Онтология для поддержки задач извлечения смысла из текста на естественном языке

Игорь Богуславский Вячеслав Диконов Светлана Тимошенко
ИППИ РАН *ИППИ РАН* *ИППИ РАН*
bogus@iitp.ru *dikonov@iitp.ru* *timoshenko@iitp.ru*

Резюме

Статья посвящена ведущейся в нашей лаборатории работе по построению большой онтологии общего назначения[1]. Создаваемый ресурс предназначен для задач семантического анализа текста на любом естественном языке и специально адаптируется для этого. Единицы онтологии соотнесены со словарями системы ЭТАП-3 и словарем искусственного языка-посредника UNL.

1. Введение

Полвека назад единственной задачей, которая предполагала автоматическую обработку свободно формулируемого текста на естественном языке, был машинный перевод. Задача эта полностью не решена до сих пор. В ходе работы обозначились два возможных подхода к ее решению: «эмпирический» (или data-based) и «эвристический» (основанный на знаниях эксперта). К первому подходу можно отнести техники SMT (Statistical Machine Translation[2]) в чистом виде и EBMT (Example-Based Machine Translation[3]). В рамках этих технологий тексты рассматриваются как цепочки символов, которые нужно оптимальным образом разделить на фрагменты и напрямую соотнести с фрагментами других — иноязычных — цепочек. При этом не делается попытки определить их реальный смысл в данном контексте. Такого рода решения могут быть эффективны за счет накопления большого объема примеров соответствий и использования подготовленных человеком примеров, например фрагментов ранее переведенного вручную похожего текста. Этот принцип используется многими современными системами машинного перевода, например Google Translate. Однако, кажущееся преодоление неоднозначности и вариативности текста основывается не на понимании машиной смысловых взаимосвязей, а на статистике. Результат перевода зависит от степени полноты и представительности используемых корпусов. Следовательно, невозможно гарантировать на практике наличие в обучающем материале достаточного количества правильных примеров, на основе которых компьютер в каждом конкретном случае выявит нужную закономерность.

Эвристический подход характеризуется полным анализом текста, направленным на выявление его морфологической и синтаксической структуры, и преобразованиями полученных структур. Благодаря анализу извлекается дополнительная информация, которая может быть использована для перевода. Например, после морфологического анализа предложения *Глядится тусклый день в стекло прозрачных льдин* нам становится известно не только то, что в него входит фрагмент *стекло*, но что он может быть либо существительным среднего рода в именительном падеже, либо тем же существительным в винительном падеже, либо глаголом в прошедшем времени. Далее, сопоставляя эту дополнительную информацию о разных словах и синтаксическую структуру, система принимает решение о том, какой перевод предпочесть - существительное *glass* или глагол *stream down*.

К системам, использующим этот подход, относится и разрабатываемая нашей лабораторией система ЭТАП. Однако результаты показывают,

[1]Работы ведутся при поддержке РФФИ (гранты №№ 11-06-00405 и 12-07-00663), а также при поддержке РГНФ (грант № 10-04-00040а)

[2] Статистический машинный перевод осуществляется на основе соответствий, которые известны системе по результатам статистической обработки параллельных корпусов.

[3] Машинный перевод по шаблону также использует параллельные корпуса, но они служат базами данных, из которых во время перевода извлекаются нужные куски.

что дальнейшее совершенствование дается ценой все больших усилий из-за возрастающей сложности. Увеличивая объем обрабатываемой информации, мы увеличиваем также число точек, где могут возникнуть ошибки. Упомянутые выше эмпирические подходы — это попытка осуществить перевод, не множа информацию о переводимом тексте. Другой возможный метод борьбы с ошибками, заключается, как ни парадоксально, в дальнейшем увеличении информации о тексте с переходом на уровень семантического анализа.

Подобно тому, как для нахождения возможных синтаксических связей в предложении необходимо знание грамматики языка, для успешного определения значений слов этого предложения и правильных связей между ними необходимо знание мира, который описывается в тексте. Такое знание предоставляется онтологией.

Онтология представляет собой формализацию знаний о взаимосвязях объектов и целых классов объектов реального мира, которая позволяет компьютеру использовать эти знания и даже дополнять информацию об отдельных объектах с помощью логического вывода.

