УДК 004.04: 519.767
DOI: 10.15827/0236-235X.116.045-057
Дата подачи статьи: 19.07.16
2016. Т. 29. № 4. С. 45–57

СЕМАНТИЧЕСКИЙ АНАЛИЗ И СПОСОБЫ ПРЕДСТАВЛЕНИЯ СМЫСЛА ТЕКСТА В КОМПЬЮТЕРНОЙ ЛИНГВИСТИКЕ

Т.В. Батура, к.ф.-м.н., старший научный сотрудник, tatiana.v.batura@gmail.com
(Институт систем информатики им. А.П. Ершова СО РАН,
просп. Лаврентьева, 6, г. Новосибирск, 630090, Россия)

Статья посвящена проблемам семантического анализа текстов. Рассмотрены различные методы: диаграммы зависимостей и семантические сети, подходы, основанные на лексических функциях и тематических классах, фреймовые и онтологические модели, логические модели представления знаний. На данный момент существуют различные методы представления смысла высказываний.

Создание новых методов семантического анализа текстов актуально для решения многих задач компьютерной лингвистики, таких как машинный перевод, автореферирование, классификация текстов и других. Не менее важна разработка новых инструментов, позволяющих автоматизировать семантический анализ.

Несмотря на то, что некоторые научные и технические идеи в области обработки текстов развиваются довольно быстро, многие проблемы семантического анализа остаются нерешенными. Большинство исследователей пришло к выводу, что словарь для поддержки семантического анализа должен оперировать смыслами и, следовательно, описывать свойства и отношения понятий, а не слов. Но возникает вопрос, как правильно структурировать и представлять информацию в подобных словарях, чтобы поиск по ним был удобным и быстрым, а кроме того, можно было бы учитывать изменения в естественном языке (исчезновение старых и возникновение новых понятий). В данной статье предпринята попытка систематизировать известные достижения в области семантического анализа и в какой-то мере найти ответ на этот и другие вопросы.

Ключевые слова: *семантический анализ, автоматическая обработка текста, извлечение информации, семантические сети, логика предикатов, представление знаний, смысл высказывания.*

Помимо знаний о структуре языка, семантика тесно связана с философией, психологией и другими науками, так как неизбежно затрагивает вопросы о происхождении значений слов, их отношении к бытию и мышлению. При семантическом анализе необходимо учитывать социальные и культурные особенности носителя языка. Процесс человеческого мышления, как и язык, который является инструментом выражения мыслей, очень гибкий и трудно поддается формализации. Поэтому семантический анализ по праву считается самым сложным этапом автоматической обработки текстов.

На данный момент существует много методов представления смысла высказываний, однако ни один из них не является универсальным. Над соотнесением смысла тексту работали многие исследователи. Так, И.А. Мельчук [1] ввел понятие лексической функции, развил понятия синтаксических и семантических валентностей и рассмотрел их в контексте толково-комбинаторного словаря, который представляет собой языковую модель. Он показал, что значения слов соотносятся не непосредственно с окружающей действительностью, а с представлениями носителя языка об этой действительности. В.Ш. Рубашкин и Д.Г. Лахути [2] ввели иерархию синтаксических связей для более эффективной работы семантического анализатора. Самыми важными являются обязательные ролевые связи, далее идут связи кореференции, затем факультативные ролевые связи и только потом предметно-ассоциативные. Известный лингвист Е.В. Падучева [3] предлагает рассматривать тематические классы слов, в частности глаголов, поскольку они несут основную смысловую нагрузку. Существенной в данном подходе является идея разделения понятий языка на некоторые семантические группы с учетом того, что эти понятия имеют некоторый нетривиальный общий смысловой компонент. Элементы таких групп склонны иметь один и тот же набор зависимых понятий.

Универсальный язык представления знаний должен быть удобным для осуществления вывода новых знаний из уже имеющихся, а значит, необходимо создать аппарат для проверки правильности высказываний. Здесь как раз полезны логические модели представления знаний. Например, семантический язык, предложенный В.А. Тузовым [4], содержит в себе формализмы логики предикатов, в нем присутствуют атомарные понятия, функции над этими понятиями и правила вывода, с помощью которых можно описывать новые понятия. Не исключено, что в направлении создания подобных семантических языков будет развиваться научная мысль в будущем.

Исследование семантики в рамках теории «Смысл ⇔ Текст»

При создании теории «Смысл ⇔ Текст» в [1] введено понятие *лексической функции*. Лексическая функция – определенное смысловое соотношение, например, «равенство по смыслу» (syn), «противоположность по смыслу» (anti), «обобщающее понятие» (gener) и др. Пусть имеется ряд лексических единиц – слов и словосочетаний. Тогда данная лексическая функция ставит в соответствие каждой из этих единиц набор лексических единиц,

находящихся с исходной единицей в соответствующем смысловом соотношении.

Значения одной лексической функции от разных аргументов могут полностью или частично совпадать; могут совпадать и значения разных функций от одного аргумента. На взгляд автора, говорить о лексических функциях как о многозначных не совсем корректно и удобно. Удобнее говорить о лексических предикатах [5]. Помимо простых лексических предикатов, для описания лексической сочетаемости могут использоваться и их комбинации – составные предикаты.

Особую роль при исследовании семантики в подходе И.А. Мельчука играют валентности слов, то есть способность слов вступать в связи с другими словами. Различают два вида валентностей слова: синтаксические и семантические. Хотя это разделение иногда довольно условно. Ситуации, описываемые словами на естественном языке, имеют, как правило, от одного до четырех смысловых компонентов, или семантических актантов. В то же время каждому слову сопоставляются глубинно-синтаксические актанты – зависимые слова, соответствующие подлежащему и дополнениям. Семантические валентности определяются лексическим анализом ситуации, задаваемой конкретным словом. Синтаксические валентности определяются количеством синтаксических актантов, представленных непосредственно в тексте и заданных контекстом.

С формальной точки зрения мы имеем конструкцию, описанную ниже. Чтобы не связывать с каждым глаголом (и другими словами) отдельный предикат, будем рассматривать предикат, размерность которого больше на 1: $P^{val}(y, x_1, x_2, \ldots, x_n)$, при этом y будет само слово, а $x_1, x_2, \ldots, x_n$ – его валентности. Отличать синтаксические и семантические актанты можно с помощью мультииндексов, указывающих, какие именно актанты заданы в тексте. Запись $P^{val}_{i_1 i_2 \ldots i_k}(y, x_{i_1}, x_{i_2}, \ldots, x_{i_k})$ означает, что заданы актанты $i_1, i_2, \ldots, i_k$. В частности, если заданы все актанты, получаем $P^{val}_{1\ldots n}(y, x_1, x_2, \ldots, x_n)$. Некоторые варианты (наборов мультииндексов) могут быть недопустимыми в языке. Если набор $i_1, i_2, \ldots, i_k$ допустим, имеет место импликация

$$\forall y \forall x_1 \ldots \forall x_n (P^{val}_{1\ldots n}(y, x_1, x_2, \ldots, x_n) \rightarrow$$
$$\rightarrow P^{val}_{i_1 i_2 \ldots i_k}(y, x_{i_1}, x_{i_2}, \ldots, x_{i_k})).$$

Одно из главных теоретических изобретений И.А. Мельчука – *толково-комбинаторный словарь*, отражающий прежде всего нетривиальную сочетаемость лексем. Получается, что язык – это очень большая модель, в которой определены лексические предикаты, действующие описанным выше образом.

Словарная статья толково-комбинаторного словаря может быть представлена в виде кортежа $A=\langle w, P_1, \ldots, P_n, Val\rangle$, где w – основное слово; $P_1, \ldots, P_n$ – лексические предикаты, связанные со словом; Val – информация о валентности слова.

В этом случае набор статей в толково-комбинаторном словаре можно считать некоторой подмоделью исходной модели, являющейся языком. Лексические предикаты, определенные теперь на более узком множестве, будут действовать аналогично.

Теория «Смысл ⇔ Текст» с самого начала создавалась для применения в прикладной проблематике автоматического перевода. По замыслу И.А. Мельчука, с ее помощью, в отличие от традиционных нестрогих теорий, следовало обеспечить построение «действующей» модели языка. Теория «Смысл ⇔ Текст» действительно была использована в некоторых системах машинного перевода, разработанных в России, – прежде всего в системе англо-русского автоматического перевода ЭТАП, созданной группой под руководством Ю.Д. Апресяна. Все эти системы относятся к экспериментальным, то есть их промышленное использование не представляется возможным. Несмотря на то, что они включают много полезной лингвистической информации, в целом ни одна из них пока не обеспечила прорыва в качестве перевода.

