УДК 004.682 *Научная статья* DOI: 10.18287/2223-9537-2023-13-4-520-530

Формирование лексического модуля прикладной онтологии для её обучения

© 2023, П.А. Ломов

Кольский научный центр Российской академии наук,
Институт информатики и математического моделирования им. В.А. Путилова, Апатиты, Россия

Аннотация

Трудоёмкость разработки онтологий связана с их обучением, которое предполагает автоматизацию решения задач извлечения онтологических понятий и отношений из естественно-языковых текстов. Для созданной онтологии, используемой в информационной системе, с большой долей вероятности возникнет необходимость расширения её понятийной системы ввиду усложнения или изменения процессов обработки данных. Последующее применение текстового анализа для обнаружения новых понятий и включения их в онтологию, т.е. связывания их с существующими понятиями, потребует идентификации последних в предложениях естественно-языковых текстов. В данной работе рассматривается задача автоматизации формирования лексического модуля прикладной онтологии, включающего формализованные представления онтологических понятий в естественно-языковых текстах. Представлен обзор работ, посвящённых использование лексической информации о компонентах онтологии при решении задач, связанных с анализом текстовых данных. Рассмотрены способы применения модели *OntoLex-Lemon* для определения структуры лексических представлений понятий онтологии. Предложена процедура формирования лексических представлений на основе анализа текстов предметной области, учитывающая наличие у понятий имён, состоящих из нескольких слов. Приведены результаты применения полученного модуля для автоматического формирования обучающего набора нейросетевой языковой модели, используемой в задаче обучения онтологии для обнаружения новых понятий в корпусе предметных текстов.

Ключевые слова: *лексический модуль, прикладная онтология, OntoLex-Lemon, обучение онтологии.*

Цитирование: *Ломов П.А.* Формирование лексического модуля прикладной онтологии для её обучения // Онтология проектирования. 2023. Т.13, №4(50). С.520-530. DOI: 10.18287/2223-9537-2023-13-4-520-530.

Конфликт интересов: автор заявляет об отсутствии конфликта интересов.

Введение

Разработка онтологий включает: определение словаря понятий онтологии, задание иерархических и предметных отношений между ними, наполнение онтологии экземплярами, формализацию ограничений предметной области (ПрО) в виде логических выражений, тестирование и др. Полученная онтология представляет знания о ПрО и предполагает её дальнейшее развитие, которое является следствием изменения условий эксплуатации и появления задач, требующих учёта дополнительных предметных знаний. В этой связи актуальным является обучение онтологий [1], т.е. автоматизация процесса создания онтологий и их последующего расширения и усложнения. Основным источником новых знаний для пополнения и/или изменения понятийной системы, представленной в онтологии, являются естественно-языковые (ЕЯ) тексты (ЕЯТ) – одна из форм фиксации знаний в различных ПрО. Поэтому обучение онтологий на основе анализа текстов и последующего представления их фрагментов в виде новых понятий, связанных отношениями, является распространённой практикой.

Обучение онтологий на основе анализа ЕЯТ можно рассмотреть с позиции триады «знак-смысл-предмет» - т.н. треугольника Фреге [2], модели, предложенной для понимания семантики знаковых выражений. В тексте в качестве знака имеется слово или словосочетание и подразумевается некоторый его смысл, зависящий от контекста употребления слова и имеющихся знаний о ПрО. С учётом этого можно указать на некоторый предмет реального мира. В онтологии знаком может выступать компонент онтологии - класс, экземпляр, свойство. Смысл знака может задаваться по-разному в зависимости от используемого языка описания онтологии: в случае *RDF/RDFS* (*Resource Description Framework*/ *RDF Schema*), *SKOS* (*Simple Knowledge Organization System*) - его положением в понятийной системе; в случае *OWL* (*Ontology Web Language*) - множеством логических выражений (*OWL*-аксиом), которое формально определяет множество интерпретации знака.

При выявлении в тексте новых понятий или отношений, которые могут быть добавлены в онтологию, необходимо для имеющихся в ней элементов (онтологических знаков) обнаружить такие слова и/или словосочетания в тексте (текстовые знаки), которые бы соответствовали друг другу, т.е. обозначали одинаковые предметы. Такое увязывание знаков и предметов в двух триадах (онтологической и текстовой) позволяет сопоставить и соответствующие им смыслы, представленные текстовым и онтологическим (т.е. набором логических выражений и фрагментом понятийной системы онтологии) контекстами. В результате сопоставления появляется возможность расширить онтологический смысл на основе текстового, т.е. добавить в онтологию новые понятия.

Для решения задачи поиска понятий онтологии в тексте требуется наличие некоторого их лексического представления, включающего компоненты соответствующей текстовой триады «знак-смысл-значение». Объём и сложность такого лексического представления значительно влияют на эффективность обучения онтологии, применяемые методы и полученные результаты. Важным требованием является также его формирование именно автором онтологии на этапе её разработки, т.к. это позволяет избежать искажения и/или утраты авторского смысла.