Невозможно точно сказать, когда появились компьютерные онтологии. В 80-ых годах этот термин стал употребляться для обозначения структурных элементов, хранящих знания о мире внутри систем искусственного интеллекта. В 90-е годы развитие онтологий набрало обороты, и исследователи стали предпринимать попытки теоретически их осмыслить и разработать правила их создания. Крупнейшей инициативой в этой области является «семантическая паутина» (Semantic Web), которая предоставляет стандартный набор языков и форматов (OWL, RDF) для создания онтологий, используемых различными и никак не связанными между собой системами обработки данных. В настоящее время в рамках SemanticWeb существует множество онтологий узких предметных областей и варианты нескольких известных онтологий широкого профиля. Однако и те, и другие ориентированы на решение прикладных задач, отличных от обработки естественного языка. Например, онтологии задействованы в приложениях, связанных с управлением знаниями, моделированием бизнес-процессов, интеллектуальной интеграцией информации, информационным поиском, интеграцией баз данных, биоинформатикой, электронной коммерцией и т.д. Для получения полного семантического анализа текстов на естественном языке этот информационный ресурс используется гораздо реже. Между тем, в перспективе он может принести большую пользу. Мы хотим исследовать перспективы применения онтологии при анализе текста.

2. Разрабатываемая онтология

Мы поставили себе задачу создать онтологию, классифицирующую объекты и факты действительности таким образом, чтобы

а) она служила удобным метаязыком для описания значений слов;

б) с ее объектами было удобно соотносить значения слов разных естественных языков (ЕЯ);

в) она предоставляла знания, полезные для разрешения неоднозначности при анализе текста произвольной тематики.

Место онтологии в системе семантического анализа и общее описание алгоритма можно найти в [1],[2].

2.1. Онтология и словари

Информация, помещаемая в лингвистической онтологии, сходна с той, которая содержится в традиционных лингвистических ресурсах. Однако имеются и важные различия. Чем онтология отличается от семантического словаря и от словаря-тезауруса?

Во-первых, предметом описания. Словари-тезаурусы описывают так называемые семантические поля — совокупности близких значений и выражающих их слов. Как правило, очертить их точные границы не представляется возможным. Поэтому составители тезаурусов, стремясь наиболее полным образом представить то или иное поле, рассматривают его максимально широко: лучше приписать лишнее, чем не дописать важного. Толковые словари описывают слова через их значения. Онтологии описывают действительность через существующие в ней объекты, их свойства и взаимосвязи. Классы объектов, как правило, организованы иерархически и образуют дерево. Свойства классов наследуются по дереву сверху вниз, то есть каждый подкласс обладает всеми свойствами класса, в который он входит. Связь между онтологией и словарем состоит в установлении соответствия между элементами онтологии и словами ЕЯ.

Во-вторых, онтология отличается от словаря степенью формализации. Толковые словари

содержат определения, предназначенные для людей. Даже если для формулирования определений используется специально выработанный лингвистами метаязык, он все равно остается адаптированным вариантом естественного человеческого языка. Онтология же предназначена для компьютерной обработки и ее выразительные средства коренным образом отличаются от естественного языка.

В-третьих, онтология обеспечивает возможность логического вывода по четко определенным законам и тем самым способна сообщить об объекте больше, чем записано «внутри» соответствующей единицы («сущности»). Статья словаря может существовать автономно от других статей и не потерять свою информационную ценность, а единица онтологии — нет.

Рассмотрим пример. У русского слова продавец несколько значений. Нам интересны следующие два:

1. Человек, являющийся сотрудником торговой организации и отпускающий товар покупателям. *Продавец положил кулек на весы.*

2. Должность в торговой организации; обязанности человека, занимающего эту должность, заключаются в том, чтобы предоставлять товар покупателям и получать от них деньги в случае покупки. Второе необязательно. *Теперь он работает продавцом.*

Иными словами, это слово может быть применено к очень разным объектам действительности, и поэтому нет смысла создавать в универсальной онтологии единицу, значение которой объединяло бы в себе эти значения русского слова. Можно создать два разных объекта, каждый из которых будет соответствовать одному из значений, но это решение представляется неэкономным, поскольку первое значение может быть истолковано через второе.

Мы предлагаем следующее решение.

Значение 2 соотносится с классом онтологии, который объединяет все должности продавцов в торговых организациях. Назовем этот класс SalesPosition[4]. Внутри онтологии этот класс будет подклассом класса JobPosition, объединяющего все возможные «работы». В русском языке ему примерно соответствует слово *должность*[5].

Класс JobPosition по своему смыслу имеет валентное свойство «в какой организации». Класс SalesPosition наследует это свойство от вышестоящего класса. Т.е. все его члены — должности, связанные с торговлей — тоже имеют свойство принадлежности к какой-либо организации. При этом может быть добавлено уточнение, что местом работы продавца должна быть торговая организация.