На взгляд автора, основная ценная идея этой теории состоит в том, что значения слов соотносятся не непосредственно с окружающей действительностью, а с представлениями носителя языка об этой действительности (иногда называемыми концептами). Природа концептов зависит от конкретной культуры; система концептов каждого языка образует так называемую наивную картину мира, которая во многих деталях может отличаться от научной картины мира, являющейся универсальной. Задача семантического анализа лексики в теории «Смысл ⇔ Текст» именно в том, чтобы обнаружить наивную картину мира и описать ее основные категории. Другими словами, важная роль этой теории состоит в описании не только объективной, но и субъективной картины мира.

Хотя интерес к теории И.А. Мельчука угасает, разметка синтаксического корпуса «Национальный корпус русского языка» [6] выполняется лингвистическим процессором ЭТАП-3, основанным на принципах теории «Смысл ⇔ Текст».

Идеи Ю.Д. Апресяна в разработке процессора ЭТАП несколько отличаются от идей И.А. Мельчука. Центральное место в исследованиях Апресяна занимает синонимический словарь нового типа [7]. Для этого словаря была разработана подробная схема описания синонимических рядов, где каждый элемент ряда характеризовался с точки зрения семантики, синтаксиса, сочетаемости и других свойств. В словаре собрано и обобщено максимальное количество информации о языковом поведении русских синонимов.

Концептуальный и прецедентный анализ

На этапе морфологического и семантико-синтаксического анализа текстов основными единицами, обозначающими понятия, являются слова. При таком подходе считается, что смысл словосочетаний и фраз может быть выражен через смыслы составляющих их слов. Такой подход опирается на предположение, что словосочетания, встречающиеся в языке, можно разделить на свободные и несвободные. Другой подход основывается на том, что неделимыми единицами смысла являются категории и понятия, состоящие не из самостоятельных слов, а из словосочетаний [8]. Такие категории и понятия называются концептами. Идея концептуального анализа как неотъемлемой составляющей семантического анализа встречается в исследованиях [2, 9, 10]. В данной работе кратко изложены взгляды на то, какие задачи должны решаться средствами концептуального семантического анализа.

С точки зрения используемых методов и средств семантический анализ должен предусматривать два этапа: этап интерпретации грамматически выраженных (синтаксических и анафорических) связей и этап распознавания связей, не имеющих грамматического выражения. Неоднозначности должны разрешаться самим процессом анализа по критерию степени смысловой удовлетворительности получаемого в каждом варианте результата.

Ключевым пунктом системы семантического анализа является эффективная словарная поддержка. В этом смысле любая система семантического анализа является тезаурусно ориентированной. Процедуры семантического анализа во всех без исключения случаях опираются на функциональность понятийного словаря. Словарь для поддержки семантического анализа должен оперировать смыслами и, следовательно, описывать свойства и отношения понятий, а не слов, поэтому его можно назвать концептуальным словарем [2]. В некотором смысле роль концептуального словаря могут выполнять семантические сети.

В семантическом интерпретаторе прежде всего следует специфицировать различаемые типы семантических отношений в тексте: ролевые (связи по валентности предиката), предметно-ассоциативные (отношения между объектами, процессами, значимые в предметной области) и др.

Принимаются следующие основные правила интерпретации синтаксических связей.

1. Тип устанавливаемого семантического отношения определяется семантическими классами и в определенных случаях более детальными семантическими характеристиками синтаксического «хозяина» и «слуги».

2. Предлоги рассматриваются не как самостоятельный объект интерпретации, а как дополнительная (семантико-грамматическая) характеристика связи между синтаксическим «хозяином» предлога и управляемым им словом.

3. Для разрешения лексической и синтаксической омонимии, фиксируемой синтаксическим анализатором, семантический интерпретатор использует систему эмпирически устанавливаемых предпочтений. На уровне типов семантических отношений устанавливается следующий порядок предпочтений (соответствует уменьшению приоритета связи): функциональные связи и связи, устанавливающие факт смысловой избыточности; ролевые связи, определяемые как обязательные, при наличии семантически согласованного актанта; связи кореференции; ролевые связи, определяемые как факультативные; специфицируемые предметно-ассоциативные связи; неспецифицируемые предметно-ассоциативные связи.

Все большее значение приобретает анализ «по образцу» (прецедентный анализ) [11], основанный на использовании корпуса предварительно размеченных текстов. Система анализа должна обеспечивать не только извлечение знаний из конкретного текста, но и накопление результатов как на синтаксическом, так и на семантическом уровне для использования их далее в качестве прецедентов.

Одним из наиболее масштабных и значимых проектов, осуществляемых в настоящее время, является создание Национального корпуса русского языка [6]. В нем участвует большая группа лингвистов многих научных центров России.

Национальный корпус русского языка – коллекция электронных текстов, снабженных обширной лингвистической и метатекстовой информацией. Корпус представляет все разнообразие стилей, жанров и вариантов русского языка 19–20 вв. В настоящее время в нем используются пять типов разметки: метатекстовая, морфологическая (словоизменительная), синтаксическая, акцентная и семантическая. Остановимся лишь на семантической разметке.

При семантической разметке большинству слов в тексте приписываются один или несколько семантических и словообразовательных признаков и пр. Разметка текстов осуществляется автоматически в соответствии с семантическим словарем корпуса. Поскольку ручная обработка семантически размеченных текстов очень трудоемка, семантическая омонимия в корпусе не снимается: многозначным словам приписываются несколько альтернативных наборов семантических признаков.

В основу семантической разметки положена система классификации русской лексики, принятая в БД «Лексикограф», которая разрабатывалась под руководством Е.В. Падучевой и Е.В. Рахилиной. Подход Е.В. Падучевой часто рассматривается как особое направление в изучении семантики русского языка. В ее работах рассмотрен большой

класс вопросов по этой теме. Наиболее интересными являются исследования тематических классов русских глаголов [3, 12, 13]. Тематический класс объединяет слова с общим семантическим компонентом, который занимает центральное место в их смысловой структуре. Различают, например, глаголы восприятия, знания, эмоций, принятия решения, речевых действий, движения, звука, бытийные глаголы и др.

Для Национального корпуса русского языка был существенно увеличен словник, расширен состав и усовершенствована структура семантических классов, добавлены словообразовательные признаки. Словник семантического словаря базируется на морфологическом словаре системы «Диалинг» (общим объемом порядка 120 тыс. слов), представляющем собой расширение грамматического словаря русского языка А.А. Зализняка. Текущая версия семантического словаря включает слова знаменательных частей речи: существительные, прилагательные, числительные, местоимения, глаголы и наречия. Лексико-семантическая информация имеет различную структуру для разных частей речи.

Существенной в данном подходе является идея разделения понятий языка на некоторые семантические группы с учетом того, что эти понятия имеют некоторый нетривиальный общий смысловой компонент. Элементы таких групп склонны иметь один и тот же набор зависимых понятий. В таком случае словарь для поддержки семантического анализа должен оперировать смыслами и, следовательно, описывать свойства и отношения понятий, а не слов. Остается вопрос, как правильно структурировать и представлять информацию в подобных словарях, чтобы поиск по ним был удобным и быстрым, а кроме того, можно было бы учитывать изменения в естественном языке (исчезновение старых и возникновение новых понятий).

При обсуждении проблем семантики часто упоминают принцип композициональности. Согласно ему, смысл сложного выражения определяется смыслами его составных частей и правилами, применяемыми для их объединения. Поскольку предложение состоит из слов, получается, что его смысл можно представить набором значений слов, входящих в него. Но не все так просто. Смысл предложения также опирается на порядок слов, фразирование и отношения между словами в предложении, то есть учитывает синтаксис.

Как видим, концептуальный анализ позволяет утверждать, что в некоторых случаях принцип композициональности нарушается. Ошибочно утверждать, что смысл словосочетаний и фраз может быть выражен через смысл составляющих их слов. Это не всегда верно. Однако главная проблема такого подхода заключается в том, что выделение тематических классов и составление семантических словарей – чрезвычайно трудоемкий процесс, сильно зависящий от индивидуального восприятия и интерпретации понятий конкретным человеком.