При решении задач с применением онтологий ввиду, как правило, отсутствия такого специального лексического представления в качестве него используют наименования понятий и/или отношений, которые трактуются как текстовые знаки, тождественные онтологическим - классам, экземплярам, отношениям или атрибутам. Наличие такого базового лексического представления позволяет с некоторой степенью достоверности обнаруживать в ЕЯТ знаки (слова или словосочетания), соответствующие онтологическим. Однако для обучения онтологий использование только онтологических наименований является недостаточным, т.к. в текстах понятия или отношения могут быть представлены различными группами слов. В данной работе рассматривается проблема автоматизации формирования модуля, содержащего лексические представления для понятий прикладной онтологии, а также возможности его дальнейшего использования на примере решения задачи обучения онтологий.

1 Обзор работ

Существует множество работ, связанных с обработкой лексической информации с применением онтологических моделей. Среди них можно выделить несколько групп в зависимости от использования существующей или создания новой лексической онтологии и способов её формирования и применения. В одну группу можно отнести работы, выполняемые в рамках концепции связанных лингвистических данных (*Linguistic Linked Open Data*) [3], направленной на представление с помощью языков семантической паутины существующих лингвистических ресурсов (словарей, тезаурусов и др.), а также результатов анализа и аннотирова-

ния ЕЯТ. Целью в этом случае является обеспечение возможности совместного использования различных лингвистических данных в веб-сервисах, ориентированных на решение задач обработки ЕЯ информации. Так, в работе [4] рассматривается отображение *Open-de-WordNet*, аналога *WordNet* для немецкого языка, в онтологическую модель *OntoLex-Lemon* (*Lexicon Model for Ontologies*), а также возможности включения в неё дополнительной лексической информации для слов помимо их лемм и тэгов.

В работах, использующих лингвистические онтологии, последние обычно представляют собой объёмные и признанные научным сообществом модели. При этом язык их описания и структура могут сильно отличаться от онтологий, представленных в виде семантической сети или таксономии. Например, работа [5] посвящена применению лексической онтологии в контексте решения проблемы извлечения мнений из ЕЯТ, относящихся к туристической тематике. В этом исследовании онтология *WordNet* используется в случаях отсутствия слова в словаре и недостаточного количества размеченных образцов на этапе обучения. Благодаря ей удаётся расширить список ключевых слов, ассоциированных с определённой категорией, путём добавления в словарь их синонимов. Это позволяет повысить точность классификации найденных мнений, включающих слова, отсутствующие в обучающем наборе.

В работах, в которых производится предварительное формирование лингвистической онтологии, рассматривается создание новой либо наполнение собранными данными существующей онтологии. Например, в работе [6] рассматривается процедура формирования в онтологии лексического слоя в виде некоторой дополнительной системы классов: «Лексическое значение», «Лексическая сущность», «Контекст», «Слово», «Фраза» и др. С их помощью осуществляется представление связей слова, употребляемого в нескольких контекстах, с понятиями онтологии и разными смыслами. Контекст и смысл в этом случае представляются в виде экземпляров соответствующих классов. Каждый экземпляр представляет некоторый набор слов. Применение данного слоя позволило авторам в ходе экспериментов определять смысл слов, встречающихся в анализируемых текстах. В качестве примера наполнения существующей онтологии можно привести работу [7]. В ней авторы констатируют утрату лексической информации при переходе от термина, представленного в тексте (в лексическом слое), к его представлению в онтологии (в онтологическом слое) и показывают необходимость создания интерфейса между ними в виде лексической онтологии. Рассматривается применение такого интерфейса в веб-сервисе, ориентированном на поиск кулинарных рецептов в сети Интернет. Анализ лексической информации (синонимов и гиперонимов терминов) позволяет находить большее количество рецептов. Онтологический слой даёт возможность расширить ЕЯ поисковый запрос и получить наиболее подходящие профилю пользователя рецепты. Основой лексической онтологии является *OntoLex-Lemon* [8], включающая набор классов и отношений для представления необходимых данных в виде экземпляров.

В приведённых работах создание лексической онтологии не является основной целью, а рассматривается в контексте решения прикладной задачи. Примером автоматического формирования лексического слоя для существующей онтологии является работа [9]. В ней исследуются возможные способы для добавления в существующую онтологию лексической информации, т.н. её лексикализации. В качестве основы её представления применяется онтология *OntoLex-Lemon*. Предлагаются два подхода получения лексической информации для классов и отношений онтологии соответственно. Первый подход включает нахождение синонимов наименований классов в некотором внешнем словаре, последующее их сохранение в виде экземпляров класса «Лексическая единица» онтологии *OntoLex-Lemon*, связываемых отношениями с понятиями исходной онтологии.