Значение 1 на языке онтологии записывается как выражение:

hasSubject[6] (HavingJobPosition, Person)

hasValue (HavingJobPosition, SalesPosition),

то есть человек (Person), имеющий характеристику по месту работы (HavingJobPosition) — продавец (SalesPosition).

Таким образом, русское слово *продавец* соотносится нетривиальным образом с классом SalesPosition. Рассмотрим этот класс подробнее.

Про продавца целесообразно знать, что его основная функция - продавать товар. Добавляем классу JobPosition слот hasFuction. SalesPosition наследует этот слот. Уточняем, что у этого класса слот заполнен: hasFuction only Selling. Однако согласно онтологии, JobPosition является абстрактным признаком, который не может самостоятельно действовать, а значит, объекты класса SalesPosition не могут выступать в роли деятеля. Однако, носители признаков класса JobPosition — люди — могут. Этот компонент толкования может быть выражен с помощью формально записанного логического утверждения следующего содержания:

‘Если человек (*Person*) занимает должность (*HavingJobPosition*), и эта должность имеет некую функцию, выраженную действием (*hasFuction X*), то этот человек (*Person*) играет роль деятеля (*hasAgent*) в X’.

Такие утверждения носят название аксиом и являются составной частью некоторых онтологий.

На примере русского слова *продавец* мы видим, как комбинация хранящихся в онтологии данных о свойствах объектов и аксиом позволяет записать ту же основную информацию, которую предоставляют толкования хорошего семантического словаря. При этом в онтологии

[4] Мы используем английские слова в качестве основы для имен классов и индивидов из соображений удобства, а также в силу ограничений стандарта языка OWL. Это ни в коем случае не означает, что в онтологию попало английское слово вместо русского. У класса с таким именем столько же общего с английским словом, сколько и с русским.

[5] Русское слово «должность» означает скорее руководящую должность — обычные сочетания с этим словом, по данным Ruscorpora: *должность директора / руководителя / начальника*. Выражения типа *?должность вахтера*, *? должность уборщицы*, *?должность дворника* хуже соответствуют нормам русского языка.

[6] Ограниченный объем статьи не позволяет нам подробно обсуждать все цитируемые элементы онтологии, поэтому мы просим читателя считать, что это просто стандартизированные бинарные связи.

смысл представляется строгим, формальным способом, с использованием богатого метаязыка. Можно сказать, что содержимое онтологии – это метаязык для определения значений слов. Причем каждый его элемент определяется только с помощью других элементов той же онтологии, и такое определение является необходимым и достаточным. Если рассматривать онтологию в такой перспективе, то формальный язык (OWL, KIF...), который используется для записи самой онтологии выступает как метаязык второго порядка.

Возможность выражать с помощью метаязыка онтологии законченные определения значений слов означает, что очень подробно и тщательно разработанная онтология может быть функциональным аналогом семантического словаря, обладая важными дополнительными возможностями логического самоконтроля и доступности для прикладных компьютерных применений.

2.2. Энциклопедическая информация

Онтология может включать любую необходимую информацию энциклопедического характера. Например, что вода при 0ºC переходит в другое агрегатное состояние, называемое «лед», а при 100ºC — в состояние, называемое «пар». В принципе, в толковом словаре эта информация иногда присутствует, но она записана не полностью и не последовательно. Одно из значений слова *лед* толкуется в [1] как 'замёрзшая и перешедшая в твёрдое состояние вода', а одно из значений слова *пар* - 'газ, в который превращается вода при нагревании'. В самой статье «вода» ни лед, ни пар не упоминаются.

Рассмотрим словосочетания: (1) в*ода из магазина* и (2) *вода из льда*. Грамматически они совершенно одинаковы, а семантически – нет. В (1) речь идет о первоначальном местоположении, а в (2) – о первоначальном состоянии. Автоматическая система сможет уловить это различие, лишь «зная» о том, что лед – это другое агрегатное состояние воды. В семантической структуре это различие может быть выражено с помощью разных имен связей: (1') hasStartingPoint (Water, Shop) и (2') PreviousState (Water, Ice).

Онтология призвана предоставить значительно больший объем обобщенных знаний о мире, которые помогут компьютеру выбирать наиболее правдоподобную интерпретацию в случае неоднозначности. В совокупности онтологическая информация позволяет перейти от формального представления собственно текста к моделированию описываемой текстом реальности (и даже к логическому выводу об этой реальности).

