гиперонимы, а ко-гипонимы, т.е. гипонимы того же гиперонима.

Лучшим векторным представлением для предсказания гиперонимов оказались эмбединги fasttext [5], рассчитанные на интернет-корпусе Common Crawl и Википедии¹. Лучшим методом в данном тестировании оказался метод, использующий разнообразные признаки, включая кандидаты на основе эмбедингов, словарные толкования Викисловаря, а также анализ результатов поиска в поисковых системах Яндекс и Google. Данный подход достиг результата 0.55 MRR для предсказания гиперонимов существительных, что соответствует нахождению в среднем первого правильного ответа на втором месте выдачи.

В продолжение данного тестирования и следуя его основной идее (тестирование пополнения таксономии на основе разных версий одного и того же ресурса), в 2021 году

¹ <https://fasttext.cc/docs/en/crawl-vectors.html>

был создан специализированный датасет, использующий три последовательных версии англоязычного WordNet и две версии RuWordNet - датасет был назван Diachronic wordnets [29]. В датасеты входят новые слова более поздних версий, которым нужно предсказать синсет-гипероним, используя более старую версию ресурсов. Таким образом, на данном датасете можно проверять разные подходы пополнения ресурсов типа WordNet параллельно на двух языках.

На созданных наборах данных лучшие результаты были получены на основе мета-эмбедингов, которые комбинируют разные векторные представления, включая эмбединги слов, полученные разными методами на разных корпусах, векторные представления вершин пополняемых ресурсов, а также информацию из толкований Викисловаря. В частности, лучший результат на датасете RUSSE-2020 удалось повысить до 0.6 MRR без всякого использования результатов поисковых систем Интернет [29].

Заключение

Таксономические отношения, связывающие сущность с ее классом, являются центральным отношением при представлении знаний во многих предметных областях. Поэтому много внимания уделяется автоматическим методам выявления таксономических отношений из текстов.

В данной статье мы представили обзор подходов к автоматическому извлечению таксономических отношений между словами, начиная от самых классических таких как лексико-семантические шаблоны. В настоящее время базой для извлечения таксономических отношений являются векторные представления слов, которые отражают сходство контекстов упоминания гипонима и его гиперонима. В статье также представлены новые подходы, работающие на основе языковых моделей типа BERT, и использующие новые варианты работы с лексико-семантическими шаблонами, в которых модель заполняет позицию гиперонима.

В статье также описаны различные способы тестирования качества извлечения таксономических отношений, а также приведены результаты этих тестирований, из которых следует, что несмотря на прогресс в развитии методов извлечения таксономических отношений, качество извлечения является недостаточным для полностью автоматического построения или пополнения таксономий.

Конфликт интересов

Процесс написания и содержание статьи не дает оснований для постановки вопроса о конфликте интересов.

Соответствие этическим стандартам

Данная статья является полностью оригинальным произведением ее авторов, прежде не опубликована и до получения решения редколлегии PRIA о непринятии ее к публикации не будет направляться в другие издания.

Список литературы

1. F. N. Al-Aswadi, H. Y Chan., and K. H. Gan. “Automatic ontology construction from text: a review from shallow to deep learning trend”, *Artificial Intelligence Review*, V. 53, №. 6, xpp. 3901–3928 (2020).
2. A.I. Aldine, M. Harzallah, G. Berio, N. Bechet and A. Faour, “Redefining Hearst patterns by using dependency relations”, *Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - Volume 2: KEOD*, pp. 148–155 (2018).
3. Y.Bai, R. Zhang, F. Kong, J. Chen and Y. Mao, “Hypernym Discovery via a Recurrent Mapping Model”, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 2912–2921 (2021).

4. M. Baroni, R. Bernardi, N.Q. Do., and C. C. Shan, "Entailment above the word level in distributional semantics", Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pp. 23–32 (2012).
5. P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information", Transactions of the Association for Computational Linguistics 5, pp. 135–146 (2017).
6. G. Bernier-Colborne, C. Barrière, "CRIM at SemEval-2018 task 9: A hybrid approach to hypernym discovery", Proceedings of the 12th International Workshop on Semantic Evaluation, pp. 725–731 (2018)
7. D. Bollegala, C. Bao, "Learning word meta-embeddings by autoencoding", Proceedings of the 27th International Conference on Computational Linguistics, pp. 1650–1661 (2018).
8. E. I. Bolshakova, N. E. Efremova, "A heuristic strategy for extracting terms from scientific texts", International Conference on Analysis of Images, Social Networks and Texts, Springer, Cham, pp. 297–307 (2015).
9. G. Bordea, P. Buitelaar, S. Faralli, and R. Navigli, "Semeval-2015 task 17: Taxonomy extraction evaluation (texeval)", Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), pp. 1081–1091 (2016).
10. G. Bordea, E. Lefever, and P. Buitelaar, "Semeval-2016 task 13: Taxonomy extraction evaluation (texeval-2)", Proceedings of the 10th international workshop on semantic evaluation (semeval-2016), pp. 1081–1091, (2016).
11. J. Camacho-Collados, C. Delli Bovi, L., Espinosa-Anke, T. Pasini, E. Santus, V. Shwartz, R. Navigli, and H. Saggion, "Semeval2018 task 9: Hypernym discovery", Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018); ACL, pp. 712–724, (2018).
12. S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis", Journal of the American society for information science, V. 41, №. 6, pp. 391–407 (1990).