Во втором подходе из исходной онтологии извлекаются триплеты вида: «экземпляр класса А» - «отношение» - «экземпляр класса Б», включающие отношения, для которых необхо-

димо сформировать лексикализацию. Из текстового корпуса извлекается набор предложений, содержащих оба экземпляра одного триплета, и создаются деревья синтаксических зависимостей, которые конвертируются в триплеты и сохраняются в хранилище триплетов. В результате получается некоторый набор триплетов, соответствующих предложениям с возможными лексикализациями. Далее определяются несколько паттернов извлечения лексикализаций в виде *SPARQL*-запросов к хранилищу триплетов. Подставляя в данные паттерны наименования отношений исходной онтологии, собираются триплеты, представляющие возможные лексикализации отношений.

Описанный подход к извлечению лексикализаций использует веб-ресурс *DBpedia* в качестве прикладной онтологии и ресурс *Wikipedia* в качестве текстового корпуса. *DBpedia,* как онтология, ввиду своего значительного объёма содержит большое множество разнообразных экземпляров отношений, что позволяет с большой вероятностью найти некоторый набор предложений, где они используются. Увеличивает данную вероятность то, что ресурс *Wikipedia*, в текстах которого осуществляется поиск, является источником информации для формирования *DBpedia*. В случае применения такого подхода для формирования лексического модуля для произвольной прикладной онтологии ввиду её сравнительно меньшего размера и возможного отсутствия экземпляров классов, а также использования небольшого текстового корпуса, включающего специализированные тексты, получение приемлемого результата маловероятно.

2 Применение *OntoLex-Lemon* для представления лексической информации

В качестве основы лексической онтологии предлагается использовать модель *OntoLex-Lemon*, имеющую широкие возможности для представления лингвистических ресурсов в виде *OWL*-онтологий. Она включает базовый модуль, содержащий основные понятия и отношения для описания лексем, их форм и связывания их с понятиями прикладных онтологий, а также набор дополнительных модулей, которые расширяют описательные возможности для представления специфических атрибутов и отношений лексем:

- *Syntax and Semantics* (*synsem*) представляет необходимые условия употребления лексемы в текстах;
- *Decomposition* (*decomp*) используется для описания словосочетаний и многословных терминов;
- *Linguistic Metadata* (*lime*) позволяет задавать метаописания для наборов лексем;
- *Lexicography Module* (*ontolex*) определяет словарь для представления различных морфологических и синтаксических атрибутов лексем, например, часть речи, падеж, залог, одушевлённость и т.д.

Основными понятиями базового модуля (рисунок 1) являются:

- Словарная единица (*Lexical Entry*) представляет собой слово, словосочетание или аффикс, которые рассматриваются в качестве лексического представлением некоторой онтологической сущности (*Ontology Entity*) – класса, отношения, атрибута и др.;
- Лексическое понятие (*Lexical Concept*) представляет собой мысленный образ, который вызывается в сознании (отношение «*evokes*») некоторой словарной единицей (лексическое понятие может быть связано с несколькими словарными единицами);
- Лексический смысл (*Lexical Sense*) определяет своего рода контекст соотнесения словарной единицы с лексическим и/или онтологическим понятием (по сути, оно является реификацией отношения «*denotes*»);
- Онтологическая сущность (*Ontology Entity*) представляет собой некоторый компонент (класс, экземпляр, отношение, атрибут и др.) из внешней онтологии.

Данные понятия позволяют описать упомянутую триаду «знак-смысл-предмет» с использованием лишь данного набора при моделировании лексического представления с различной степенью детализации. Например, в простом случае можно связать отношением «*denotes*» онтологическую сущность со словарной единицей, представляющей её в тексте. Далее можно дополнить такое описание лексическим смыслом, содержащим, например, словарную статью, и тем самым дать ответ на вопрос: почему данная словарная единица и онтологическая сущность соответствуют друг другу?

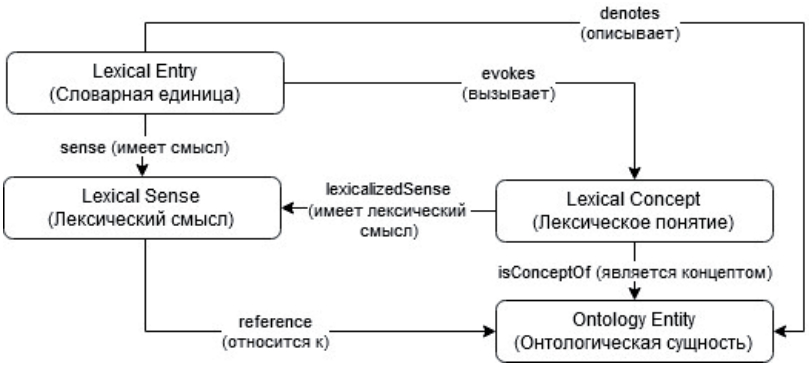

Рисунок 1 - Основные понятия онтологии *OntoLex-Lemon*

Отдельно следует рассмотреть представление смысла синонимов. Авторами *OntoLex-Lemon* указывается на то, что лексический смысл лексикализируется в одной словарной единице, поэтому он не может быть соотнесён с несколькими единицами. В этом случае для объединения синонимов следует использовать лексическое понятие, которое может иметь несколько лексических смыслов, связанных непосредственно со словарными единицами, представляющими синонимы.