3. OWL и SUMO

3.1. OWL

В качестве языка для разработки нашей онтологии на данном этапе мы выбрали OWL [4], широко распространенный язык для разработки онтологий. Он является стандартом SemanticWeb, что обеспечивает наиболее широкую программную поддержку, а это важное достоинство.

Минусом OWL без специфических расширений является монотонное наследование при отсутствии стандартных средств описания исключений. Есть понятия, которые по совокупности большинства значимых для человека признаков отнесены к некоему классу, но имеют также и несовместимые с определением класса свойства. За этим часто скрывается расхождение между объективной и «наивной» картинами мира. Примеры — бесконечны: пингвины и страусы — птицы, хотя в рамках «наивной» картины мира птица — это существо, которое летает; с точки зрения биологов, бактерии несомненно живые, а для языка это не так (к предложению *Болезнь вызвана бактериями* нельзя поставить вопрос *Кем вызвана болезнь?*), а с куклами и роботами все наоборот...

Формат OWL не определяет языка для написания аксиом и не предусматривает специально места для их включения непосредственно в онтологию. Для этой цели используются другие языки, такие как SWRL (SemanticWebRuleLanguage).

В рамках нашего проекта, основанного на системе ЭТАП, используется логический язык ЭТАП — ФОРЕТ и семантические словари ЭТАП. При этом аксиомам онтологии соответствуют правила ЭТАП. Правило, осуществляющее перевод русского слова «продавец» во втором значении и сопоставляющее ему целое выражение на языке онтологии, является частью русско-онтологического словаря, в то время как информация о том, что классу SalesPosition соответствует русское слово «продавец» в первом значении может быть записана в соответствующей зоне комментария в онтологии.

3.2. Существующие онтологии

Мы ознакомились с некоторыми онтологиями, открытыми для некоммерческого использования (SUMO [3], GUM [4], PROTON [5], OMEGA [6], OpenCyc [7], FunGram [8], Dolce [9]) Кроме этого мы учли опыт создания коммерческих онтологий (MikroKosmos) [10]. В качестве основы для дальнейшей работы была выбрана онтология SUMO.

При сопоставлении онтологий принимались в расчет следующие критерии:

1. Достаточно полное покрытие концептуального поля.
2. Богатство информации о свойствах классов объектов действительности в терминах их отношений и атрибутов.
3. Связь единиц онтологии с лексической базой данных WordNet.

3.3. SUMO: достоинства и недостатки

Данная онтология написана на языке KIF (Knowledge Interchange Format) [11]. Внутри онтологии выделяются три уровня классификации: верхний, средний и нижний. Верхний и средний уровни составляют базовую часть онтологии, а нижний уровень покрывается разнообразными расширениями. Каждое расширение представляет собой описание отдельной области знаний о действительности. Таким образом, SUMO строится из модулей. Каждый модуль-расширение обязательно «привязан» к базовой части. Кроме того, расширения могут быть связаны между собой, однако это связи необязательные и трудно прослеживаемые. Если бы не эта особенность, можно было бы считать, что SUMO — это не одна онтология, а целое семейство онтологий: взяв базовую часть и необходимые расширения, каждый пользователь может создать онтологию, адаптированную для решения его задач. На данный момент модульный подход возможен, но связан с определенным риском: в производной, «неполной» онтологии могут быть утрачены некоторые связи.

Существенным отличием SUMO от некоторых других онтологий, в значительной степени определившим наш выбор, является то, что ее классы поставлены в соответствие с элементами лексической базы данных WordNet. Поскольку словари системы ЭТАП-3 также частично соотнесены с этой базой данных, то появляется возможность полуавтоматически связывать SUMO со словами естественного языка. Говоря «полуавтоматически», мы имеем в виду, что связи сначала устанавливаются в автоматическом режиме, а затем их достоверность проверяется экспертом.

Надо отметить, однако, что онтология SUMO создавалась в первую очередь для задач логического моделирования. Обработка текстов на ЕЯ также интересовала авторов SUMO. Решая эту задачу, они установили связи между онтологией и синсетами[7] WordNet [12]. Но в сложных случаях, когда затруднительно установить соответствие, решение часто принималось «в пользу онтологии». Так, например, все слова, значение которых связано с оценкой, «привязаны» к одному- единственному классу SubjectiveAssessmentAttribute, причем строгих критериев нет. Например, *difficult* считается SubjectiveAssessmentAttribute, а *dangerous* – нет, хотя никаких лингвистических или психологических доказательств того, что чувство опасности менее субъективно, чем чувство сложности, у нас нет. Некоторые другие существенные для языка различия также сознательно игнорируются, например, различие между единичной сущностью и родом. Поэтому разработка онтологии на базе SUMO разделяется на два этапа:

1) конвертация в OWL-формат (с неизбежными потерями - дополнительные сложности связаны с тем, что не все аксиомы SUMO можно адекватно перевести на язык OWL); возможность перехода обратно к KIF сохраняется;

2) содержательные модификации, приспосабливающие ее к решению задач анализа естественного языка. Следует отметить, что степень сложности большинства онтологий SemanticWeb ниже, так как они описывают сильно ограниченные и очень хорошо формализованные предметные области.