Сетевые модели представления знаний

Тезаурусы, семантические сети, фреймовые и онтологические модели. *Тезаурус* – разновидность словаря общей или специальной лексики, в котором указаны семантические отношения между лексическими единицами. В отличие от толкового словаря тезаурус позволяет выявить смысл не только с помощью определения, но и посредством соотнесения слова с другими понятиями и их группами, благодаря чему может использоваться для наполнения баз знаний систем искусственного интеллекта. В тезаурусах обычно используются следующие основные семантические отношения: синонимы (*смелый–храбрый*), антонимы (*добрый–злой*), гипонимы (*животное–собака*), гиперонимы (*собака–животное*), меронимы (*автомобиль–двигатель*, *колесо*), холонимы (*двигатель*, *колесо–автомобиль*) и паронимы (*индеец–индиец*).

Пример тезауруса – WordNet [14]. Базовой словарной единицей WordNet является синонимический ряд (синсет), объединяющий слова со схожим значением. Синсеты состоят из слов, принадлежащих той же части речи, что и исходное слово. Каждый синсет сопровождается небольшой формулировкой, разъясняющей его значение. Синсеты связаны между собой различными семантическими отношениями. WordNet содержит около 155 тысяч различных лексем и словосочетаний, организованных в 117 тысяч синсетов. Вся БД разбита на три части: существительные, глаголы и прилагательные/наречия. Слово или словосочетание может находиться более чем в одном синсете и принадлежать более чем одной категории части речи.

Семантическая сеть – модель предметной области, имеющая вид ориентированного графа, вершины которого соответствуют объектам предметной области, а дуги (ребра) задают отношения между ними [15]. Объектами могут быть понятия, события, свойства, процессы. Таким образом, семантическая сеть отражает семантику предметной области в виде понятий и отношений. Причем в качестве понятий могут выступать как экземпляры объектов, так и их множества.

Семантические сети возникли как попытка визуализации математических формул. За визуальным представлением семантической сети в виде графа стоит математическая модель, в которой каждая вершина соответствует элементу предметного множества, а дуга – предикату. Терминология, использующаяся в этой области, различна. Чтобы добиться некоторой однородности, узлы, соединенные дугами, принято называть графами, а структуру, где имеется целое гнездо из узлов или где существуют отношения различного порядка между графами, сетью.

Заметим, что среди семантических отношений, применяемых для описания сетей, могут быть не только семантические отношения, используемые в тезаурусах, но и другие связи: функциональные (определяемые обычно глаголами *производит*, *влияет*, …), количественные (*больше*, *меньше*, *равно*, …), пространственные (*далеко от*, *близко от*, *под*, *над*, …), временные (*раньше*, *позже*, *в течение*, …), атрибутивные (*иметь свойство*, *иметь значение*), логические (*И*, *ИЛИ*, *НЕ*) и пр.

Несмотря не некоторые различия, сети удобны для чтения и обработки компьютером, являются наглядным и достаточно универсальным средством представления семантики естественного языка. Однако их формализация в конкретных моделях представления, использования и модификации знаний оказывается достаточно трудоемкой, особенно при наличии множественных отношений между ее элементами. Сеть способна разрастись до большого размера, и, как следствие, поиск вывода в ней будет слишком сложным.

В сложных семантических сетях, включающих множество понятий, процесс обновления узлов и контроль связей между ними, как видим, усложняют процедуру обработки информации. Стремление устранить эти недостатки послужило причиной появления особых типов семантических сетей, таких как фреймовые модели.

Фрейм – структура для описания понятия или ситуации, состоящая из характеристик этой ситуации и их значений [16]. Фрейм можно рассматривать как фрагмент семантической сети, предназначенный для описания понятий со всей совокупностью присущих им свойств. Особенность фреймовых моделей представления знаний в том, что все понятия, описываемые в каждом из узлов модели, определяются набором атрибутов и их значениями, которые содержатся в слотах фрейма ⟨*имя фрэйма*, *слот* 1, *слот* 2, …, *слот N*⟩. Графически это выглядит аналогично семантической сети, но принципиальное отличие заключается в том, что каждый узел во фреймовой модели имеет обобщенную структуру, состоящую из множества слотов, каждый из которых содержит имя, указатель наследования, указатель типа данных и значение.

Слот – это атрибут, связанный с узлом в модели, основанной на фреймах, и являющийся составляющей фрейма. *Имя слота* должно быть уникальным в пределах фрейма. *Указатель наследования* показывает, какую информацию об атрибутах слотов во фрейме верхнего уровня наследуют слоты с теми же именами во фрейме более низкого уровня. *Указатель типа данных* содержит информацию о типе данных, включаемых в слот. Обычно используются следующие типы данных: указатель на имя фрейма верхнего уровня, текст, список, таблица, присоединенная процедура и др. *Значением слота* могут быть экземпляр атрибута, другой фрейм или фасет, оно должно соответствовать указанному типу данных и условию наследования. Помимо конкретного значения, в слоте могут храниться процедуры и правила, которые вызываются при необходимости вычисления этого значения. Таким образом, слот может содержать не только конкретное значение, но и имя процедуры, позволяющей вычислить его по заданному алгоритму, а также одну или несколько продукций, с помощью которых это значение определяется. В слот могут входить несколько значений. Иногда слот включает компонент, называемый фасетом, который задает диапазон или перечень его возможных значений.

Различают фреймы-образцы (прототипы), хранящиеся в базе знаний, и фреймы-экземпляры, создаваемые для отображения реальных ситуаций на основе поступающих данных. Фреймовые модели являются достаточно универсальными, поскольку позволяют отразить все многообразие знаний о мире через фреймы-структуры (для обозначения объектов и понятий: *заем, залог, вексель*), фреймы-роли (*менеджер, кассир, клиент*), фреймы-сценарии (*банкротство, собрание акционеров, празднование именин*), фреймы-ситуации (*тревога, авария, рабочий режим устройства*) и др.

Важнейшим свойством теории фреймов является заимствованное из теории семантических сетей наследование свойств. И во фреймах, и в семантических сетях наследование происходит по *ISA*. Слот *ISA* указывает на фрейм более высокого уровня иерархии, откуда неявно наследуются, то есть переносятся, значения аналогичных слотов.

Основными преимуществами фреймов как модели представления знаний являются соответствие современным представлениям об организации долговременной памяти человека, а также ее гибкость и наглядность. Достоинства фреймовых моделей представления знаний проявляются, если родовидовые связи изменяются нечасто и предметная область насчитывает немного исключений.

Недостаток фреймовых моделей в их относительно высокой сложности, что проявляется в снижении скорости работы механизма вывода и увеличении трудоемкости внесения изменений в сформированную иерархию [17]. Поэтому при разработке фреймовых систем большое внимание уделяется наглядным способам отображения и эффективным средствам редактирования фреймовых структур. Фреймовые модели не позволяют организовать гибкий механизм логического вывода, поэтому фреймовые системы либо представляют собой объектно-ориентированные БД, либо требуют интеграции с другими средствами обработки знаний, например с логическими моделями.

В инженерии знаний под онтологической моделью понимается детальное описание некоторой предметной или проблемной области, которое используется для формулирования утверждений общего характера. Онтологии позволяют представить понятия в таком виде, что они становятся пригод-

ными для машинной обработки. Обычно выделяют следующие основные элементы онтологий: экземпляры, классы объектов (понятий), атрибуты (описывают свойства классов и экземпляров), функции (описывают зависимости между классами и экземплярами), аксиомы (дополнительные ограничения) [18].

Как видим, в центре большинства онтологий находятся классы, описывающие понятия предметной области. Атрибуты описывают свойства классов и экземпляров. Здесь прослеживаются аналогии с фреймовым подходом к формализации знаний. Многие понятия и принципы реализации, а также графическая форма представления на начальном этапе структуризации в онтологиях сходны с семантическими сетями. Основным отличием является ориентация онтологий на использование непосредственно компьютером, то есть структуры данных описаны не на естественном языке (как это принято в семантических сетях и тезаурусах), а на специальном формальном. С тезаурусами онтологии тоже имеют много общего. Но в отличие от них для онтологических моделей необходимыми требованиями являются внутренняя полнота, логическая взаимосвязь и непротиворечивость используемых понятий. В тезаурусах эти требования могут не выполняться.

Специализированные (предметно-ориентированные) онтологии – это представление какой-либо области знаний или части реального мира. В такой онтологии содержатся специальные для этой области значения терминов. К примеру, слово *поле* в сельском хозяйстве означает участок земли, в физике – один из видов материи, в математике – класс алгебраических систем.