В контексте данного исследования важным является определение способа моделирования лексического представления онтологических понятий, которые имеют имена, состоящие из нескольких слов. Для этого предлагается использовать схему, представленную на рисунке 2. Согласно приведённой схеме ЕЯ наименование-словосочетание представляется экземпляром класса «Многословная словарная единица» (*MultiWordExpression*), который является подклассом *Lexical Entry*. Для каждого из составляющих его слов-частей создаётся экземпляр класса «Часть» (*Component*) и экземпляр класса *Lexical Entry*. Оба экземпляра связываются отношением «Соответствует» (*correspondsTo*). Такое представление слова-части в виде двух экземпляров обусловлено тем, что в составе словосочетания оно будет иметь конкретные характеристики (например, род, число, порядковый номер), которые будут не релевантны её общему представлению в виде *Lexical Entry*. На схеме также присутствует класс «Форма представления» (*Form*), экземпляры которого являются письменным представлением лексической единицы и/или её леммой.

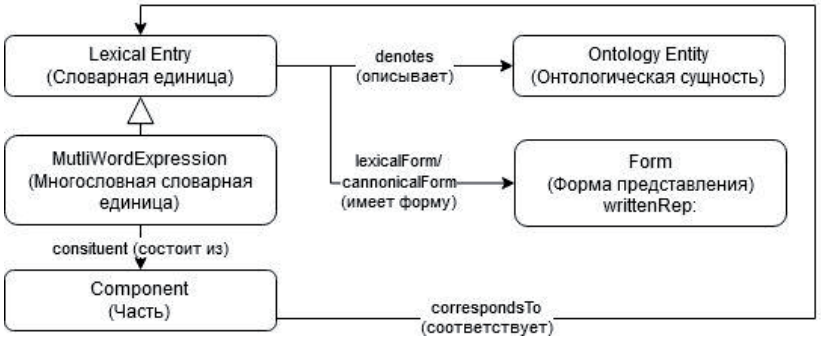

Рисунок 2 - Схема представления многословных наименований понятий с помощью *OntoLex-Lemon*

3 Процедура формирования лексического модуля

Формирование лексического модуля онтологии начинается с анализа её классов и экземпляров и последующего создания на основе их наименований исходных лексических пред-

ставлений. В результате лексический модуль будет включать множество экземпляров классов *Lexical Entry* и *MultiWordExpression*, соответствующих наименованиям, состоящим из одного или нескольких слов.

На следующем этапе производится расширение модуля путём добавления в него альтернативных лексических представлений понятий. Для этого выполняется автоматический анализ корпуса текстов ПрО, из которых извлекаются предложения, содержащие наименования понятий онтологии. Для каждого предложения с помощью программной библиотеки *spacy* [10] формируется дерево синтаксических зависимостей, которое является результатом синтаксического разбора предложения и отражает синтаксические отношения между его словами в виде древовидного графа. Его корневой (главной) вершиной обычно выступает сказуемое, к которому присоединяются подлежащее и другие члены предложения.

В ходе дальнейшего анализа такого дерева из соответствующего ему предложения извлекается именная группа, имеющая в качестве главного слова наименование понятия или, в случае наименования-словосочетания, главное слово этого словосочетания. При этом различия порядка и формы слов, а также наличие слов, не присутствующих в исходном наименовании, игнорируются.

Например, на рисунке 3 представлено синтаксическое дерево для предложения «Международный аэропорт Шереметьево является крупнейшим в России». В случае наименования понятия «аэропорт Шереметьево» в данном дереве находятся вершины, соответствующие словам «аэропорт» и «Шереметьево». Из них выбирается вершина, которая соответствует главному слову наименования - «аэропорт». В результате она и все её потомки, составляют результирующую именную группу «Международный аэропорт Шереметьево».

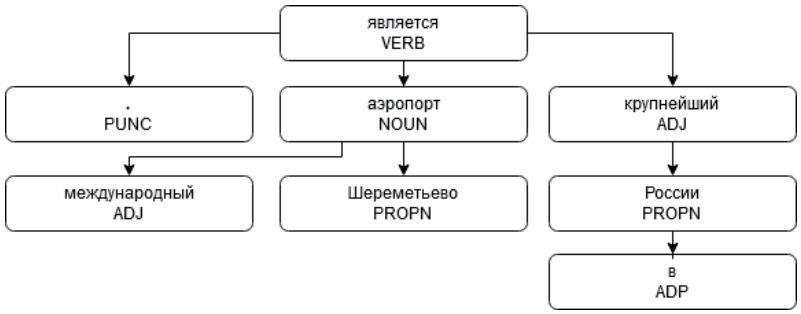

Рисунок 3 - Пример дерева синтаксических зависимостей предложения

Полученные таким образом именные группы включаются в лексический модуль как экземпляры *MultiWordExpression* (рисунок 4).