4. Преобразование SUMO

Мы разработали инструмент-конвертер, «переводящий» SUMO с формального языка KIF на принятый на текущей стадии нашего проекта формальный язык OWL. Наш конвертер позволяет преобразовывать как саму базовую часть SUMO, так и базовую часть вместе с произвольным набором входящих в SUMO подчиненных расширений, а также «переводить» на OWL новые расширения и добавлять их в уже имеющуюся

[7]Набор слов-синонимов, которые имеют общее значение при заданном Wordnet уровне детальности описания. Каждый синсет имеет идентификатор, а в рамках SemanticWeb — URI.

онтологию. С помощью этой программы в любой момент может быть получена OWL-версия онтологии SUMO, которая служит основой для дальнейшей разработки. В процессе конверсии она может быть обогащена некоторыми видами дополнительной информации.

Особенностями данного процесса являются а) возможность внесения собственных изменений и дополнений в ходе процесса преобразования и б) возможность изменения формата вывода программы с целью перехода с одной версии OWL на другую или к другому языку представления знаний.

Все изменения и дополнения OWL-онтологии выражаются в виде списков элементарных действий: удалить/добавить/переименовать класс, индивид или OWL-свойство, удалить или добавить связь между объектами онтологии или объектом онтологии и единицей языка. Формирование таких списков позволяет изменять содержание получаемой онтологии при сохранении возможностей импорта новых знаний из SUMO и переноса накопленной информации в другие форматы.

Данное преобразование не является полным, так как исходный язык KIF обладает большими выразительными возможностями, чем OWL. В частности, не подлежат переводу аксиомы SUMO и некоторые предикаты KIF, которым нет прямых аналогов в OWL.

Значительная часть добавляемых OWL-свойств призвана формально отразить ключевые общие признаки объектов, объединяемых классами SUMO. Эта информация присутствует в SUMO в форме словесных определений классов и аксиом (правил для логического вывода) и ее трудно или невозможно адекватно перенести в OWL в автоматическом режиме. Например, один из основополагающих классов в SUMO «Physical» объединяет все объекты и явления действительности, которые имеют определенные координаты в пространстве и времени. В SUMO это формально выражено через аксиому, что существование любого объекта класса «Physical» неизбежно означает существование двух переменных ?LOC и ?TIME. Их значения являются координатами объекта, возвращаемыми KIF-функциями located и time. Однако, в онтологии на языке OWL это следует записать как свойства класса «Physical», так как в ней нет таких выразительных средств, как переменные и функции языка KIF. При добавлении новых свойств особое внимание уделялось тем свойствам, которые выражают валентности предикатов. Здесь под валентностью мы понимаем семантические валентности или набор обязательных участников ситуации.

Наша OWL-версия сохраняет те данные SUMO, которые можно интерпретировать на языке OWL как классы, индивиды и их свойства. Наряду с упомянутыми потерями процесс конверсии также позволяет исправлять некоторые ошибки и обогащать получаемую онтологию новой информацией. В частности, можно автоматически обнаруживать некоторые разрывы связей между классами, часто возникающие при исключении отдельных расширений SUMO, а также случаи ошибочного отнесения значений слов ЕЯ к классам SUMO.

5. Связи с лексикой различных ЕЯ

Важным направлением обогащения онтологии является связывание включенных в нее объектов (понятий) со значениями слов различных естественных языков. Такое связывание необходимо для перехода от слов к понятиям в ходе анализа текста на ЕЯ. Многозначность слов ЕЯ мешает непосредственно связывать каждое слово текста с каким-либо одним классом онтологии, а при связывании многозначного слова с многими классами одновременно свойства этих классов часто начинают противоречить друг другу. Так, например, русское слово «собачка» может означать животное или деталь механизма. Эти два значения относятся к несовместимым между собой онтологическим классам, противопоставленным по признакам «живое» - «неживое», а также «естественное» - «искусственное». Попытка конъюнктивно использовать оба класса при анализе текста, который содержит это слово, немедленно приведет к логической ошибке — объекту с невозможным сочетанием свойств. Дизъюнктивная интерпретация списка возможных классов избегает этой ошибки, но связана с некоторыми ограничениями, так как позволяет рассматривать лишь одиночные классы и не дает возможности использовать альтернативную множественную классификацию (разбиение списка возможных классов на несколько альтернативных групп).

