Семантические технологии для семантических приложений. Часть 1. Базовые компоненты семантических технологий*

В.И. Городецкий [I], О.Н. Тушканова [II,III]

[I] TRA Robotics Ltd., г. Санкт-Петербург, Россия
[II] Санкт-Петербургский институт информатики и автоматизации РАН, г. Санкт-Петербург, Россия
[III] Санкт-Петербургский политехнический университет Петра Великого, г. Санкт-Петербург, Россия

Аннотация. В работе обсуждаются базовые аспекты современного понимания семантических вычислений, семантических технологий и семантических приложений в искусственном интеллекте. Приводится базовая терминология, используемая в работе, и даются конкретные примеры семантических приложений, в том числе, индустриального уровня. Показано, что базовыми компонентами семантических технологий искусственного интеллекта являются онтологии и семантические модели их использования, семантические ресурсы, которые содержат знания о семантике слов и других сущностей естественного языка и средства уточнения этой семантики, а также семантическая компонента технологии, которая используется для формального описания смысла сущностей естественного языка и численной оценки их попарной семантической близости. Обсуждаются доступные семантические ресурсы, и дается их сравнительный анализ. Приводятся сведения о типах сущностей естественного языка (примитивах), которые далее практически используются для построения моделей формального описания смысла текстов в различных семантических приложениях. Последние компоненты описания семантики текстов составляют содержание второй части данной работы.

Ключевые слова: семантика естественного языка, семантические технологии, семантические вычисления, семантическое приложение, онтология, семантический ресурс.

DOI 10.14357/20718594180406

Введение

На современном этапе развития методов и средств искусственного интеллекта, в частности, методов и средств обработки больших данных, отмечается значительное повышение интереса исследователей и разработчиков к моделям и методам работы с данными в семантическом стиле, другими словами, – на естественном языке (ЕЯ). Анализ показывает, что к настоящему времени накоплен достаточный научный потенциал теоретических знаний и алгоритмов работы с семантикой ЕЯ-текстов, который совместно с достижениями в области получения, преставления и использования знаний, в частности, в области инжиниринга и практического использования онтологий, позволяет сделать качественный скачок в области алгоритмов работы со смыслом ЕЯ уже в сугубо прикладных задачах искусственного интеллекта и использовать эти достижения на индустриальном уровне. Реальность вычислений в семантическом стиле подтверждается материалами сборника [1], в котором анализируется состояние исследований и индустриальных разработок в этой области.

*Работа выполнена при поддержке программы Президиума РАН №26 "Фундаментальные основы создания алгоритмов и программного обеспечения для перспективных сверхвысокопроизводительных вычислений".

✉ *Тушканова Ольга Николаевна. E-mail: tushkanova.on@gmail*

Семантические вычисления стали возможными благодаря успехам в трех областях, а именно:

– в области методов и средств *практического инжиниринга онтологий* как структур для интеграции и совместного представления гетерогенных распределенных *знаний* и *данных*, которые делают знания и данные с ЕЯ-семантикой одинаково доступными как человеку, так и компьютеру;

– в области использования веб-ресурсов, лавинообразный рост объема и разнообразия которых превратил их практически в универсальный источник знаний о смысле слов, понятий и других сущностей ЕЯ. Важно отметить, что эти ресурсы стали доступными в реальном времени благодаря мощным поисковым средствам типа поисковой машины Google. Важно также, что веб-ресурсы включают в себя не только веб-страницы (их число в настоящее время оценивается величиной порядка 10 млрд), но и другие разнообразные ресурсы различной степени структуризации типа Википедии, семантического веб, веба связанных массивов данных (англ. *Linked Data Web*) и других не менее значимых веб-ресурсов;

– в области методов, алгоритмов и инструментальных средства обработки больших объемов текстовых данных, представленных на ЕЯ, и слабоструктурированных данных, представленных в таких форматах, как XML, Json и др. Благодаря этому большие данные, в состав которых входят и ресурсы веб, стали уникальным источником ранее недоступных научных, инженерных и технологических знаний, что, в свою очередь, стимулировало развитие семантических технологий.

Перечисленные достижения делают реальным использование методов и средств выявления и формального представления семантики различных сущностей ЕЯ (слов, понятий, фраз, искусственно создаваемых паттернов, текстов в целом) и семантических вычислений с ними в приложениях искусственного интеллекта промышленного масштаба.

В данной работе семантика ЕЯ рассматривается в ограниченном понимании. Она ограничена *прикладными задачами искусственного интеллекта*, в которых обычно не решаются задачи глубокого анализа семантики естественного языка, например, для перевода ЕЯ-текстов с одного языка на другой и т.п., а, в основном – обработки данных, представленных текстами, возможно, совместно с другими параметрами. Среди них в искусственном интеллекте рассматриваются два основных типа задач семантической обработки данных, а именно, задачи семантической структуризации (кластеризации) больших объемов текстовых документов (множеств текстов, которые специалисты по большим данным часто образно называют «озером данных»), возможно, с последующим тематическим (смысловым) анализом, реферированием или индексированием кластеров. А также задачи семантической классификации текстов, в которых центральной проблемой является классификация новых текстовых документов, не входивших в обучающую выборку, т.е. отнесение их к наиболее близкому по смыслу (семантике) классу (кластеру). В этих приложениях пока онтология практически не имеет конкурентов при описании семантики сущностей ЕЯ (слов, понятий, текстовых паттернов и текстов в целом), поэтому в данной работе основное внимание уделяется семантическим моделям, которые опираются на онтологический подход к описанию смысла сущностей ЕЯ.