Общие онтологии используются для представления понятий, общих для большого числа областей. Такие онтологии содержат базовый набор терминов, глоссарий или тезаурус, используемый для описания терминов предметных областей.

Современные онтологические модели являются модульными, то есть состоят из множества связанных между собой онтологий, каждая из которых описывает отдельную предметную область или задачу. Онтологические модели не являются статичными, они постоянно меняются.

Если использующая специализированные онтологии система развивается, может потребоваться объединение онтологий. Недостатком онтологических моделей является сложность их объединения. Онтологии даже близких областей могут быть несовместимы друг с другом. Объединение онтологий выполняют как вручную, так и в полуавтоматическом режиме. В целом это трудоемкий, медленный и дорогостоящий процесс.

Одна из существующих проблем в онтологическом подходе – представление знаний о времени и об изменениях знаний с течением времени. Однако большинство применяемых на практике языков описания онтологий (например OWL и RDF) основываются на логике предикатов первого порядка и используют унарные или бинарные отношения. В этом случае для описания бинарных отношений с учетом времени требуется вводить в отношения дополнительный параметр, соответствующий времени. При этом бинарные отношения превращаются в тернарные и выходят за рамки описательных возможностей языка.

Еще одной важной задачей является описание знаний о времени с учетом возможной неполноты этих знаний. Эта задача обычно решается в рамках модальных темпоральных логик [19], например LTL, при помощи определенных модальных операторов. Но, поскольку язык описания знаний OWL основан на дескриптивной логике, воспользоваться таким решением для OWL-онтологий невозможно. Интересный способ представления знаний о времени с учетом неопределенности в онтологиях описан в работе А.Ф. Тузовского [20].

Семантические роли и семантические ограничения. Семантические сети позволяют представлять семантику отдельно взятого слова согласно его внутренней структуре. Если вместе с этой структурой учитывать грамматические особенности, то смысл высказывания может быть представлен в терминах семантических ролей и связанных с ними семантических ограничений.

Помимо термина «семантические роли», в литературе используются также понятия: тематические роли, тета-роли, глубинные падежи. Основоположниками данного направления исследований семантики принято считать Дж. Грубера и Ч. Филлмора. По своей сути эти понятия близки к семантическим и глубинно-синтаксическим актантам, исследованием которых занимался И.А. Мельчук. Приведем некоторые семантические роли, рассмотренные в работах [21, 22].

Агенс – одушевленный инициатор и контролер действия. Бенефактив (реципиент, посессор) – участник, чьи интересы косвенно затронуты в ситуации (получает пользу или вред). Инструмент – стимул эмоции или участник, с помощью которого выполняется действие. Источник – место, из которого осуществляется движение. Контрагент – сила или сопротивляющаяся среда, против которой выполняется действие. Объект – участник, который передвигается или изменяется в ходе события. Пациенс – участник, претерпевающий существенные изменения. Результат – участник, который появляется в результате события. Стимул – внешняя причина или объект, вызывающие это состояние. Цель – место, в которое осуществляется движение.

В соответствии с числом аргументов и их семантическими свойствами множество глагольных лексем можно разбить на классы: глаголы физического воздействия (*рубить*, *пилить*), глаголы восприятия (*видеть*, *слышать*), глаголы способа речи (*кричать*, *шептать*) и др. Внутри каждого класса

существует более точное деление. Среди глаголов физического воздействия похожую семантическую предикатно-аргументную структуру имеют глаголы вида глагол (агенс, инструмент, объект): *break – разбить, crack – расколоть* и т.д. Другая предикатно-аргументная структура характерна для глаголов вида глагол (агенс, инструмент, цель): *hit – ударить, slap – шлепнуть, strike – ударить* и пр.

Было замечено, что существуют корреляции между морфологическими падежами, предлогами, синтаксическими ролями, с одной стороны, и семантическими ролями, с другой стороны, например, «cut *with* a knife», «work *with* John». Кроме того, следует учитывать, что у одного предикатного слова не может быть двух актантов с одной и той же семантической ролью. Различия в наборах ролей затрагивают в основном периферийные семантические роли (контрагент, стимул, источник) или сводятся к объединению/фрагментации ядерных ролей.

К сожалению, в результате многократных исследований не удалось установить взаимно-однозначное соответствие между семантическими ролями и падежами. Ситуация осложняется еще и тем, что сами роли нетривиально связаны между собой, а в естественных языках распространены такие генеративные приемы, как метафора и метонимия, которые порождают множество новых смыслов и не могут в принципе отражаться в статическом лексиконе.

Логические модели представления знаний

При построении логических моделей представления знаний вся информация, необходимая для решения прикладных задач, рассматривается как совокупность фактов и утверждений, которые представляются в виде формул в некоторой логике. Знания отображаются совокупностью таких формул, а получение новых знаний сводится к реализации процедур логического вывода. В основе логических моделей представления знаний лежит понятие формальной теории, задаваемое кортежем $S = <B, F, A, R>$, где B – счетное множество базовых символов (алфавит); F – множество, называемое формулами; A – выделенное подмножество априори истинных формул (аксиом); R – конечное множество отношений между формулами, называемое правилами вывода.

Основной подход к представлению смысла в компьютерной лингвистике включает в себя создание представления смысла в формальном виде. Такое представление описывается языком представления смысла. Он необходим для того, чтобы ликвидировать разрыв между естественным языком и общесмысловыми знаниями о мире. Поскольку предполагается использовать этот язык для автоматической обработки текстов и при создании систем искусственного интеллекта, необходимо учитывать вычислительные требования семантической обработки, такие как необходимость определять истинность высказываний, поддерживать однозначность представления, представлять высказывания в канонической форме, обеспечивать логический вывод и быть выразительными.

В естественных языках существует большое разнообразие приемов для передачи смысла. Среди наиболее важных – способность передавать предикатно-аргументную структуру. Учитывая вышесказанное, получаем, что в качестве инструмента для представления смысла высказываний хорошо подходит логика предикатов первого порядка. С одной стороны, она относительно легко понимается человеком, с другой, хорошо поддается обработке (вычислительной). При помощи логики первого порядка могут быть описаны важные смысловые классы, включающие события, время и другие категории. Однако следует помнить, что высказывания, соответствующие таким понятиям, как убеждения и желания, требуют выражений, включающих модальные операторы. Язык представления смысла, как и любой язык, должен иметь свой синтаксис. Например, в работе [23] можно найти описание контекстно-свободной грамматики для исчисления предикатов первого порядка.

Семантические сети и фреймы могут быть рассмотрены в рамках логики предикатов первого порядка [24]. Например, смысл предложения *У меня есть книга* можно записать четырьмя различными способами с использованием четырех различных языков представления смысла (рис. 1).

Несмотря на то, что все эти четыре подхода различны, на абстрактном уровне они представляют собой общепринятое фундаментальное обозначение того, что представление смысла состоит из структур, составленных из множества символов. Эти символьные структуры соответствуют объектам и отношениям между объектами. Все четыре представления состоят из символов, соответствующих «говорящему», «книге» и набору отношений, обозначающих обладание одного другим. Важным здесь является то, что все эти четыре представления позволяют связать, с одной стороны, выразительные особенности естественного языка, а с другой – реальное состояние дел в мире.

Логические модели представления знаний обладают рядом преимуществ. Во-первых, в качестве «фундамента» здесь используется классический аппарат математической логики, методы которой достаточно хорошо изучены и формально обоснованы. Во-вторых, существуют достаточно эффективные процедуры вывода синтаксически правильных высказываний. В-третьих, такой подход позволяет хранить в базах знаний лишь множество аксиом, а все остальные знания (в том числе факты и сведения о людях, предметах, событиях и процессах) получать из этих аксиом по правилам вывода.