Для нейтрализации противоречий мы выбрали подход, основанный на использовании лексических значений слов в качестве промежуточных единиц, которые могут быть ассоциированы с классами или индивидами онтологии или даже включены в онтологию в качестве объектов на самом низком уровне

иерархии. В роли словаря значений выступает словарь языка UNL [13]. Любое известное в системе значение слова может быть связано с несколькими непротиворечащими классами, что позволяет получить для логического вывода больше информации об упоминаемом объекте действительности за счет объединения свойств этих нескольких классов.

Основная единица UNL — это «универсальное слово» (UniversalWord, UW). Каждое UW представляет собой значение какого-либо слова какого-либо ЕЯ. Таким образом UNL позволяет пользоваться более широким набором значений, чем Wordnet, включая те, которых нет в английском языке. А также позволяет учитывать мелкие различия значений внутри синсета Wordnet. Мы уже построили ранее в рамках других проектов большой словарь универсальных слов, который применяется для перевода и аннотации текстов на нескольких естественных языках. В настоящий момент словарь UNL позволяет связывать с разрабатываемой онтологией и SUMO слова русского, английского, французского, испанского, вьетнамского и малайского языков.

6. Расширение SUMO

Преобразование онтологии SUMO в OWL не является новым достижением само по себе, но необходимо для развития нескольких проектов семантического анализа на базе ЭТАП-3. А вот обогащение полученной онтологии новой информацией, важной для анализа текста, ведет к созданию нового ресурса. Новая информация, которая вносится в порождаемую онтологию, включает в себя новые свойства существующих классов SUMO, соответствия между объектами онтологии и значениями слов ЕЯ а также новые классы объектов.

Заметим, что в рамках разработки системы ЭТАП уже предпринимались шаги по обогащению анализа семантической составляющей. На данный момент в ЭТАПе действует очень простая система семантических дескрипторов — ярлыков, приписанных словам. Практика показывает, что даже такая примитивная система неиерархизованных пометок позволяет значительно улучшить качество синтаксического анализа и перевода.

Онтология представляет собой гораздо более сложную и разветвленную семантическую систему, которая в перспективе займет место этого набора дескрипторов.

6.1. Фундаментальная классификация предикатов

Ведется эксперимент по внедрению в онтологию фундаментальной классификации предикатов (ФКП), разработанной Ю.Д.Апресяном [14]. Данная классификация создавалась для решения практических задач лексикографии. Она является фундаментальной в том смысле, что ее понятия «используются во всех лингвистических правилах — морфологических (...), словоообразовательных, синтаксических, семантических, прагматических, коммуникативно-просодических, сочетаемостных и других» [14: 28]. Иными словами, положение лексемы в этой классификации должно почти полностью определять ее свойства. Однако основания для классификации — строго семантические, и благодаря этому мы теоретически можем соотнести ее с единицами онтологии, представляющими соответствующие смыслы, и обогатить онтологию информацией, релевантной для обработки ЕЯ.

Напрашивающееся решение — поискать прямые соответствия. При первом взгляде на SUMO оно кажется обманчиво доступным: довольно высоко в иерархическом дереве помещен класс Process, обнаруживаются также классы Perception, StateOfMind и некоторые другие, вроде бы близкие по смыслу традиционным семантическим классам ФКП. На самом деле ни один из этих классов не соответствует классам и подклассам ФКП. Process в онтологии — это, дословно, «все, что происходит во времени, но не является объектом», в то время как в ФКП процесс понимается уже, как изменение (возникновение/исчезновение) чего-либо, происходящее без чьего-либо намерения/желания, само по себе. Перебирая подклассы Process в поисках класса, чье значение могло бы соответствовать процессам, как они понимаются ФКП, находим InternalChange. По определению, это процесс, вызывающий изменение собственного свойства объекта, но не процессы, вызывающие изменение отношений объектов, в частности, изменение положения в пространстве или во времени. В соответствии с этим определением, подклассом InternalChange является класс Killing, 'убивать', который с точки зрения ФКП является не процессом, а действием. А вот класс событий, которые могут быть описаны русским глаголом *падать*, никак не может быть включен в InternalChange.