Рис. 1 дает представление о составе компонент семантических технологий и все они представляются одинаково важными. Однако в рамках одной работы невозможно сколько-нибудь полно описать все эти компоненты и варианты подходов к их построению. Поэтому далее основное внимание уделяется обзору семантических ресурсов и типовых вариантов формальных моделей семантики ЕЯ-сущностей.

В Разделе 1 уточняется используемая терминология. В Разделе 2 кратко описываются примеры семантических приложений, разработанных к настоящему времени. В Разделе 3 дается характеристика семантических ресурсов, доступных разработчиками приложений для выявления семантики ЕЯ-сущностей, а также оценки их семантической близости. В Разделе 4 описываются варианты семантических сущностей ЕЯ, которые используются в качестве примитивов для описания смысла текста в векторном пространстве (далее – семантических примитивов, для краткости). Семантические ресурсы и примитивы текста формируют базис для дальнейшего построения моделей смысла сущностей ЕЯ и численной оценки их попарной семантической близости, существенных для решения задач кластерного анализа и ма-

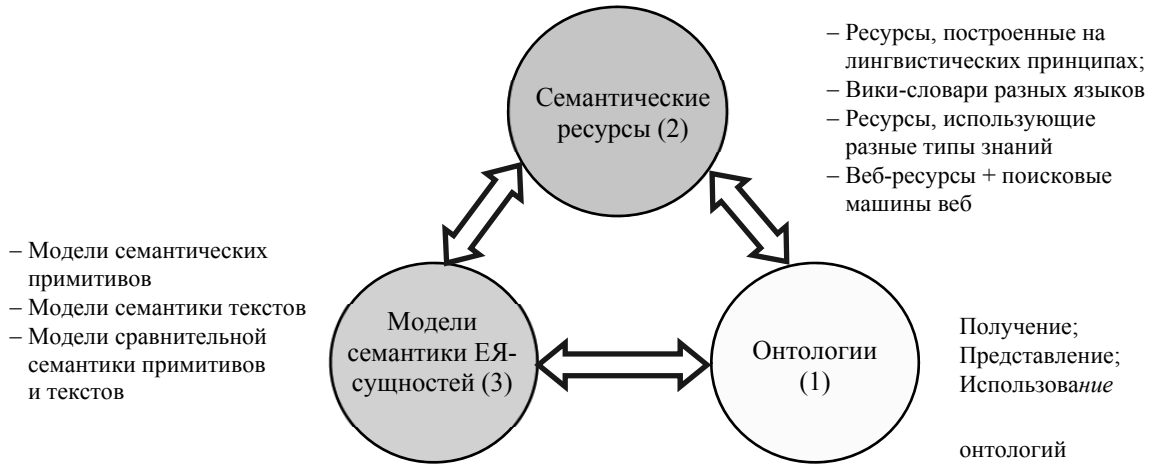

Рис. 1. “Три кита” семантических технологий, вычислений и приложений

шинного обучения классификации текстов по их смыслу. Последние модели составляют содержание второй части данной работы. В заключении обзора в первой части работы формулируются ее результаты и указываются проблемы, сдерживающие развитие семантических технологий и их практическое использование в приложениях.

1. Используемая терминология

Содержание всей используемой и вводимой далее терминологии описывается на интуитивном уровне и основывается на “здравом смысле”, а потому их не следует рассматривать в качестве определений. Естественно, что терминология не претендуют на строгость, как и любые сущности, ЕЯ, а только очерчивают смысл, который вкладывается в эту терминологию в рамках данной работы. Материал данного раздела существенно базируется на работе [1].

Семантика ЕЯ как научное направление имеет целью выявить и представить формально *отношения* между сущностями ЕЯ (символами, словами, понятиями, группами слов, паттернами текстов и текстами в целом) и их смыслом, т.е. значением, которое они несут в себе для человека-носителя ЕЯ. Все сущности ЕЯ имеют неточный, а зачастую – многозначный смысл. *Семантика сущности ЕЯ* определяется (уточняется) в таких случаях контекстом, который позволяет установить ее смысл с помощью других слов, синтаксически и семантически связанных с нею.

Формальное описание/представление семантики сущностей ЕЯ выполняется, как правило, с использованием онтологии, которая в современной литературе рассматривается как средство установления *единого, одинакового понимания* всеми (пользователями и компьютерами) смысла сущности ЕЯ и/или текста в целом [2]. Смысл сущности в конкретном тексте обычно устанавливается с точностью до множества синонимов, а варианты его интерпретации иллюстрируются *примерами* этой сущности, которые являются частью онтологии. В онтологии их иногда называют также *фактами*.

Под термином *семантические вычисления* далее понимается комплекс методов, алгоритмов и программ обработки данных, включая текстовые данные, в которых однозначно задается и интерпретируется ЕЯ-смысл исходных, промежуточных и конечных данных вычислений. Слова *смысл* или *понимание* ЕЯ-сущностей подразумевают обычно их понимание человеком.

С другой стороны, онтология тоже оперирует только понятиями ЕЯ, представленными в человеко- и машино-понятной формах. При этом для вычислений с понятиями онтологии должны автоматически решаться конфликты, связанные с потенциальной неоднозначностью их смысла. Одним из обычных средств разрешения таких конфликтов является использование примеров понятия. Другой вариант – это выявление смысла понятия с помощью специальных алгоритмов анализа контекста, которые называют WSD-алгоритмами (от англ. *Word Sense Disambiguation*).