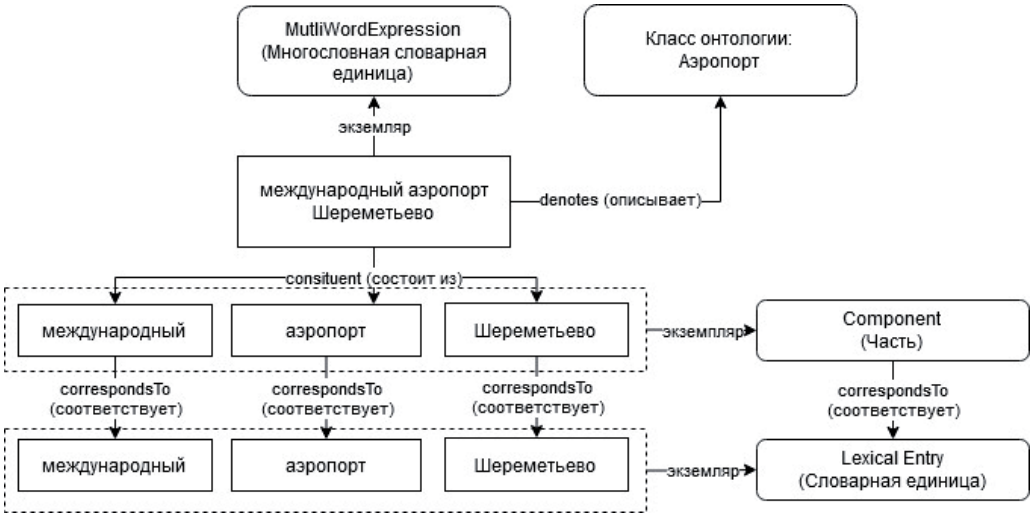

Рисунок 4 - Пример задания лексического представления для онтологической сущности - класса "Аэропорт" в лексическом модуле в виде экземпляра MultiWordExpression

На финальном этапе производится экспертная проверка расширенного модуля, в ходе которой удаляются и/или корректируются добавленные лексические представления.

4 Применение лексического модуля при обучении онтологии

Проверка целесообразности формирования лексического модуля онтологии проверялась путём его использования для решения задачи обучения онтологий с помощью технологии, предложенной в работе [11]. Её применение предполагает автоматизированное создание обучающего набора для тренировки нейросетевой языковой модели, ориентированной на решение задачи извлечения именованных сущностей. Обучающий набор строится на основе текстов ПрО путём представления предложений, содержащих наименования классов и экземпляров онтологии в виде троек: предложение, границы понятия, категория. Предполагается, что обученная с его помощью модель будет извлекать из текстов новые понятия на основе контекстов, сходных с теми, в которых употребляются понятия обучающего набора.

Важным является определение границ понятия в случаях, когда понятие задано в разных предложениях разным количеством слов. Ошибка в определении границ приводит к неверному определению контекста понятия, снижает качество обучения и способность модели правильно обнаруживать новые понятия для пополнения онтологии.

Для решения данной задачи на этапе создания обучающего набора был использован лексический модуль. В качестве исходного множества понятий обучаемой онтологии были приняты понятия, характерные для новостных текстов, такие как: президент, государство, событие, компания, персона и др. Формирование модуля осуществлялось на основе фрагмента специализированного текстового корпуса (корпус русскоязычных статей Интернет-издания *Lenta.ru*) [12]. В состав модуля вошли различные лексические представления понятий онтологии, например: президент страны, акции компании, акция протеста, президент компании, международная компания и др.

Применение лексического модуля на этапе формирования обучающего набора позволило в значительно большем числе предложений правильно определить границы для понятий с наименованиями-словосочетаниями, т.к. в модуле содержалось несколько лексических представлений, из которых выбиралось наиболее подходящее, тогда как использование только наименования понятия давало лишь одно.

Для последующего сравнения результатов был сформирован обучающий набор без использования лексического модуля, т.е. с учётом только наименований понятий. Было обучено две модели с применением лексического модуля и без него. Оценка эффективности обнаружения новых понятий моделями проводилась на основе сравнения результатов работы моделей с эталонным набором, сформированным экспертом ПрО и включающим понятия, которые должны быть извлечены языковой моделью из текстов.

Сравнение производилось по следующим критериям:

- полнота = R/N, где R - количество понятий эталонного набора, обнаруженных моделью, N – общее количество понятий в эталонном наборе;
- точность = R/M, где R - количество понятий эталонного набора, обнаруженных моделью, M – общее количество понятий, обнаруженных моделью.

Были получены следующие результаты работы моделей:

- модель, обученная без использования лексического модуля: точность 0.002, полнота 0.03;
- модель, обученная с использованием лексического модуля: точность 0.06, полнота 0.65.