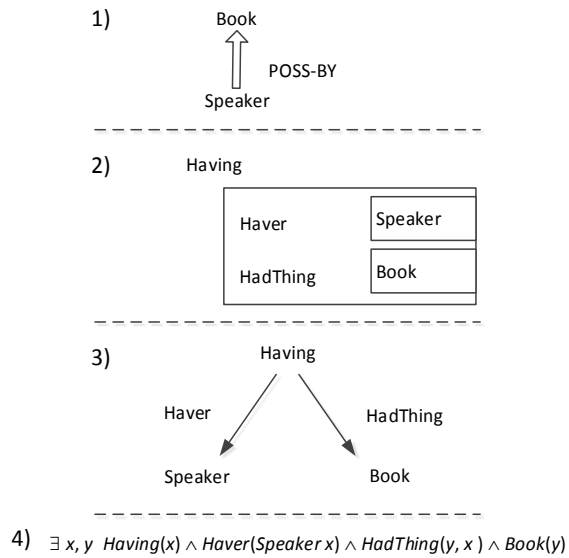

Рис. 1. Четыре языка представления смысла высказывания: *1) концептуальная диаграмма зависимостей, 2) представление, основанное на фреймах, 3) семантическая сеть, 4) исчисление предикатов первого порядка*

Fig. 1. Four languages of statement meaning representation: *1) a conceptual dependency graph, 2) a representation based on frames, 3) a semantic network, 4) first-order predicate calculus*

Вывод синтаксически правильных высказываний в логических моделях представления знаний опирается на правило резолюций, разработанное Дж. Робинсоном в 1965 году. Оно утверждает: если группа выражений, образующая посылку, является истинной, то применение правила вывода гарантированно обеспечит получение истинного выражения в качестве заключения. Результат применения правила резолюций называют резольвентой.

Метод резолюций (или правило устранения противоречий) позволяет проводить доказательство истинности или ложности выдвинутого предположения методом от противного. В методе резолюций множество предложений обычно рассматривается как составной предикат, который содержит несколько предикатов, соединенных логическими функциями и кванторами существования и всеобщности. Так как одинаковые по смыслу предикаты могут иметь разный вид, предложения сначала необходимо привести к унифицированному виду (к дизъюнктивной или конъюнктивной нормальной форме), то есть удалить кванторы существования, всеобщности, символы импликации, эквивалентности и др. Правило резолюций содержит в левой части конъюнкцию дизъюнктов. Поэтому приведение посылок, используемых для доказательства, к виду, представляющему собой конъюнкции дизъюнктов, является необходимым этапом практически любого алгоритма, реализующего логический вывод на базе метода резолюций [25].

Именно правило резолюций послужило основой для создания языка программирования Prolog. В языке Prolog факты описываются в форме логических предикатов с конкретными значениями. Правила вывода описываются логическими предикатами с определением правил вывода в виде списка предикатов над базами знаний и процедурами обработки информации. Интерпретатор языка Prolog самостоятельно реализует вывод, подобный вышеописанному. Чтобы инициировать вычисления, выполняется специальный запрос к базе знаний, на который система логического программирования генерирует ответы «истина» и «ложь».

Метод резолюций легко программируется, это одно из важнейших его достоинств, однако он применим только для ограниченного числа случаев, так как для его применения доказательство не должно иметь большую глубину, а число потенциальных резолюций не должно быть большим.

После того как язык Prolog приобрел большую популярность, в начале 80-х годов прошлого века появился термин «компьютеры пятого поколения». В то время ожидалось создание следующего поколения компьютеров, ориентированного на распределенные вычисления. Вместе с этим считалось, что пятое поколение станет основой для создания устройств, способных имитировать процесс человеческого мышления. Тогда же возникла идея создания аппаратной поддержки параллельных реляционных БД Grace и Delta [26, 27] и параллельного логического вывода (Parallel Inference Engine, PIE), опирающаяся на принципы языка Prolog. Каждый блок логического вывода сообщает о своей текущей рабочей нагрузке таким образом, чтобы работа могла быть передана в блок логического вывода с наименьшей нагрузкой [28]. Но, как известно, подобные попытки не позволили создать искусственный интеллект, а лишь послужили очередным подтверждением того, что человеческое мышление еще недостаточно изучено.

Системы с компонентами семантического анализа

Проект Open Cognition. В рамках проекта Open Cognition [29] разрабатывается анализатор Link Grammar Parser, который отвечает за обработку естественного языка. Link Grammar Parser начал разрабатываться в 1990-е гг. в университете Карнеги–Меллона [30]. Данный подход отличается от классической теории синтаксиса. Система приписывает предложению синтаксическую структуру, которая состоит из множества помеченных связей (коннекторов), соединяющих пары слов. Link Grammar Parser использует информацию о типах связей между словами. В настоящий момент поддерживаются словари для иврита, английского,

немецкого, русского, турецкого, персидского, арабского, латышского и вьетнамского языков.

Главной причиной, по которой анализатор называют семантической системой, можно считать уникальный по полноте набор связей (около 100 основных, причем некоторые из них имеют 3-4 варианта). В некоторых случаях тщательная работа над разными контекстами привела авторов системы к переходу к почти семантическим классификациям, построенным на синтаксических принципах.

Проект Open Cognition, в рамках которого развивается Link Grammar Parser, открытый и бесплатный, что является большим преимуществом для проведения исследований. Довольно подробное описание и исходный код можно найти на сайте [31]. Open Cognition продолжает развиваться, что также важно, поскольку есть возможность взаимодействовать с разработчиками. Наравне с Link Grammar ведется разработка анализатора RelEx [32], который позволяет извлекать отношения семантической зависимости в высказываниях на естественном языке и в результате представлять предложения в виде деревьев зависимостей. Он использует несколько наборов правил для перестроения графа с учетом синтаксических связей между словами. После каждого шага, согласно набору правил сопоставления, в полученном графе добавляются теги структурных характеристик и отношений между словами. Однако некоторые правила, наоборот, могут сокращать граф. Таким образом происходит преобразование графа. Этот процесс применения последовательности правил напоминает метод, используемый в ограничительных грамматиках. Главное отличие состоит в том, что RelEx работает с графовым представлением, а не с простыми наборами тегов (обозначающими отношения). Эта особенность позволяет применять более абстрактные преобразования при анализе текстов. Другими словами, основная идея состоит в том, чтобы использовать распознавание образов для преобразования графов. В отличие от других анализаторов, которые полностью опираются на синтаксическую структуру предложения, RelEx больше ориентирован на представление семантики, в частности, это касается сущностей, сравнений, вопросов, разрешения анафор и лексической многозначности слов.

Система «Диалинг». Эта автоматическая система русско-английского перевода разрабатывалась в 1999–2002 гг. в рамках проекта «Автоматическая обработка текста». В разное время в работе над ней принимали участие двадцать два специалиста, большинство из которых известные ученые-лингвисты. За основу системы «Диалинг» были взяты система французско-русского автоматического перевода, разработанная в ВЦП совместно с МГПИИЯ им. М. Тореза в 1976–1986 гг., и система анализа политических текстов на русском языке «Политекст», разработанная в Центре информационных исследований в 1991–1997 гг.

Система «Политекст» была направлена на анализ официальных документов на русском языке и содержала полную цепочку анализаторов текста: графематический, морфологический, синтаксический и частично семантический. В системе «Диалинг» был частично заимствован графематический анализ, но адаптирован под новые стандарты программирования. Программа морфологического анализа была написана заново, поскольку скорость работы была низкой, но сам морфологический аппарат не изменился [33].

На графематическом уровне константами являются графематические дескрипторы: ЛЕ (лексема) – присваивается последовательностям, состоящим из кириллических символов; ЦК (цифровой комплекс) – присваивается последовательностям, состоящим из цифр, и т.д. На морфологическом уровне для обозначений используются граммемы: тв – творительный падеж, мн – множественное число, но – неодушевленность, св – совершенный вид, пе – переходность глагола и т.д. Возможные типы фрагментов на этапе фрагментационного анализа: главные предложения, придаточные предложения в составе сложного, причастные, деепричастные и другие обособленные обороты. Про каждый фрагмент известно, какие фрагменты в него непосредственно вложены и в какие он непосредственно вложен.

Основными составляющими применяемого в «Диалинге» семантического аппарата являются семантические отношения и семантические характеристики. Примеры семантических отношений: ИНСТР – «инструмент», ЛОК – «локация, местоположение», ПРИНАДЛ – «принадлежность» и пр. Они довольно универсальны и имеют сходство с предикатами и семантическими ролями. Семантические характеристики позволяют строить формулы с использованием логических связок «и» и «или». Каждому слову приписывается некоторая формула, составленная из семантических характеристик. В семантическом словаре «Диалинга» содержится около 40 семантических характеристик. Примеры семантических характеристик: ГЕОГР – географический объект; ДВИЖ – глаголы движения; ИНТЕЛ – действия, связанные с мыслительной деятельностью; НОСИНФ – носители информации; ЭМОЦ – прилагательные, которые выражают эмоции, и т.д. Некоторые характеристики являются составными, так как их можно выразить через другие. Семантические характеристики наравне с грамматическими характеристиками обеспечивают проверку согласования слов при интерпретации связей в тексте.