Соответственно, *семантическими приложениями* будем называть приложения, в которых компоненты формальных моделей на концептуальном уровне описаны множеством понятий ЕЯ и отношений на них, а процессы преобразования данных модели выполняются с помощью семантических вычислений. В работе [1, с. 4] класс приложений, которые называются семантическими, определяется более кратко: "Семантическое приложение есть программное приложение, которое явно или неявно использует семантику предметной области" (англ. A *semantic application* is a software application which explicitly or implicitly uses the semantics of a domain). В качестве одного из примеров семантического приложения в работе [1, с. 4] используется семантический поиск, в котором синонимы и относящиеся к нему другие слова используются для обогащения результатов простого поиска по запросу, представленному текстом. Некоторые другие характерные примеры семантических приложений рассматриваются ниже.

Центральной компонентой любого семантического приложения искусственного интеллекта является онтология, представляющая ЕЯ-семантику задач приложения, и именно она задает смысловую интерпретацию компонент модели, а также ее исходных, промежуточных и конечных результатов.

Во всех задачах семантических вычислений, например, в задачах семантической кластеризации или классификации речь идет о ЕЯ-семантике, представленной в человеко-понятных ЕЯ-терминах. Поэтому методы установления семантики сущностей текста, их сравнения по семантическим мерам и метрикам близости, интерпретация результатов семантических вычислений должны базироваться на адекватных источниках *семантических знаний*, явно или неявно описывающих ЕЯ-семантику языка. Эти ресурсы принято называть семантическими.

Для пояснения существа введенной терминологии, в частности, для пояснения сущности семантических вычисления далее рассматриваются варианты соответствующих им прикладных систем.

2. Семантические приложения. Примеры

В соответствии с терминологией, принятой в данной работе, приложение называется семантическим, если в нем пользователь общается с программой приложения с помощью запросов, содержащих понятия и отношения ЕЯ, программа приложения общается с пользователем аналогичным образом, а пользовательские и программные интерфейсы построены так, что перевод с человеко-понятного языка на машино-понятный и обратно для исходных данных, заданий, промежуточных и конечных результатов выполняются автоматически.

Как уже отмечалось, в данной работе рассматриваются семантические технологии, в основе которых лежат онтологии, выполняющие интеграцию данных и знаний, а также обеспечивающие интерфейсы пользователя и программ к онтологиям. Другими словами, пользователь и программа семантического приложения общаются, главным образом, через онтологию. Кроме того, онтология в таких приложениях выступает в роли метамодели данных и знаний, которые могут иметь различную структуру (или вообще не иметь структуры) и хранится в распределенных базах данных. Таким образом, онтология семантического приложения решает две главные роли - роль семантического интегратора знаний и данных и роль посредника между пользователями и программами приложения.

Традиционная интеграция данных включает извлечение, трансформацию и загрузку данных (англ. *Extraction*, *Transformation*, and *Loading*, ETL). Семантическая интеграция добавляет к этим процедурам ЕЯ-интерпретацию данных, знаний и результатов вычислений для пользователя и их трансформацию в машино-понимаемую форму и обратно. В настоящее время семантические приложения, главным образом, рассматриваются как средство семантической интеграции данных. Особенно актуальна эта функция при хранении и обработке больших данных, причем здесь речь идет не только о представленных текстами ЕЯ, но и о слабоструктурированных данных типа XML, JSON и др. Характерной чертой большинства платформ обработки и хранения больших данных является недостаточная полнота или полное отсутствие их схем. Нехватка семантической информации, описывающей содержимое таких хранилищ, создает трудности для эффективной обработки данных и использования ее результатов. Практически всегда большие данные генерируются и используются различными

приложениями или группами пользователей, при этом в интересах разных пользователей они и обрабатываются. Роль семантических метаданных и семантических вычислений становится ключевой при проектировании, наполнении и поддержке "озер данных", где семантические модели и вычисления потенциально позволяют обеспечить прозрачность, согласованность и лучшее понимание данных "озера".

Среди множества конкретных приложений, в которых была успешно использована семантическая интеграция данных, можно упомянуть, например, разработку системы, осуществляющей сбор текстовых документов о медицинских устройствах, удовлетворяющих законодательству, проверку актуальности этих документов и их погружение в индексированный репозиторий [3]. В [4] представлена семантическая система управления контентом для веб-портала Leipzig Health Atlas, который содержит только верифицированный контент из области здравоохранения. В [5] рассмотрено семантическое приложение, предназначенное для создания архивов культурного наследия с помощью сопоставления словарей различных культурных проектов.

Важной областью применения семантических вычислений является аннотирование контента. Аннотации, которые чаще представляют собой теги или понятия онтологии, связанные с экземплярами данных, позволяют упростить поиск информации, поскольку "индексируют" семантику, "спрятанную" в данных. Онтологии и графы знаний могут использоваться как для обогащения тегов и разметки, выполненной пользователем вручную, так и в алгоритмах автоматического аннотирования. Сам процесс обогащения разметки фактически формирует новый источник знаний, которые в дальнейшем могут использоваться для решения задач, например, тематического моделирования.

В предыдущем разделе в качестве типового примера семантического приложения упоминался семантический поиск. Сам термин "семантический поиск" неоднозначен, и в общем случае он может подразумевать технологию поиска информации, которая использует любые модели семантики, например, таксономии, тезаурусы, онтологии или графы знаний. В более узком смысле семантический поиск должен справляться не только с запросами на формальном языке типа SPARQL или GraphQL, или с запросами специального вида от технических экспертов, но и с более размытыми формулировками непрограммирующих пользователей, включающими, например, только ключевые слова, которые не обязательно встречаются в анализируемых документах, однако имеют синонимы, относящиеся к интересующей пользователя тематике. Заметим, что в работе [6] в качестве компоненты семантического приложения рассматривается программа, которая помогает пользователю сформулировать запрос к веб-ресурсам более точно, предлагая ему несколько возможных вариантов.