Разница между полученными результатами позволяет сделать вывод о повышении эффективности обучения при использовании набора, полученного с применением информации из лексического модуля онтологии.

Наряду с использованием модуля для установления правильных границ понятия в предложении при формировании обучающего набора, его можно также применять, например, в задачах связывания именованных сущностей (*Named Entity Linking*) [13] или их разрешения (*Entity Resolution*) [14]. Эти задачи предполагают сопоставление сущности с некоторым фрагментом данных (текстом, строкой в таблице и т.п.). Для этого может быть использована информация из лексического модуля, например, лексические представления понятий.

Следует заметить, что для данных задач существуют специальные подходы с применением нейронных сетей [15, 16]. Однако их внедрение потребует создание и сопровождение соответствующих программных компонентов, тогда как использование лексического модуля в качестве источника данных о соответствии онтологических и текстовых знаков может обеспечить приемлемый уровень качества решения с меньшими затратами труда и времени.

Еще одним направлением использования лексического модуля может быть генерация так называемых затравок для больших языковых моделей [17], получивших распространение в области анализа ЕЯТ в последние годы. Ввиду существенных требований к аппаратной платформе для проведения их дообучения путём тонкой настройки [18] популярность приобрёл альтернативный способ, называемый контекстным обучением [19]. В этом способе вместе с входными данными передаётся некоторая дополнительная (подсказывающая) информация – затравка (*prompt*), которая может быть строкой ключевых слов, иногда включающей форматирование. Это позволяет направить модель на генерацию более точного и полного ответа. Для создания затравок в ряде работ рассматривается применение онтологий [19, 20] в качестве источника данных. В этом случае целесообразным является применение для их создания лексического представления понятия из лексического модуля онтологии в качестве дополнения или альтернативы онтологическим элементам (идентификатору *IRI*, аннотациям *rdfs:label, rdfs:comment*).

В этом контексте лексический модуль можно рассматривать уже как интерфейс между концептуальным уровнем, представленным понятийной системой онтологии, и большой языковой моделью, которая выступает в качестве некоторой замены огромного множества ЕЯТ, т.к. потенциально может сгенерировать их. Однако такая модель явным образом оперирует лишь элементами последовательностей – токенами, которые представляют собой упорядоченные множества символов, часто встречающихся вместе [21]. Таким образом, «понимание» моделью синтаксиса и семантики языка сводится к её способности предоставлять список наиболее вероятных токенов для их добавления в определённую позицию последовательности.

Заключение

В работе рассматривается задача автоматизации формирования лексического модуля для прикладной онтологии. Его создание представляет собой один из этапов разработки онтологии, который позволяет упростить её дальнейшее использование при решении задач, связанных с обработкой ЕЯТ, т.к. особенности сопряжения текстовых и онтологических знаков, характерных для ПрО, будут явно заданы разработчиком.

Применение такого модуля для обучения онтологии предполагает использование нейросетевой языковой модели. При обучении на наборе данных, подготовленном с применением лексического модуля, возрастает эффективность работы по критериям полноты и точности обнаружения новых понятий онтологии.

Список источников

[1] ***Wong W., Liu W., Bennamoun M.*** Ontology learning from text: A look back and into the future // ACM Comput. Surv. 2012. Vol.44(4). P.1-36. DOI:10.1145/2333112.2333115.

[2] ***Фреге Г.*** Смысл и денотат. Пер. с нем. Е.Э. Разлоговой // Семиотика и информатика. 1977. Вып.8. С.181–210.

[3] ***Cimiano P. et al.*** Linguistic linked open data cloud // Linguistic linked data: Representation, generation and applications. Cham: Springer International Publishing, 2020. P.29–41.

[4] ***Declerck T., Siegel M., Gromann D.*** OntoLex-Lemon as a possible bridge between WordNets and full lexical descriptions. Conference: 10th Global WordNet Conference (GWC)At: Wroclaw, Poland. 2019. 7 p.

[5] ***Chen L.J., Hoon G.K.*** Feature Expansion using Lexical Ontology for Opinion Type Detection in Tourism Reviews Domain: 8 // International Journal of Advanced Computer Science and Applications (IJACSA). The Science and Information (SAI) Organization Limited, 2020. Vol.11(8). DOI:10.14569/IJACSA.2020.0110877.

[6] Lexical Ontology Layer – A Bridge between Text and Concepts | SpringerLink. https://link.springer.com/chapter/10.1007/978-3-642-34624-8_20.

[7] ***Badra F.*** Ontology and Lexicon: The Missing Link // 9th International Conference on Terminology and Artificial Intelligence, TIA 2011, Paris, 10 November 2011. P.16–18.

[8] ***Klimek B. et al.*** Challenges for the representation of morphology in ontology lexicons. Proceedings of eLex 2019. P.570-591.