В данный момент все инструменты, разработанные в рамках проекта «Автоматическая обработка текста» (в том числе система «Диалинг»), являются свободным кроссплатформенным ПО.

Демоверсия и подробная документация доступны на сайте [34].

Другие системы семантического анализа. Существуют и другие системы, содержащие компоненты семантического анализа. Однако они имеют существенные недостатки для исследований: сложно найти описание, не являются бесплатными и свободно распространяемыми или не работают с текстами на русском языке. К таким системам относятся *OpenCalais* (http://www.opencalais.com/opencalais-api/), *RCO* (http://www.rco.ru/?page_id=3554), *Abbyy Compreno* (https://www.abbyy.com/ru-ru/isearch/compreno/), *SemSin* (http://www.dialog-21.ru/media/1394/kanevsky.pdf), *DictaScope* (http://dictum.ru/) и др.

Следует упомянуть систему извлечения данных из неструктурированных текстов Pullenti (http://semantick.ru/). Она заняла первое место на дорожках T1, T2, T2-m и второе место на T1-l на конференции «Диалог-2016» в соревновании FactRuEval. На сайте разработчиков системы Pullenti есть также демоверсия семантического анализатора, позволяющего по предложению строить семантическую сеть.

Инструментальная среда «ДЕКЛ» (http://ipiran-logos.com/) разработана в конце 90-х годов и использована для построения экспертных систем, оболочек для экспертных систем, логико-аналитических систем, лингвистических процессоров, обеспечивающих обработку и автоматическое извлечение знаний из потоков неформализованных документов на естественном языке.

Система машинного перевода «ЭТАП-3» предназначена для анализа и перевода текстов на русском и английском языках. Система использует преобразование текстов на естественном языке в их семантическое представление на языке Universal Networking Language. Как уже говорилось ранее, разметка синтаксического корпуса «Национальный корпус русского языка» [6] выполняется лингвистическим процессором ЭТАП-3, основанным на принципах теории «Смысл ⇔ Текст».

В последнее время появляется все больше систем представления баз знаний в виде графов. Поскольку объемы информации постоянно увеличиваются с невероятной скоростью, такие системы должны поддерживать построение и пополнение баз знаний в автоматическом режиме. Автоматическое построение баз знаний может осуществляться на основе структурированных источников данных. Примерами таких систем являются *Yago* (http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/), *DBpedia* (http://wiki.dbpedia.org/), *Freebase* (https://developers.google.com/freebase/), *Google's Knowledge Graph* (https://developers.google.com/knowledge-graph/), *OpenCyc* (http://www.opencyc.org/). Другой подход позволяет извлекать информацию из открытых ресурсов в Интернете без участия человека: *ReadTheWeb* (http://rtw.ml.cmu.edu/rtw/), *OpenIE* (http://nlp.stanford.edu/software/openie.html), *Google Knowledge Vault* (https://www.cs.ubc.ca/~murphyk/Papers/kv-kdd14.pdf). Подобные системы являются экспериментальными, каждая из них имеет свои особенности. Например, Knowledge Vault пытается учитывать неопределенности, каждому факту ставятся в соответствие коэффициент доверия и происхождение информации. Таким образом, все утверждения делятся на те, которые имеют высокую вероятность быть истинными, и те, которые могут быть менее вероятными. Предсказание фактов и их свойств осуществляется методами машинного обучения на основе очень большого количества текстов и уже имеющихся фактов. В данный момент Knowledge Vault содержит 1,6 млрд фактов. Система NELL, разрабатываемая в рамках проекта ReadTheWeb университетом Карнеги–Меллона, содержит более 50 млн утверждений с разными степенями доверия. Около 2 млн 800 тыс. фактов имеют высокую степень доверия. Процесс обучения NELL также еще не завершен.

Применение гибридного подхода в системах семантического анализа текстов. Учитывая вышесказанное, получаем, что при работе с семантикой текстов приходится иметь дело с гибкими, постоянно меняющимися структурами. Если на семантическом уровне представлять последовательность высказываний в виде графа, становится ясно, что этот граф нужно постоянно перестраивать по ходу повествования или диалога. Добавляется новая информация, исчезает старая и изменяется уже имеющаяся. Поэтому подходы, основанные на наборах заранее заданных правил, применяемых ранее для обработки текстов, дают недостаточно хорошие результаты. Основным их недостатком является невозможность эффективно осуществлять вывод новых правил.

Одним из возможных решений является применение гибридного подхода в системах семантического анализа текстов. Подразумевается, что гибридный подход сочетает в себе методы машинного обучения и методы, основанные на правилах. На рисунке 2 представлена основная идея применения гибридного подхода для систем семантического анализа текстов.

Семантический анализ текста сводится к анализу семантического пространства, то есть смысловой модели текста. Для изучения свойств семантического пространства может быть введено понятие размерности. Размерность семантического пространства – количество возможных вариантов сопоставления смысла тексту. Для синтаксического пространства размерность – количество возможных синтаксических ролей, корректно приписанных словам. Тогда можно утверждать, что размерность семантического пространства больше размерности синтаксического пространства ввиду многозначности не только семантических правил,

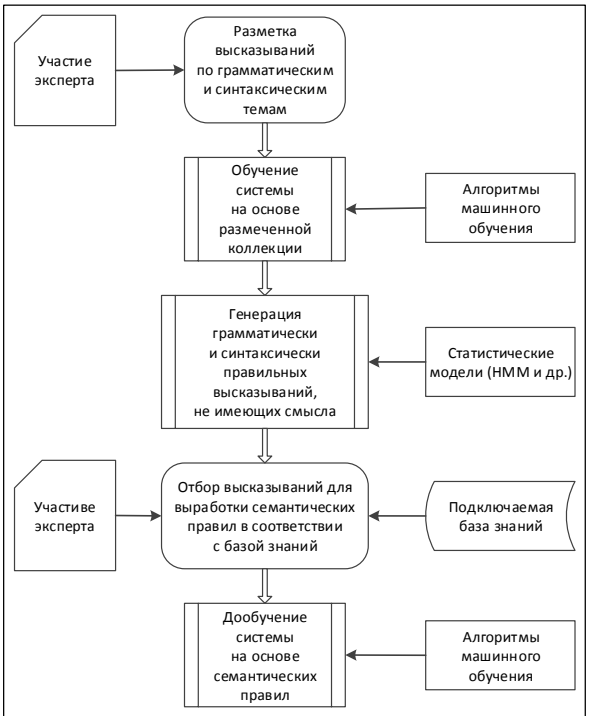

Рис. 2. Применение гибридного подхода для семантического анализа текстов

Fig. 2. Using a hybrid approach for semantic intelligence

но и лексических единиц. Для текстов возникают также понятия семантических подпространств и проекций семантических пространств. В будущем планируется формализовать эти и другие понятия, исследовать свойства семантических пространств и подпространств, реализовать прототип системы семантического анализа.

В заключение отметим, что с развитием компьютерных технологий и постоянным ростом объемов текстовой информации исследования в области автоматической обработки текстов сфокусировались на прикладных аспектах. Однако в настоящее время возможности большинства программных инструментов ограничиваются морфологическим и синтаксическим анализом в сочетании с методами из теории вероятностей и математической статистики. Таким образом, лишь избранная часть относительно простых задач оказалась решенной, множество проблем предстоит решить в будущем.

Очевидно, причин для этого много. Например, существует мнение, что каждое правило в синтаксисе имеет свой аналог в семантике. Этот постулат называют гипотезой «правило к правилу» (rule-to-rule hypothesis) [35]. На самом деле это соответствие не является взаимно-однозначным, и в этом состоит главная сложность. Действительно, каждому синтаксическому правилу (дереву разбора) можно сопоставить семантическое правило (дерево разбора), но оно не будет единственным. Аналогично семантическому правилу сопоставляется синтаксическое правило, но необязательно единственное. Именно эта неоднозначность приводит к неразрешимым на сегодняшний день проблемам в области автоматической обработки текстов. В связи с этим рассуждением возникает вопрос выбора нужного сопоставления из большого количества возможных вариантов.