Перебрав множество классов SUMO, мы убедились, что классов, соответствующих классам ФКП, в SUMO нет, поскольку в ней почти не учитывается специфически «языковая» информация, как например, целенаправленность/нецеленаправленность действия. Следующий возможный способ интеграции ФКП в SUMO — добавление недостающих классов. Например, можно добавить класс Action, соответствующий действию ФКП. В рамках ФКП у него выделяются три конституирующих признака:

в вершине ассертивной части толкования на последней ступени семантической редукции обнаруживается семантический примитив 'делать';

время существования ситуации укладывается в один раунд наблюдения;

главную роль играет агент, целенаправленно изменяющий мир.

Убийство удовлетворяет этим условиям, и поскольку в OWL возможно сделать класс подклассом более чем одного класса, ничто не мешает нам подчинить класс Killing, наряду с Destruction, которому он непосредственно подчиняется, еще и новому классу Action. В этом случае Killing унаследует свойства от обоих родителей. Однако тут возникает другая сложность: ФКП по форме – это нестрогая многоуровневая иерархия с многочисленными пересечениями классов. Так, например, *разбить* (*разбить стекло*) относится либо к действиям, либо к процессам, в зависимости от намеренности/ненамеренности агента. Если отнести соответствующий класс к обоим над-классам, то обеспечиваемый форматом OWL монотонный вывод в данном случае приведет к ошибке, потому что будут приписаны взаимоисключающие свойства.

На данный момент самым гибким способом соединения ФКП с онтологией нам представляется следующий: классы ФКП вводятся в SUMO в виде индивидов и используются как «ярлыки». То есть в онтологию вводится класс FundamentalPredicateClass, индивидами которого являются конкретные элементы ФКП, а также свойство belongToFPC, принимающего значения из этого класса.

В том случае, если принадлежность какого-либо класса онтологии к определенному классу ФКП представляется неоспоримой (или целесообразно указать то значение, которое этот параметр принимает по умолчанию), например, что любая болезнь — это процесс, соответствующая информация может быть непосредственно приписана классу. Если же однозначное определение затруднено, или зависит от контекста, создаются правила, дополняющие этим знанием конкретную семантическую структуру.

Рассмотрим два предложения: (3) *Геологи ищут алмазы*[8] и (4) *Эти геологи сейчас ищут алмаз*. В (3) *ищут* - это деятельность по ФКП, а в (4) — действие. За этим различием скрывается различие в смысле предложений; (3) естественно понимать так: 'поиск алмазов является для геологов типичной работой', а (4) значит, что конкретная группа геологов ищет конкретный алмаз. Информация, полученная из (3) при правильном анализе, представляет собой факт, достойный включения в базу фактов, а из (4) при правильном анализе следует, что геологи либо найдут алмаз, либо нет. В (3) почти нет контекста, указывающего, что *ищут* — это деятельность, но вот (4) обладает контекстом, позволяющим довольно надежно определить, что искать в данном случае — действие. Подобные примеры предполагается в дальнейшем обслуживать правилами.

6.2. Дескрипторы словаря лексической игры

Вторым источником лингвистической информации для нас является словарь лексической игры, разработанный Ю.Д. Апресяном [15, 16]. В нем также есть система дескрипторов. Она в несколько раз превосходит систему дескрипторов ЭТАПа по объему — в ней порядка 300 семантических единиц. Но главное ее достоинство заключается в том, что будучи приписанными к слову, дескрипторы образуют не просто пучок ярлыков, а семантическое выражение со своим синтаксисом. Так, первый дескриптор в ряду — это вершина толкования. Например,

NAME: СТРЕЛЯТЬ DES: 'уничтожать', 'оружие', 'расстояние'
NAME: РАССТРЕЛ DES: 'уничтожать'
NAME: РАССТРЕЛИВАТЬ DES: 'уничтожать'
NAME: ОТМЕНЯТЬ DES: 'уничтожать'
NAME: СТИРАТЬ DES: 'уничтожать'
NAME: ЗАЧЕРКИВАТЬ DES: 'уничтожать'

Легко видеть, что слова, имеющие дескриптор 'уничтожать' на первом месте, объединяются в класс, обладающий общими свойствами: главную роль всегда играет разумный агент, всегда имеется объект, который после того, как подвергся описанному действию перестает существовать.

[8] Пример Ю.Д. Апресяна.