13. J. Devlin, Ming-Wei Chang, K. Lee, K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding”, Proceedings of NAACL-2019, pp. 4171–4186, (2019).
14. K. Erk, “Vector space models of word meaning and phrase meaning: A survey”, Language and Linguistics Compass, V. 6, №. 10, pp. 635–653 (2012).
15. R. Fu, J. Guo, B. Qin, W. Che, H. Wang, T. Liu, “Learning semantic hierarchies via word embeddings”, Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1199–1209 (2014).
16. A. Gómez-Pérez, D. Manzano-Macho, “An overview of methods and tools for ontology learning from texts”, The knowledge engineering review, V. 19, №. 3, pp. 187–212 (2004).
17. Z. Harris, “Distributional structure”, Word, 10 (23), pp. 146–162 (1954).
18. M. Hearst, “Automatic acquisition of hyponyms from large text corpora”, In Proceedings of the 14th conference on Computational linguistics, V. 2, pp. 539–545 (1992).
19. D. Jurgens, M. Pilehvar, Semeval-2016 task 14: semantic taxonomy enrichment, In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), 1092–1102 (2016).
20. Y. Kiselev, S. Porshnev, and M. Mukhin, “Method of extracting hyponym-hypernym relationships for nouns from definitions of explanatory dictionaries” (in Russian), Software Engineering 10, 38–48 (2015).
21. A. Kutuzov, E. Kuzmenko, “WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models”, In: Ignatov D. et al. (eds) Analysis of Images, Social Networks and Texts. AIST 2016. Communications in Computer and Information Science, vol 661. Springer, Cham (2017).
22. A. Lenci, “Distributional semantics in linguistic and cognitive research”, Italian journal of linguistics, V. 20, №. 1, pp. 1–31 (2008).

23. O. Levy, S. Remus, Ch. Biemann, and I. Dagan, “Do supervised distributional methods really learn lexical inference relations?”, In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 970–976 (2015).
24. N. Loukachevitch, G. Lashevich, A. Gerasimova, V. Ivanov, and B. Dobrov, “Creating Russian wordnet by conversion”, In Computational Linguistics and Intellectual Technologies: papers from the Annual conference” Dialogue, 2016. pp. 22–30 (2016).
25. T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. “Distributed representations of words and phrases and their compositionality”, In Advances in neural information processing systems, 3111–3119 (2013).
26. G. A. Miller, “WordNet: a lexical database for English”, Communications of the ACM,. 1995. V. 38, №. 11. – pp. 39-41. (1995).
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. – pp. 1135-1145.
28. Nikishina I., Logacheva, V., Panchenko, A., Loukachevitch, N. RUSSE’2020: Findings of the First Taxonomy Enrichment Task for the Russian Language // Computational Linguistics and Intellectual Technologies: papers from the Annual conference “Dialogue”, 2020.
29. Nikishina I., Panchenko, A., Logacheva, V., Loukachevitch, N. Studying Taxonomy Enrichment on Diachronic WordNet Versions // Proceedings of the 28th International Conference on Computational Linguistics: Association for Computational Linguistics, 12.2020.
30. Nikishina I., Tikhomirov, M., Logacheva, V., Nazarov, Y., Panchenko, A., & Loukachevitch, N. Taxonomy Enrichment with Text and Graph Vector Representations // Semantic Web, 2022. V. 13, № 3. – pp. 441-475.