В работе [7] описана семантическая технология сопоставления ключевых слов с фразами из запросов пользователей и добавления синонимов, акронимов и других слов, отсутствующих в анализируемых текстах, но семантически близких к ним (англ. *hidden terms*), с понятиями доменной онтологии приложения для обогащения результатов поиска. Авторы отмечают, что эта технология также может быть использована для создания семантического интерфейса более понятного пользователю.

В [8] показано как использование контекстной информации, структурированной с помощью онтологии, позволяет возвращать пользователю информацию в нужное время и в нужном месте, и таким образом улучшать результаты семантического поиска.

Обратим внимание на то, что во всех приведенных примерах созданных семантических приложений используется только семантическая технология интеграции распределенных знаний и данных. И это отражает реально слабое состояние исследований в области семантических интерфейсов онтологии, необходимых для работы с ней в более сложных случаях использования. Однако эта очень важная задача находится вне темы данной работы и потому не анализируется, хотя и заслуживает отдельного рассмотрения.

3.Семантические ресурсы

Понятие семантики текста сложно определить сколько-нибудь точно, а семантику его сущностей и их семантическую близость непросто измерить численно. Но потенциальное качество решения всех таких задач определяется тем источником знаний о семантике ЕЯ, ко-

торый положен в основу вычисления этих характеристик.

Известно, что различные источники знаний обладают разными семантическими ресурсами, и, соответственно, потенциально разными возможностями при решении задач семантических вычислений. Эти возможности определяются двумя главными факторами: мощностью множества понятий и их примеров, явно или неявно представленных в источнике знаний, и детальностью их описания: множеством отношений между понятиями, которые в них определены. Среди ресурсов, доступных в настоящее время, авторы работы [9] различают:

1. Ресурсы, которые построены на лингвистических принципах (англ. *linguistically constructed*). Примерами являются семантические словари типа WordNet в английском [10], европейском [11] и русском [12, 13] вариантах. Эти ресурсы содержат понятия, множество их синонимов (англ. *synsets*) и ряд семантических отношений, а также текстовые описания понятий и другие.

2. Вики-словари для различных языков (англ. *Wiktionary*), которые представляют собой многофункциональные словари и тезаурусы. Словарные статьи в них содержат информацию о синтаксических и морфологических свойствах слова и информацию о его семантических свойствах, в частности, описывают его значение, представляют список синонимов, антонимов, гиперонимов (более общих), гипонимов (более специализированных) понятий, а также различные родственные слова. К ресурсам такого же уровня информативности относят предметно-ориентированные онтологии. Ресурсы Вики-словарей и онтологий с их информационными и структурными отношениями намного богаче чисто лингвистических ресурсов типа WordNet.

3. Ресурсы, которые построены на совместном использовании разных типов знаний (англ. *collaborative resources*). Примером таких ресурсов является Wikipedia с ее инструментами типа DBpedia и ее варианты на различных языках, включая русский. Этот ресурс содержит словарные статьи (как и в Вики-словарях) для категорий понятий и их синонимов со ссылками на более общие и более специализированные понятия, тексты для решения WDA-проблемы, а также ряд информационных и структурных отношений. Заметим, что словарный запас Wikipedia намного богаче, чем у всех выше названных ресурсов, и, что еще важнее, он постоянно пополняется.

4. Ресурсы веб в целом совместно с множеством поисковых машин, например, Google, Яндекс и др. Эти ресурсы являются наиболее богатыми и весьма разнообразными, например, они включают в себя Википедию, все множество веб-страниц, ресурсы семантического веба, примерами которых являются данные *Linked Open Data*, представленные в облаке. Использование поисковых машин дает возможность получить в явной форме те знания, которые различными способами представлены в веб-ресурсах, а это практически все знания человечества. С другой стороны, поисковым машинам доступны только те знания, которые имеются в веб-ресурсах, индексированных этими машинами. Заметим, что обычно поисковые машины активно используют средства собственных онтологий, графов знаний и другие семантические средства поиска, которые скрыты от пользователей. Другими словами, поисковые машины могут обладать дополнительными семантическими ресурсами.

Оригинальный путь формирования фрагмента семантических ресурсов, который нужен для работы с конкретной лексикой, рассмотрен в работе [14]. Для того чтобы отнести новое слово предметной области к некоторому понятию онтологии, авторы предлагают найти семантическую близость этого слова со словами, которые уже связаны с имеющейся онтологией и лексически в ней представлены. Онтология при этом рассматривается как дерево принятия решений, в каждом узле которого решается задача о выборе очередного потомка (более частного понятия), к которому принадлежит новое слово. Для решения этой задачи в каждом узле дерева формируется обучающее множество из слов, которые являются примерами соответствующего понятия онтологии. Далее, для каждого слова из онтологии и нового слова формируется документ *X.txt*, который состоит из списка определений, выдаваемых поисковой машиной Google по запросу define: X. Таким способом формируется дополнительная компонента семантического ресурса, необходимая для определения смысла нового слова в контексте существующей онтологии.