Кроме того, если дескрипторов больше одного, то они образуют выражение, которое можно рассматривать как грубое толкование понятия. СТРЕЛЯТЬ — это 'уничтожать с помощью какого-либо оружия, находясь на расстоянии'. Такое определение не позволяет различить *стрелять* и *бомбить*, но для обработки текста его точность удовлетворительна.

Кроме того, задан образец, по которому можно расширять класс. Например, *сжигать* - 'уничтожать', 'огонь'; *рвать* - 'уничтожать', 'части'.

Интеграцию дескрипторов в SUMO мы представляем себе так: дескрипторы, занимающие первую позицию, становятся классами; слова, у которых некоторый дескриптор стоит на первом месте, соотносятся с каким-либо подклассом класса-дескриптора. Остальные дескрипторы заполняют какое-либо свойство.

Сам словарь игры невелик — порядка 3000 единиц, но интегрировав систему дескрипторов игры в онтологию, мы сможем соотнести эти 3000 лексических значений русского языка с классами онтологии, не прибегая к UNL. Последующее сравнение с результатами автоматического связывания значений тех же русских слов позволит выявить и исправить часть возможно накопившихся ошибок.

6.3. Новое по сравнению с SUMO

Важным новшеством является соотнесение значений русских слов с единицами онтологии SUMO. Это позволит использовать SUMO и связанные с ней ресурсы для анализа русского текста. Русские слова Связываются с SUMO посредством словаря UNL. Сохраняемая связь SUMO-UNL также позволяет связать с онтологией слова других естественных языков. Кроме того, некоторую ценность имеет и сам подход к преобразованию онтологии, позволяющий одновременно с извлечением данных из SUMO вносить собственные новые данные и получать новую онтологию из любой версии SUMO без повторного ручного редактирования.

Примеры новой информации, которую мы добавляем в онтологию:

1) информация о том, к каким семантически релевантым классам фундаментальной классификации предикатов относятся элементы SUMO,

2) данные о важнейших составных частях различных объектов действительности,

3) дополнительные знания о валентностях предикатов,

4) дополнительные классы, основанные на системе дескрипторов лексической игры, разработанной Ю.Д. Апресяном.

Литература

[1] I. Boguslavsky, L. Iomdin, V. Sizov, S. Timoshenko. *Interfacing the Lexicon and the Ontology in a Semantic Analyzer.* In: COLING 2010. Proceedings of the 6th Workshop on Ontologies and Lexical Resources (Ontolex 2010), Beijing, August 2010, pages 67–76.

[2] Igor Boguslavsky. Semantic Analysis based on linguistic and ontological resources In: Proceedings of the 5th International Conference on the Meaning - Text Theory. Barcelona, September 8 – 9, 2011. Igor Boguslavsky and Leo Wanner (Eds.), p. 25-36.

[3] С. А. Кузнецов и др. *Большой толковый словарь русского языка*.

[4] http://www.w3.org/TR/owl-guide/

[5] http://www.ontologyportal.org/

[6] http://www.fb10.uni-bremen.de/anglistik/langpro/webspace/jb/gum/index.hm

[7] http://proton.semanticweb.org/

[8] http://omega.isi.edu/doc/

[9] http://www.opencyc.org/cb/welcome

[10] http://www.fungramkb.com/

[11] http://www.loa.istc.cnr.it/DOLCE.html

[12] Raskin, V. and Nirenbburg, S. *Ontological Semantics*. Cambridge, MA: MIT Press, 2004.

[13] http://suo.ieee.org/SUO/KIF/suo-kif.html

[14] Niles, I. and Pease, A. Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. In Proceedings of the 2003 International Conference on Information and Knowledge Engineering (IKE 03), Las Vegas, Nevada, June 23-26, 2003.

[15] Boguslavsky I.M., Dikonov V.G. Universal Dictionary of Concepts MONDILEX First Open Workshop “Lexicographic Tools and Techniques“. Moscow, Russia. October 3-4, 2008. С. 31-41.

[16] Апресян, Ю. Д. *Исследования по семантике лексикографии. Том 1. Парадигматика.* Москва, Языки славянских культур, 2009.

[17] Apresjan, J. *Enseignement du lexique assisté par ordinateur* Dans : Lexicomatique et dictionnairiques: IVes Journées scientifiques du réseau thématique « Lexicologie, Terminologie, Traduction » Lyon, France, 28-30 sept 1995. Montréal, 1996. P. 1 - 10.

[18] Апресян Ю.Д., Дяченко П.В., Лазурский А.В., Цинман Л.Л. *О компьютерном учебнике лексики русского языка* «Русский язык в научном освещении» №2, 2008. с. 48 – 112.