Из всего вышесказанного можно сделать еще один очень важный вывод. Не следует рассматривать процессы генерации и интерпретации высказывания отдельно, так как они неразрывно связаны между собой. Выражая свою мысль, человек ориентируется на то, поймет ли его собеседник. В процессе генерации высказывания человек как бы перепроверяет себя, моделируя, как собеседник воспримет информацию. Похожий механизм действует при интерпретации высказывания. При осмыслении услышанного или прочитанного мы опять же сверяемся с нашими знаниями и представлениями о мире. Только благодаря этому нам удается выбрать подходящий смысл.

Современные исследователи склоняются к мысли, что правильный выбор можно сделать, имея дополнительную базу знаний о мире. Такая база знаний должна содержать общесмысловую информацию о понятиях и отношениях между ними, чтобы при обращении к ней можно было определить подходящий контекст высказывания в автоматическом режиме. Она помогла бы учитывать накопленные знания о мире, которые в явном виде не присутствуют в конкретном высказывании, но непосредственно влияют на его смысл.

В данной статье предпринята попытка систематизировать известные на сегодняшний день достижения в области машинно-ориентированного семантического анализа. Расширенный вариант статьи доступен по ссылке http://swsys-web.ru/methods-and-systems-of-semantic-text-analysis.html.

Литература

1. Мельчук И.А. Опыт теории лингвистических моделей «Смысл–Текст». М.: Школа «Языки русской культуры», 1999. 346 с.

2. Лахути Д.Г., Рубашкин В.Ш. Семантический (концептуальный) словарь для информационных технологий // Научно-техническая информация. 2000. № 7. С. 1–9.

3. Падучева Е.В. Динамические модели в семантике лексики. М.: Языки славянской культуры, 2004. 608 с.

4. Тузов В.А. Компьютерная семантика русского языка. СПб: Изд-во СПбГУ, 2003. 391 с.

5. Батура Т.В., Мурзин Ф.А. Машинно-ориентированные логические методы отображения семантики текста на естественном языке: монография. Новосибирск: Изд-во НГТУ, 2008. 248 с.

6. Национальный корпус русского языка. URL: http://www.ruscorpora.ru/ (дата обращения: 22.06.2016).

7. Апресян В.Ю., Апресян Ю.Д., Бабаева Е.Э., Богуславская О.Ю., Галактионова И.Г., Гловинская М.Я., Григорьева С.А., Иомдин Б.Л. и др. Новый объяснительный словарь синонимов русского языка. М.–Вена: Языки славянской культуры–Венский славистический альманах, 2004. 1488 с.

8. Хорошилов А.А. Методы автоматического установления смысловой близости документов на основе их концептуаль-

ного анализа // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: тр. XV Всерос. науч. конф. RCDL' 2013. Ярославль: Изд-во ЯрГУ, 2013. С. 369–376.

9. Рубашкин В.Ш. Представление и анализ смысла в интеллектуальных информационных системах. М.: Наука, 1989. 189 с.

10. Лахути Д.Г., Рубашкин В.Ш. Средства и процедура концептуальной интерпретации входных сообщений на естественном языке // Изв АН СССР: Сер. Технич. киберн. 1987. № 2. С. 49–59.

11. Рубашкин В.Ш. Семантический компонент в системах понимания текста // КИИ-2006: тр. 10 Национ. конф. по искусствен. интеллекту с междунар. участ. 2006. URL: http://www.raai.org/resurs/papers/kii-2006/#dokladi (дата обращения: 23.06.2016).

12. Падучева Е.В. Семантика вида и точка отсчета // Изв. АН СССР: Сер. лит. и яз. 1986. Т. 45. № 5. С. 18–25.

13. Падучева Е.В. Отпредикатные имена в лексикографическом аспекте // Научно-техническая информация. 1991. Сер. 2. № 5. С. 21–31.

14. WordNet. A lexical database for English. URL: http://wordnet.princeton.edu/ (дата обращения: 23.06.2016).

15. Семантическая сеть. URL: https://ru.wikipedia.org/wiki/Семантическая_сеть (дата обращения: 23.06.2016).

16. Хабаров С.П. Представление знаний в информационных системах: конспекты лекций. URL: http://www.habarov.spb.ru/bz/bz07.htm (дата обращения: 23.06.2016).

17. Луценко Е.В. Представление знаний в информационных системах: электрон. учеб. пособие для студентов. Краснодар: Изд-во КубГАУ, 2010. 428 с.

18. Константинова И.С., Митрофанова О.А. Онтологии как системы хранения знаний // Информационно-телекоммуникационные системы: Всерос. конкурс. отбор статей, 2008. 54 с.

19. Темпоральная логика. URL: https://ru.wikipedia.org/wiki/Темпоральная_логика (дата обращения: 23.06.2016).

20. Разин В.В., Тузовский А.Ф. Представление знаний о времени с учетом неопределенности в онтологиях Semantic WEB // Докл. ТУСУР. 2013. № 2 (28). С. 157–162.

21. Fillmore Ch. The Case for Case. Proc. Texas Sympos. on Language Universals, 1967, 134 p.

22. Филлмор Ч. Дело о падеже // Новое в зарубежной лингвистике. М.: Прогресс, 1981. С. 369–495.

23. Рассел С., Норвиг П. Искусственный интеллект: современный подход. М.: Вильямс, 2007. 1408 с.

24. Jurafsky D., Martin J. Speech and language processing: an introduction to natural language processing, computational linguistics and speech recognition. 2008, 1024 p.

25. Вывод в логических моделях. Метод резолюций. URL: http://www.aiportal.ru/articles/knowledge-models/method-resolution.html (дата обращения: 11.07.2016).

26. Boral H., Redfield S. Database Machine Morphology. Proc. 11th Intern. Conf. Very Large Data Bases, 1985, pp. 59–71.

27. Fushimi S., Kitsuregawa M., Tanaka H. An overview of the system of a parallel relational database machine GRACE. Proc. 12th Intern. Conf. Very Large Data Bases, 1986, pp. 209–219.

28. Tanaka H. Parallel Inference Engine. IOS Press Publ., 2000, 296 p.

29. Open Cognition. URL: http://opencog.org/ (дата обращения: 23.06.2016).

30. Link Grammar Parser. AbiWord, 2014. URL: http://www.abisource.com/projects/link-grammar/ (дата обращения: 20.06.2016).

31. The CMU Link Grammar natural language parser. URL: https://github.com/opencog/link-grammar/ (дата обращения: 22.06.2016).

32. RelEx Dependency Relationship Extractor. OpenCog. URL: http://wiki.opencog.org/wikihome/index.php/Relex (дата обращения: 22.06.2016).

33. Сокирко А.В. Семантические словари в автоматической обработке текста (по материалам системы ДИАЛИНГ): дисс. ... канд. тех. наук. М.: МГПИИЯ, 2001. 120 с.

34. Автоматическая обработка текста. URL: http://aot.ru/ (дата обращения: 23.06.2016).

35. Prószéky G. Machine Translation and the rule-to-rule hypothesis. New Trends in Translation Studies (In Honour of Kinga Klaudy). Budapest: Akadémiai Kiadó, 2005, pp. 207–218.

Software & Systems
DOI: 10.15827/0236-235X.116.045-057

Received 19.07.16
2016, vol. 29, no. 4, pp. 45–57

SEMANTIC ANALYSIS AND METHODS OF TEXT MEANING REPRESENTATION IN COMPUTER LINGUISTICS

T.V. Batura[1]*, Ph.D. (Physics and Mathematics), Senior Researcher, tatiana.v.batura@gmail.com*
[1] *A.P. Ershov Institute of Informatics Systems (IIS), Siberian Branch of the Russian Federationn Academy of Sciences, Lavrentev Av. 6, Novosibirsk, 630090, Russian Federation*

Abstract. The paper is devoted to the problems of semantic analysis of texts. The article discusses different methods, such as dependency diagrams, semantic network, approaches based on lexical functions and thematic classes, frame, ontological and logical models of knowledge representation. At the moment, there are many methods of representing sentence meaning.

Creating new methods of semantic analysis is significant in solving many problems of computational linguistics such as machine translation, automatic summarization, text classification and others. Development of new tools for semantic analysis is equally important.