В связи с осознанием важной роли источников семантических знаний о ЕЯ в различных

задачах искусственного интеллекта, можно ожидать появление и других, семантически более богатых ресурсов. В качестве варианта можно рассматривать обогащение веб-ресурсов встроенными средствами автоматического решения WSD-проблемы, а также отношениями семантической близости понятий разного уровня агрегирования. Важно заметить, что технологии больших данных могут сыграть важную роль в обогащении семантических ресурсов, включая веб-ресурсы.

4. Модели семантики естественного языка. Семантические примитивы текстов

Когда говорят о формальных моделях семантики ЕЯ-сущностей, например, отдельного слова, предложения, фрагмента текста или текста в целом, то обычно имеют в виду такие их представления, которые, с одной стороны, однозначны по смыслу, т.е. все компоненты, модели которых определены однозначно с точностью до их синонимов и отделены по смыслу от возможных омонимов. Кроме того, компонентам модели семантики ЕЯ-сущностей текста могут быть поставлены в соответствие значения весовых коэффициентов, определяющих их важность в формировании смысла текста. Однако на практике, к сожалению, приходится иметь дело и с более неопределенными моделями, в которых смысл ЕЯ-сущностей текста не детерминирован однозначно, весовые коэффициенты определены достаточно грубо, отсутствует информация о взаимозависимости смысла различных компонент модели и т.п. Конкретным примером такой грубой модели является рассматриваемая далее Google-семантика ЕЯ-сущностей текста, которая, как правило, не различает смысл омонимов. Что касается “весовых коэффициентов”, то здесь можно говорить только о более или менее удачном их выборе. Обычно удачность выбора модели описания семантики ЕЯ-сущностей текста вместе с числовыми оценками их важности с точки зрения представления смысла можно оценить только по результатам экспериментов в конкретном приложении или с помощью тестовых множеств. Сейчас трудно отдать предпочтение какой-то одной модели семантики, и этим можно объяснить наличие огромного количества предложенных моделей, число которых продолжает увеличиваться.

Однако несмотря на большое разнообразие моделей формализации семантики ЕЯ-текстов и существующих алгоритмов работы с ними, в предложенных моделях можно выделить некоторые базовые семантические единицы разного уровня сложности, которые используются в качестве своего рода *семантических примитивов текстов* (СПТ)[1] для описания их смысла . СПТ могут представлять собой просто слова и отдельные понятия, фрагменты текста или/и результаты некоторых трансформаций фрагментов текста или текста в целом. Эти примитивы обычно интерпретируются как строковые (номинальные) переменные-координаты векторного пространства, которые могут иметь атрибуты, например, весовые коэффициенты, о которых уже шла речь. Эти переменные рассматриваются в качестве признаков, описывающих семантику текста в пространстве примитивов. Такие модели представляют смысл текста точкой в том или ином векторном пространстве. Их называют моделями векторного пространства (англ. *Vector Space Model*, *VSM*) и в семантических приложениях они сейчас используются наиболее часто.

Рассмотрим варианты СПТ для описания семантики текстов, которые в настоящее время в области искусственного интеллекта используются наиболее активно. На начальном этапе развития технологии обработки текстов использовались примитивы, называемые *N-граммами*, под которыми понимали последовательности из *N* слов, извлеченных из текста. Каждой N-грамме ставилось в соответствие значение выборочной оценки некоторой статистики, например, частоты ее встречаемости в тексте или другой оценки. Обычно число N выбиралось в пределах от двух до четырех. Уже для небольшого текста общее число таких N-грамм оказывалось значительным, поэтому подход с их использованием в качестве примитивов приводит к большим размерностям про-

[1] В области лингвистики существует термин “семантический примитив”, описанный А. Вежбицкой в ряде ее работ. Семантические примитивы по А. Вежбицкой – это некоторые слова ЕЯ, при помощи которых можно объяснять значения всех остальных слов ЕЯ, не прибегая к герменевтическому кругу. Предлагаемое в данной работе понятие *семантического примитива текста* относится не к языку в целом, а к конкретному тексту, и оно не имеет отношения к термину А. Вежбицкой.

странства описания текстов и к вычислительно сложным процедурам работы с ними в прикладных задачах.

Похожий подход используется и в настоящее время с той разницей, что длина N-граммы может быть переменной и ее примитивы формируются более сложным образом. За счет активного использования семантических ресурсов они позволяют более точно задать и более компактно представить смысл примитивов и их веса. В качестве веса, как и четверть века тому назад, наибольшей популярностью пользуется мера TF-IDF [15]. Ее популярность объясняется тем, что она дополнительно к оценке частоты встречаемости примитива (эту весовую функцию примитива называют CountBOW), учитывает с меньшим весом те из них, которые встречаются также и в других текстах, акцентируя тем самым внимание на различиях в специфичности примитивов для разных текстов. Это позволяет отфильтровывать служебные и малоинформативные слова (их называют stop-words), а также семантически значимые слова, которые, однако, встречаются в большинстве текстов, а потому неинформативны при решении прикладных задач типа кластеризации и классификации. Иногда используется также мера BM25, предложенная в [16], которая базируется на оценке TF-IDF, но вводит для нее нормировку в виде ограничения сверху на ее значение. Детальное описание этих и других мер оценки семантической важности дано в [15].

В современной терминологии N-граммы чаще называют терминами (англ. *term*) [17, 18]. Однако до настоящего времени понятие термина трактуется неоднозначно и продолжает оставаться предметом дискуссий. Достаточно подробный анализ понятия *термин* в различном его понимании имеется в работе [17]. Подробный обзор алгоритмов автоматического выявления терминов разной структуры из множества текстов на ЕЯ дано в [18]. В этих работах в качестве терминов рассматривают N-граммы, составленные из слов или понятий онтологии. В [18] можно также найти информацию о программных средствах, свободно доступных в Интернете, которые предназначены для извлечения терминов, и их сравнительные оценки. Примером терминов являются также латентные факторы [19], которые позволяют строить более компактные модели семантики текстов.