[9] ***Walter S., Unger C., Cimiano P.*** M-ATOLL: A framework for the lexicalization of ontologies in multiple languages // The semantic web – ISWC 2014 / ed. Mika P. et al. Cham: Springer International Publishing, 2014. P.472–486. DOI:10.1007/978-3-319-11964-9_30.

[10] ***Honnibal M., Montani I.*** spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. 2017.

[11] ***Lomov P., Malozemova M., Shishaev M.*** Training and Application of Neural-Network Language Model for Ontology Population // Software Engineering Perspectives in Intelligent Systems / ed. Silhavy R., Silhavy P., Prokopova Z. Cham: Springer International Publishing, 2020. P.919–926. DOI:10.1007/978-3-030-63319-6_85.

[12] Корпус русскоязычных статей Интернет-издания Lenta.ru: https://kaggle.com/yutkin/corpus-of-russian-news-articles-from-lenta.

[13] ***Tedeschi S. et al.*** Named Entity Recognition for Entity Linking: What Works and What's Next // Findings of the Association for Computational Linguistics: EMNLP 2021. Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021. P.2584–2596. DOI:10.18653/v1/2021.findings-emnlp.220.

[14] ***Bhattacharya I., Getoor L.*** Entity Resolution // Encyclopedia of Machine Learning / ed. Sammut C., Webb G.I. Boston, MA: Springer US, 2010. P. 321–326.

[15] ***Barlaug N., Gulla J.A.*** Neural Networks for Entity Matching: A Survey // ACM Trans. Knowl. Discov. Data. 2021. Vol.15(3). P.52:1-52:37.

[16] ***Gottapu R.D., Dagli C., Ali B.*** Entity Resolution Using Convolutional Neural Network // Procedia Computer Science. 2016. Vol.95. P.153–158. DOI:10.1016/j.procs.2016.09.306.

[17] ***Hadi M.U. et al.*** Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects. 2023. DOI:10.36227/techrxiv.23589741.

[18] ***Lv K. et al.*** Full Parameter Fine-tuning for Large Language Models with Limited Resources: arXiv:2306.09782. arXiv, 2023. DOI: 10.48550/arXiv.2306.09782.

[19] ***Shin T. et al.*** AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts: arXiv:2010.15980. arXiv, 2020. DOI:10.18653/v1/2020.emnlp-main.346.

[20] ***Li F., Hogg D.C., Cohn A.G.*** Ontology Knowledge-enhanced In-Context Learning for Action-Effect Prediction: Proceedings Paper // Advances in Cognitive Systems. Arlington, Virginia: ACS-2022, 2022.

[21] ***Zouhar V. et al.*** A Formal Perspective on Byte-Pair Encoding: arXiv:2306.16837. arXiv, 2023.

Сведения об авторе

Ломов Павел Андреевич, 1984 г. рождения, к.т.н., старший научный сотрудник Института информатики и математического моделирования имени В.А. Путилова Кольского научного центра Российской академии наук. Области научных интересов: представление знаний, онтологическое моделирование, семантическая сеть. AuthorID (РИНЦ): 8479-8320. Author ID (Scopus): 55350587100; ORCID: 0000-0002-0924-0188; Researcher ID (WoS): P-6627-2015. *lomov@iimm.ru.*

Поступила в редакцию 28.09.2023, после рецензирования 22.10.2023. Принята к публикации 01.11.2023.

Scientific article DOI: 10.18287/2223-9537-2023-13-4-520-530

Formation of the lexical module of applied ontology for its learning

© 2023, P.A. Lomov

Kola Science Centre of the Russian Academy of Science,
Institute for Informatics and Mathematical Modeling named after V.A. Putilov, Apatity, Russia

Abstract

The complexity of developing ontologies is associated with their learning, which involves the automation of solving problems related to the extraction of ontological concepts and relationships from natural language texts. For a created ontology used in an information system, there will most likely be a need to expand its conceptual system due to the complication or change in data processing. Subsequent application of text analysis to discover new concepts and incorporate them into the ontology, that is, linking them with existing concepts, will require identifying the latter in sentences of natural language texts. This paper considers the problem of automating the formation of the lexical module of applied ontology, which includes formalized representations of ontological concepts in natural language texts. A brief review of existing works devoted to the use of lexical information about ontology components in solving problems related to the analysis of textual data is presented. The ways of using the OntoLex-Lemon model for determining the structure of lexical representations of ontology concepts are considered. A procedure is proposed for the formation of lexical representations based on the analysis of texts in the subject area, taking into account the case when concepts have names consisting of several words. The results of applying the obtained module for the automatic formation of a training set of a neural network language model used in the ontology learning task for discovering new concepts in the corpus of subject texts are presented.

Key words: *lexical module, applied ontology, OntoLex-Lemon, ontology learning.*

For citation: *Lomov PA.* Formation of the lexical module of applied ontology for its learning [In Russian]. *Ontology of designing*. 2023; 13(4): 520-530. DOI:10.18287/2223-9537-2023-13-4-520-530.