Despite the fact that some of the scientific and technical ideas in natural language processing are evolved, many problems of semantic analysis remain unsolved. Most of researchers came to the conclusion that the dictionary for semantic analysis has to operate meanings and, therefore, describe the properties and relationships between concepts, rather than individual words. But there is a question: how to organize and represent information in these dictionaries to search it fast and conveniently, and in addition, take into account the changes in the natural language (the disappearance of old and the emergence of new concepts). This paper attempts to answer this and other questions. The article undertakes an attempt to systematize known achievements in the field of a semantic analysis, and in any measure to find the answer to this and other questions.

Keywords: semantic analysis, natural language processing, information retrieval, semantic networks, predicate logic, knowledge representation, meaning of the sentence.

References

1. Melchuk I.A. *Opyt teorii lingvisticheskikh modeley "Smysl–Tekst"*. Moscow, Yazyki russkoy kultury Publ., 1999, 346 p.
2. Lakhuti D.G., Rubashkin V.Sh. Semantic (conceptual) dictionary for information technologies. *Nauchno-tekhnicheskaya informatsiya* [Scientific and Technical Information]. 2000, no. 7, pp. 1–9 (in Russ.).
3. Paducheva E.V. *Dinamicheskie modeli v semantike leksiki* [Dynamic Models in Lexis Semantics]. Moscow, Yazyki russkoy kultury Publ., 2004, 608 p.
4. Tuzov V.A. *Kompyuternaya semantika russkogo yazyka* [Russian Language Computer Semantics]. St. Petersburg, SPbGU Publ., 2003, 391 p.
5. Batura T.V., Murzin F.A. *Mashinno-orientirovannye logicheskie metody otobrazheniya semantiki teksta na estestvennom yazyke* [Computer Oriented Logical Methods of Text Semantics Representation on a Natural Language]. Monograph. Novosibirsk, NGTU, 2008, 248 p.
6. *Russian National Corpus*. Available at: http:// http://www.ruscorpora.ru/en/index.html/ (accessed June 22, 2016).
7. Apresyan V.Yu., Apresyan Yu.D., Babaeva E.E., Boguslavskaya O.Yu., Galaktionova I.G., Glovinskaya M.Ya., Grigoreva S.A., Iomdin B.L., Krylova T.V., Levontina I.B., Ptentsova A.V., Sannikov A.V., Uryson E.V. *Novy obyasnitelny slovar sinonimov russkogo yazyka* [The New Explanatory Dictionary of Russian Synonyms]. 2nd ed., Moscow, Vena, 2004, 1488 p.
8. Khoroshilov A.A. Methods of automatic setting documents' semantic adjacency based on their conceptual analysis. *Elektronnye biblioteki: perspektivnye metody i tekhnologii, elektronnye kollektsii: XV Vseross. nauch. konf. RCDL' 2013* [Proc. 15th All-Russian Scientific Conf. "Electronic Libraries: Prospect Methods and Technologies, Electronic Collections"]. Yaroslavl, YarGU Publ., 2013, pp. 369–376 (in Russ.).
9. Rubashkin V.Sh. *Predstavlenie i analiz smysla v intellektualnykh informatsionnykh sistemakh* [Meaning Representation and Analysis in Intelligent Information Systems]. Moscow, Nauka Publ., 1989, 189 p.
10. Lakhuti D.G., Rubashkin V.Sh. Tools and a Procedure of Conceptual Interpretation of Input Messages on a Natural Language. *Izvestiya AN SSSR. Tekhnicheskaya kibernetika* [News of USSR Academy of Sciences. Technical Cybernetics]. 1987, no. 2, pp. 49–59 (in Russ.).
11. Rubashkin V.Sh. Semantic component in text understanding systems. *Tr. 10 nats. konf. po iskusstvennomu intellektu s mezhdunar. uchastiem (KII-2006)* [Proc. 10th National Conf. on Artificial Intelligence with Int. Participation (KII-2006)]. 2006. Available at: http://www.raai.org/resurs/papers/kii-2006/#dokladi (accessed June 22, 2016).
12. Paducheva E.V. Semanticsc of a kind and a reference point. *Izvestiya AN SSSR. Seriya lit. i yaz*. [News of USSR Academy of Sciences. Literature and Language]. 1986, vol. 45, no. 5, pp. 18–25 (in Russ.).
13. Paducheva E.V. Edited names in a lexicographical aspect. *Nauchno-tekhnicheskaya informatsiya* [Scientific and Technical Information]. 1991, iss. 2, no. 5, pp. 21–31 (in Russ.).
14. *WordNet. A lexical database for English*. Available at: http://wordnet.princeton.edu/ (accessed June 23, 2016).
15. *Semantic network. Wikipedia – The Free Encyclopedia*. Available at: https://en.wikipedia.org/wiki/Semantic_network (accessed June 23, 2016).
16. Khabarov S.P. *Predstavlenie znany v informatsionnykh sistemakh* [Knowledge Representation in Information Systems]. Lecture notes. Available at: http://www.habarov.spb.ru/bz/bz07.htm (accessed June 23, 2016) (in Russ.).
17. Lutsenko E.V. *Predstavlenie znany v informatsionnykh sistemakh* [Knowledge Representation in Information Systems]. Electronic study guide. Krasnodar, KubGAU Publ., 2010, 428 p.
18. Konstantinova I.S., Mitrofanova O.A. Ontologies as knowledge storage systems. *Vseross. konkursny otbor obzorno-analiticheskikh statey po prioritetnomu napravleniyu "Informatsionno-telekommunikatsionnye sistemy"* [All-Russian Competitive Selection of Review and Analytical Articles on Priority Area "Information and Analytical Systems"]. 2008, 54 p.
19. *Tepmoral logic. Wikipedia – The Free Encyclopedia*. Available at: https://en.wikipedia.org/wiki/Temporal_logic (accessed June 23, 2016).
20. Razin V.V., Tuzovsky A.F. Time knowledge representation taking into account uncertainty in Semantic WEB ontologies. *Doklady TUSUR* [Proc. of TUSUR Univ.]. 2013, no. 2 (28), pp. 157–162 (in Russ.).
21. Fillmore Ch. The Case for Case. *Proc. of the Texas Symp. on Language Universals*. 1967, 134 p.
22. Fillmore Ch. *The Case for Case. Universals in Linguistic Theory*. In Bach and Harms (Ed.). NY, Holt, Rinehart, and Winston Publ., 1968 (Russ.ed.: Moscow, Progress Publ., 1981, pp. 369–495).
23. Russell S., Norvig P. *Artificial Intelligence: A Moderm Approach*. 2nd ed., Prentice Hall Publ., 2002, 1132 p.
24. Jurafsky D., Martin J. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. 2008, 1024 p.
25. *Vyvod v logicheskikh modelyakh. Metod rezolyutsy* [Output in Logical Models. Resolution Method]. Available at: http://www.aiportal.ru/articles/knowledge-models/method-resolution.html (accessed Jule 11, 2016).
26. Boral H., Redfield S. Database Machine Morphology. *Proc. 11th Int. Conf. on Very Large Data Bases*. 1985, pp. 59–71.
27. Fushimi S., Kitsuregawa M., Tanaka H. An overview of the system of a parallel relational database machine GRACE. *Proc. 12th Int. Conf. on Very Large Data Bases*. 1986, pp. 209–219.
28. Tanaka H. *Parallel Inference Engine*. 2000, 296 p.
29. *Open Cognition*. Available at: http://opencog.org/ (accessed June 23, 2016).
30. Link Grammar Parser. *AbiWord*. 2014. Available at: http://www.abisource.com/projects/link-grammar/ (accessed June 20, 2016).
31. *The CMU Link Grammar natural language parser*. Available at: https://github.com/opencog/link-grammar/ (accessed June 22, 2016).
32. RelEx Dependency Relationship Extractor. *OpenCog*. Available at: http://wiki.opencog.org/wikihome/index.php/Relex (accessed June 22, 2016).
33. Sokirko A.V. *Semanticheskie slovari v avtomaticheskoy obrabotke teksta (po materialam sistemy DIALING)* [Semantic Dictionaries in Text Automatic Processing (adapted from DIALING system)]. PhD thesis, 2001, 120 p. (in Russ.).
34. *Avtomaticheskaya obrabotka teksta* [Text Automatic Processing]. Available at: http://aot.ru/ (accessed June 23, 2016).
35. Prószéky G. Machine Translation and the Rule-to-Rule Hypothesis. *New Trends in Translation Studies (In Honour of Kinga Klaudy)*. Budapest, Akadémiai Kiadó Publ., 2005, pp. 207–218.