В современных семантических приложениях используются, главным образом, два основных вида терминов, а именно *понятия* некоторого семантического словаря или онтологии и некоторые “семантически насыщенные” агрегированные примитивы, которые строятся с привлечением тех или иных семантических ресурсов ЕЯ (СПТ-агрегаты). Заметим, что понятия онтологии тоже могут рассматриваться как агрегированные СПТ, поскольку в конкретном контексте каждому понятию ЕЯ отвечает множество его синонимов, примеров и слов с близким смыслом. Эта информация позволяет эффективно решать WSD-проблему, ключевую для семантических вычислений.

Что касается СПТ-агрегатов, то в научной литературе предложено много их различных вариантов, некоторые из которых будут представлены далее. Среди них следует особо выделить частный случай СПТ-агрегатов, который называют лексической цепью.

Концепция лексической цепи была предложена в лингвистике как СПТ, который описывает семантику текста и удобен также для оценки сходства и различия разных документов по семантике. Иногда говорят, что лексические цепи позволяют оценивать меру “связности” (англ. *cohesion*) текстов [20]. *Лексической цепью* называется последовательность терминов (слов, понятий), выбранных из предложения или из текста в целом, в которой слова записываются в порядке их встречаемости в тексте, а соседние и, возможно, другие слова связаны некоторыми бинарными отношениями, имеющими семантическую окраску. Разные авторы предлагают свои трактовки и виды лексической цепи. Они отличаются выбором типа терминов, из которых составляется лексическая цепь, а также набором бинарных отношений, которые представлены в цепи. Лексическая цепь обычно используется как семантический примитив текстов, для компонент которой решена WSD-проблема. Но обычно лексическая цепь содержит в себе достаточно информации, для того чтобы решить эту проблему только на основе содержания самой цепи [21]. По сути, лексическая цепь в сжатой форме интегрирует в себе совместно последовательность терминов, их *контекст* и его *семантическую окраску*, позволяя тем самым однозначно устанавливать смысл многозначных терминов цепи. Один из вариантов построения лексической цепи с параллельным решением WSD-проблемы описан в [21], другой вариант ее построения представлен в [22].

Важно отметить, что от выбора типа СПТ и модели описания семантики документов в целом во многом зависит как качество решения задач семантических вычислений, так и их вычислительная эффективность.

Заключение

Семантические технологии, которые в настоящее время имеют большую перспективу развития, достигли современного уровня зрелости благодаря достижениям в области онтологий, развитию моделей семантики ЕЯ и механизмов ее выявления, формального описания и практического использования, а также благодаря созданию семантических ресурсов ЕЯ (Рис. 1). При этом, современный уровень развития семантических ресурсов ЕЯ во многом определяют ресурсы веб, включая его компоненты, структурированные человеком, типа Википедии и семантического веб и др., а также семантические механизмы поиска информации в веб-ресурсах типа поисковой машины Google.

По-видимому, основные препятствия широкого практического использования семантических технологий обусловлены все-таки недостаточной структуризацией ресурсов веб для того, чтобы они могли служить полноценным и легко доступным источником информации для более точного установления семантики обрабатываемых текстов и установления различного рода семантических отношений между разными сущностями ЕЯ. Например, в структурированных веб-ресурсах семантические отношения на множестве сущностей ЕЯ представлены все-таки достаточно бедно. Например, Википедия, которая является наиболее богатым источником знаний о семантике ЕЯ-сущностей, содержит информацию об отношениях понятий в таксономии, о синонимии, приводит антонимы, паронимы и омонимы понятий. Однако эти отношения пока не в состоянии представить все оттенки семантических отношений между ЕЯ-сущностями. Представляется, что обогащение и расширение семантических ресурсов ЕЯ является важной исследовательской проблемой в области семантических технологий. Проект Linked Data Web предназначен именно для обогащения и расширения семантических ресурсов ЕЯ, и прежде всего, в части заранее подготовленных паттернов онтологий.

Другая важная проблема – это поиск более информативных форм СПТ и эффективных алгоритмов их вычисления. Представляется, что одной из наиболее перспективных моделей СПТ является модель лексических цепей, которая способна представить некоторую “агрегированную единицу смысла” текста (ее называют темой текста) в виде графа отношений на множестве лексических компонент обрабатываемого текста. Однако в предложенных вариантах структур лексических цепей (см., например [21, 22]) они, так или иначе, построены на эвристиках и достаточно сложны с вычислительной точки зрения.

Онтология как неотъемлемая часть семантических технологий выдвигает свои проблемы, в частности, в построении механизмов поиска информации по ЕЯ-запросам и механизмов выдачи результатов вычислений также на ЕЯ. Однако это тема отдельного глубокого исследования, которое выходит за рамки данной работы.