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

List of figures

References

[1] ***Wong W, Liu W, Bennamoun M.*** Ontology learning from text: A look back and into the future // ACM Comput. Surv. 2012; 44(4): 1-36. DOI:10.1145/2333112.2333115.

[2] ***Frege G.*** Meaning and denotation [In Rusian. Translation from German by E.E. Razlogova]. Semiotics and computer science.1977; 8: 181–210.
[3] ***Cimiano P. et al.*** Linguistic linked open data cloud // Linguistic linked data: Representation, generation and applications. Cham: Springer International Publishing, 2020. P.29–41.
[4] ***Declerck T, Siegel M, Gromann D.*** OntoLex-Lemon as a possible bridge between WordNets and full lexical descriptions. 2019.
[5] ***Chen LJ, Hoon GK.*** Feature Expansion using Lexical Ontology for Opinion Type Detection in Tourism Reviews Domain: 8 // International Journal of Advanced Computer Science and Applications (IJACSA). The Science and Information (SAI) Organization Limited, 2020; 11(8). DOI:10.14569/IJACSA.2020.0110877.
[6] Lexical Ontology Layer – A Bridge between Text and Concepts | SpringerLink [Electronic resource]. https://link.springer.com/chapter/10.1007/978-3-642-34624-8_20 (accessed: 15.10.2023).
[7] ***Badra F.*** Ontology and Lexicon: The Missing Link // 9th International Conference on 2011.
[8] ***Klimek B. et al.*** Challenges for the representation of morphology in ontology lexicons. 2019.
[9] ***Walter S, Unger C, Cimiano P.*** M-ATOLL: A framework for the lexicalization of ontologies in multiple languages // The semantic web – ISWC 2014 / ed. Mika P. et al. Cham: Springer International Publishing, 2014. P.472–486.
[10] ***Honnibal M., Montani I.*** spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. 2017.
[11] ***Lomov P, Malozemova M, Shishaev M.*** Training and Application of Neural-Network Language Model for Ontology Population // Software Engineering Perspectives in Intelligent Systems / ed. Silhavy R., Silhavy P., Prokopova Z. Cham: Springer International Publishing, 2020. P. 919–926. DOI:10.1007/978-3-030-63319-6_85.
[12] Corpus of Russian-language articles of the online publication Lenta.ru - https://kaggle.com/yutkin/corpus-of-russian-news-articles-from-lenta (date of access: 05/12/2022).
[13] ***Tedeschi S. et al.*** Named Entity Recognition for Entity Linking: What Works and What's Next // Findings of the Association for Computational Linguistics: EMNLP 2021. Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021. P. 2584–2596. DOI:10.18653/v1/2021.findings-emnlp.220.
[14] ***Bhattacharya I, Getoor L.*** Entity Resolution // Encyclopedia of Machine Learning / ed. Sammut C., Webb G.I. Boston, MA: Springer US, 2010. P. 321–326.
[15] ***Barlaug N, Gulla JA.*** Neural Networks for Entity Matching: A Survey // ACM Trans. Knowl. Discov. Data. 2021; 15(3): 52:1-52:37.
[16] ***Gottapu RD, Dagli C, Ali B.*** Entity Resolution Using Convolutional Neural Network // Procedia Computer Science. 2016; 95: 153–158. DOI:10.1016/j.procs.2016.09.306.
[17] ***Hadi MU. et al.*** Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects. 2023. DOI:10.36227/techrxiv.23589741.
[18] ***Lv K. et al.*** Full Parameter Fine-tuning for Large Language Models with Limited Resources: arXiv:2306.09782. arXiv, 2023. DOI: 10.48550/arXiv.2306.09782.
[19] ***Shin T. et al.*** AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts: arXiv:2010.15980. arXiv, 2020. DOI:10.18653/v1/2020.emnlp-main.346.
[20] ***Li F, Hogg DC, Cohn AG.*** Ontology Knowledge-enhanced In-Context Learning for Action-Effect Prediction: Proceedings Paper // Advances in Cognitive Systems. Arlington, Virginia: ACS-2022, 2022. https://advancesincognitivesystems.github.io/acs2022/papers/ (accessed: 21.10.2023).
[21] ***Zouhar V. et al.*** A Formal Perspective on Byte-Pair Encoding: arXiv:2306.16837. arXiv, 2023.

About the author

Pavel Andreevich Lomov (b.1984), PhD, a senior researcher of Institute for Informatics and Mathematical Modeling named after V.A. Putilov – Subdivision of Kola Science Centre of the Russian Academy of Sciences. The research interests include knowledge representation, ontological modeling, and semantic web. AuthorID (RSCI): 8479-8320. Author ID (Scopus): 55350587100; ORCID: 0000-0002-0924-0188; Researcher ID (WoS): P-6627-2015. *lomov@iimm.ru.*

Received September 28, 2023, Revised October 22, 2023. Accepted November 01, 2023.