31. Stephen Roller, Katrin Erk, and Gemma Boleda. Inclusive yet selective: Supervised distributional hypernymy detection // In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, 2014. – pp 1025–1036.
32. Roller S., Kiela D., Nickel M. Hearst Patterns Revisited: Automatic Hypernym Detection from Large Text Corpora // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2018. – pp. 358-363.
33. Sabirova K., Lukanin A. Automatic Extraction of Hypernyms and Hyponyms from Russian Texts // Supplementary Proceedings of the 3rd International Conference on Analysis of Images, Social Networks and Texts (AIST 2014). 2014. – pp. 35-40.
34. Schick T., Schutze H. Exploiting cloze-questions for few-shot text classification and natural language inference // Proceedings EACL-2021, 2021. – pp. 255–269, 2021.
35. Shwartz V., Goldberg Y., Dagan I. Improving hypernymy detection with an integrated path-based and distributional method. // In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016.. – pp. 2389–2398.
36. Shwartz V., Santus E., Schlechtweg D. Hypernyms under Siege: Linguistically-motivated Artillery for Hypernymy Detection // Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, 2017. – pp. 65-75.
37. Seitner, J., Bizer, C., Eckert, K., Faralli, S., Meusel, R., Paulheim, H., Ponzetto, S. P. A large database of hypernymy relations extracted from the web // LREC-2016, 2016.
38. Shen J., Shen, Z., Xiong, C., Wang, C., Wang, K., Han, J. TaxoExpan: Self-supervised taxonomy expansion with position-enhanced graph neural network // Proceedings of The Web Conference 2020, 2020. – pp. 486-497.
39. Snow R., Jurafsky D., Ng A. Learning syntactic patterns for automatic hypernym discovery // Advances in neural information processing systems, 2004.

40. Snow R., Jurafsky D., Ng A. Semantic taxonomy induction from heterogeneous evidence // In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, 2006. – pp. 801–808.
41. Tian F., Yuan C., Ren F. Hyponym extraction from the web by bootstrapping //IEEEJ transactions on electrical and electronic engineering, 2012. V. 7, №. 1. – pp. 62-68.
42. Tikhomirov M., Loukachevitch N. Meta-Embeddings in Taxonomy Enrichment Task // Computational Linguistics and Intellectual Technologies: papers from the Annual conference Dialogue, 2021. – pp. 681-691.
43. Turney P. D., Pantel P. From frequency to meaning: Vector space models of semantics //Journal of artificial intelligence research. – 2010. – T. 37. – C. 141-188.
44. Ustalov D., Arefyev, N., Biemann, C., & Panchenko, A. Negative Sampling Improves Hypernymy Extraction Based on Projection Learning // Proceedings of EACL 2017, 2017. – pp. 543.
45. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. Attention is all you need //Advances in neural information processing systems, 2017. – V. 30.
46. Xu, H., Chen, Y., Liu, Z., Wen, Y., & Yuan, X. TaxoPrompt: A Prompt-based Generation Method with Taxonomic Context for Self-Supervised Taxonomy Expansion.
47. Yamane J., Takatani, T., Yamada, H., Miwa, M., & Sasaki, Y. Distributional hypernym generation by jointly learning clusters and projections //Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, 2016. – pp. 1871-1879.
48. Yang H., Callan J. A metric-based framework for automatic taxonomy induction // 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. – pp. 271-279.

AUTOMATIC METHODS FOR EXTRACTING TAXONOMICAL RELATIONS FROM TEXTS

This paper provides an overview of approaches to automatic extraction of taxonomic relationships between words, including classical methods based on lexico-semantic patterns, vector representations of words, and their modern development. It also describes new approaches based on language models like BERT and using new options for working with lexico-semantic templates, in which the model fills the target position in the template. The paper presents various ways to test the quality of extracting taxonomic relations and the results of these tests. According to the results currently achieved, it can be concluded that despite the progress in the development of methods for extracting taxonomic relations, the quality of extraction is insufficient for fully automatic construction or enrichment of taxonomies..

Keywords: taxonomy, class-subclass relations, hypernyms, knowledge acquisition from texts

Информация об авторах



Наталья В. Лукашевич (год рождения 1964) закончила факультет вычислительной математики и кибернетики МГУ имени М.В. Ломоносова, защитила диссертацию на степень доктора технических наук в 2016 году. В настоящее время является ведущим научным сотрудником НИВЦ МГУ имени М.В. Ломоносова. Читает лекции по автоматической обработке текстов и информационному поиску на факультете вычислительной математики и кибернетики и филологическом факультете МГУ имени М.В. Ломоносова, а также в МГТУ имени Н.Э. Баумана. Имеет более 300 публикаций в области автоматической обработки текстов, информационного поиска и представления знаний. Научные интересы: автоматическая обработка текстов, информационный поиск, онтологии.