Литература

1. Bartussek W., Bense H., Hoppe T., Humm B.G., Reibold A., Schade U., Siegel M., Walsh P. Introduction to Semantic Applications. In Thomas Hoppe, Bernhard Humm, Anatol Reibold (Eds.). Semantic Applications. Methodology, Technology, Corporate Use. Springer-Verlag GmbH Germany, part of Springer Nature 2018.
2. Gruber T.R. A translation approach to portable ontologies. Knowledge Acquisition. Academic Press, 1993, 5(2), pp. 199-220.
3. Uciteli A., Goller C., Burek P., Siemoleit S., Faria B., Galanzina H., Weiland T., Drechsler-Hake D., Bartussek W., Herre H.. Search ontology, a new approach towards semantic search. In: Plödereder E, Grunske L, Schneider E, Ull D (eds) FoRESEE: Future Search Engines, 2014, 44. Annual meeting of the GI, Stuttgart – GI edition proceedings LNI. Köllen, Bonn, pp. 667-672.
4. Beger C, Uciteli A., Herre H. Light-weighted automatic import of standardized ontologies into the content management system Drupal. Stud Health Technol Inform 243, 2017, pp. 170-174.
5. Diwisch K., Engel F., Watkins J., Hemmje M. Managing Cultural Assets: Challenges for Implementing Typical Cultural Heritage Archive’s Usage Scenarios. In: Hoppe T., Humm B., Reibold A. (eds) Semantic Applications. Springer Vieweg, Berlin, Heidelberg, 2018, pp. 219-229.
6. Bollegala D., Yutaka Matsuo Y., Ishizuka M. WebSim: a Web-based Semantic Similarity Measure. The 21st Annual Conference of the Japanese Society for Artificial Intelligence, 2007, pp. 1-4.
7. Humm B.G., Ossanloo H. A semantic search engine for software components. IADIS Press, Mannheim, 2016, pp. 127–135.

8. Fillies C., Weichhardt F., Straub H. The Semantic Process Filter Bubble. In: Hoppe T., Humm B., Reibold A. (eds) Semantic Applications. Springer Vieweg, Berlin, Heidelberg, 2018, pp. 231-241.
9. Feng Y., Bagheri E., Ensan F., Jovanovic J. The state of the art in semantic relatedness: a framework for comparison. Knowledge Engineering Review, 2017, pp. 1-30.
10. Miller A., R. Beckwith R., Fellbaum C.D., Gross D., Miller K. WordNet: An online lexical database. International Journal Lexicograph. 1990, 3, pp. 235-244.
11. Европейский WordNet. http://www.illc.uva.nl/EuroWordNet.
12. Азарова И.В., Митрофанова О.А., Синопальникова А.А. Компьютерный тезаурус русского языка типа WordNet. http://project.phil.spbu.ru/RussNet/papers/RussNet_dialog2003.pdf
13. Сухоногов А.М., Яблонский С.А. Разработка русского WordNet. http://www.interface.ru/iarticle/files/36053_78003690.pdf.
14. Gupta A., Oates T. Using Ontologies and the Web to Learn Lexical Semantics. Proceedings of the 20th international joint conference on Artificial intelligence, January 06-12, 2007, Hyderabad, India, pp. 1618-1623.
15. Пархоменко П.А., Григорьев А.А., Астраханцев Н.А. Обзор и экспериментальное сравнение методов кластеризации текстов. Труды ИСП РАН, 2017, том 29, выпуск 2, с. 161-200.
16. Slonim N., Friedman N., Tishbyi N. Document Clustering Using Word Clusters via the Information Bottleneck Method. Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM. 2000, pp. 208-215.
17. Астраханцев Н.А. Методы и программные средства извлечения терминов из коллекции текстовых документов предметной области. Диссертация на соискание ученой степени кандидата физико-математических наук. Институт системного программирования РАН, Москва, 2014.
18. Kosa V., Chaves-Fraga D., Naumenko D., Yuschenko E., Badenes-Olmedo C., Ermolayev V., Birukou A. Cross-Evaluation of Automated Term Extraction Tools. Technical Report TS-RTDC-TR-2017-1, 30.09.2017. Dept of Comp. Science, Zaporizhzhia National University, Ukraine, 61 p.
19. Deerwester S., Dumais S.T., Furnas G.W., et al. Indexing by latent semantic analysis. Journal of the American Society for Information Science, volume 41, issue 6, 1990, pp. 391-407.
20. Morris J., Hirst G. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. Computational Linguistics, 1991, vol. 17, 1, pp. 21-43.
21. Ткач С.С. Применение лексических цепочек для разрешения лексической многозначности на основе Русского Викисловаря. Магистерская диссертация. Петрозаводский Государственный университет. Петрозаводск, 2016. 60 с.
22. Wei T., Lu Y., Chang H., Zhou Q., Bao X. A semantic approach for text clustering using WordNet and lexical chains. Expert Systems with Applications 42, 2015, pp. 2264-2275.

Semantic technologies for semantic applications. Part 1. Basic components of semantic technologies

V.I. Gorodetsky[I], O.N. Tushkanova [II,III]

[I]TRA Robotics Ltd., St. Petersburg, Russia
[II] St. Petersburg Institute for Informatics and Automation of Russian Academy of Sciences, St. Petersburg, Russia
[III] Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia

The paper discusses the basic aspects of modern understanding of semantic computations, semantic technologies and semantic applications in the field of artificial intelligence. The basic terminology used in the work is introduced, and concrete examples of semantic applications, including the industrial ones, are given. It is shown that the basic components of semantic technologies are ontologies and their semantic models, semantic resources and semantic component. The semantic resources contain knowledges about word semantics and means for refinement of this semantics. The semantic component of the technology is used to formally describe the meaning of NL-entities and numerically evaluate their pairwise semantic similarity. The available semantic resources are discussed and their comparative analysis is given. Information is given on the types of NL-entities (primitives), which are then practically used to build models of text meaning formal description in various semantic applications. The last components of the text semantics description constitute the content of the second part of this paper.

Keywords: natural language semantics, semantic technology, semantic computing, semantic application, ontology, semantic resource.

DOI 10.14357/20718594180406

References

1. Bartussek W., Bense H., Hoppe T., Humm B.G., Reibold A., Schade U., Siegel M., Walsh P. Introduction to Semantic Applications. In Thomas Hoppe, Bernhard Humm, Anatol Reibold (Eds.). Semantic Applications. Methodology, Technology, Corporate Use. Springer-Verlag GmbH Germany, part of Springer Nature 2018.

2. Gruber T.R. A translation approach to portable ontologies. Knowledge Acquisition. Academic Press, 1993, 5(2), pp. 199-220.
3. Uciteli A., Goller C., Burek P., Siemoleit S., Faria B., Galanzina H., Weiland T., Drechsler-Hake D., Bartussek W., Herre H. Search ontology, a new approach towards semantic search. In: Plödereder E, Grunske L, Schneider E, Ull D. (eds) FoRESEE: Future Search Engines, 2014, 44. Annual meeting of the GI, Stuttgart – GI edition proceedings LNI. Köllen, Bonn, pp. 667-672.
4. Beger C, Uciteli A., Herre H. Light-weighted automatic import of standardized ontologies into the content management system Drupal. Stud Health Technol Inform 243, 2017, pp. 170-174.
5. Diwisch K., Engel F., Watkins J., Hemmje M. Managing Cultural Assets: Challenges for Implementing Typical Cultural Heritage Archive's Usage Scenarios. In: Hoppe T., Humm B., Reibold A. (eds) Semantic Applications. Springer Vieweg, Berlin, Heidelberg, 2018, pp. 219-229.
6. Bollegala D., Yutaka Matsuo Y., Ishizuka M. WebSim: A Web-based Semantic Similarity Measure. The 21st Annual Conference of the Japanese Society for Artificial Intelligence, 2007, pp. 1-4.
7. Humm B.G., Ossanloo H. A semantic search engine for software components. IADIS Press, Mannheim, 2016, pp. 127–135.
8. Fillies C., Weichhardt F., Straub H. The Semantic Process Filter Bubble. In: Hoppe T., Humm B., Reibold A. (eds) Semantic Applications. Springer Vieweg, Berlin, Heidelberg, 2018, pp. 231-241.
9. Feng Y., Bagheri E., Ensan F., Jovanovic J. The state of the art in semantic relatedness: a framework for comparison. Knowledge Engineering Review, 2017, pp. 1-30.
10. Miller A., R. Beckwith R., Fellbaum C.D., Gross D., Miller K. WordNet: An online lexical database. International Journal Lexicograph. 1990, 3, pp. 235-244.
11. Europe WordNet. http://www.illc.uva.nl/EuroWordNet.
12. Azarova I.V., Mitrophanova O.A., Sinopolnikova A.A. Komp'yuternyy tezaurus russkogo yazyka tipa WordNet [WordNet-like computer thesaurus for Russian] http://project.phil.spbu.ru/RussNet/ papers/RussNet dialog2003.pdf
13. Sukhonogov A.M., Yablonsky S.A. Razrabotka russkogo WordNet [Russian WordNet development]. http://www.interface.ru/iarticle/files/36053_78003690.pdf.
14. Gupta A., Oates T. Using Ontologies and the Web to Learn Lexical Semantics. Proceedings of the 20th international joint conference on Artificial intelligence, January 06-12, 2007, Hyderabad, India, pp. 1618-1623.
15. Parkhomenko P.A., Grigor'yev A.A., Astrakhantsev N.A. Obzor i eksperimental'noye sravneniye metodov klasterizatsii tekstov [Review and experimental comparison of methods of text clustering]. Trudy ISP RAN [ISP RAS Proceddings]. 2017, 29 (2), pp. 161-200.
16. Slonim N., Friedman N., Tishbyi N. Document Clustering Using Word Clusters via the Information Bottleneck Method. Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM. 2000, pp. 208-215.
17. Astrakhantsev N.A. Metody i programmnyye sredstva izvlecheniya terminov iz kollektsii tekstovykh dokumentov predmetnoy oblasti [Methods and software for extracting terms from the collection of text documents of the domain]. PhD thesis. ISA RAN, Moscow, 2014.
18. Kosa V., Chaves-Fraga D., Naumenko D., Yuschenko E., Badenes-Olmedo C., Ermolayev V., and Birukou A. Cross-Evaluation of Automated Term Extraction Tools. Technical Report TS-RTDC-TR-2017-1, 30.09.2017, Dept of Comp. Science, Zaporizhzhia National University, Ukraine, 61 p.
19. Deerwester S., Dumais S.T., Furnas G.W. et al. Indexing by latent semantic analysis. Journal of the American Society for Information Science, vol. 41, issue 6, 1990, pp. 391-407.
20. Morris J., Hirst G. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. Computational Linguistics, 1991, vol. 17, 1, pp. 21–43.
21. Tkach S.S. Primeneniye leksicheskikh tsepochek dlya razresheniya leksicheskoy mnogoznachnosti na osnove Russkogo Viki-slovarya [Application of lexical chains for solving lexical polysemy based on the Russian Wiki Dictionary]. Master's thesis. Petrozavodsk State University, Petrozavodsk, 2016, 60 p.
22. Wei T., Lu Y., Chang H., Zhou Q., Bao X. A semantic approach for text clustering using WordNet and lexical chains. Expert Systems with Applications 2015, 42, pp. 2264-2275.