Эксперимент по автоматизированному нахождению правил для разрешения неоднозначности на основе семантических классов значений слов

Диконов Вячеслав Григорьевич
ИППИ РАН
dikonov@iitp.ru

Аннотация

В статье описывается начальная часть опыта по автоматизированному нахождению правил разрешения лексической/синтаксической многозначности, которые могли бы применяться в системе автоматического анализа текста и перевода ЭТАП-3. Основным принципом формирования правил является поиск семантически обусловленной сочетаемости между словами в определенном значении (лексемами). Для этого на основе данных словаря Кросслексика и модифицированной онтологии SUMO, которая задает классы значений слов, формируется набор шаблонов словосочетаний. Эти шаблоны позволяют компьютеру определить, похожи ли встречаемые в тексте синтаксически связанные пары слов на примеры словосочетаний в словаре, и определить значения входящих в них слов. Качество и продуктивность шаблонов оценивается с помощью корпуса Синтагрус. Удачные шаблоны словосочетаний могут быть впоследствии преобразованы в словарные правила системы ЭТАП.

1. Введение

Разрешение лексической и синтаксической неоднозначности является одной из наиболее сложных задач в области анализа естественного языка, при этом ни один из существующих методов не достигает полностью удовлетворительного для практического использования уровня эффективности. Основой проводимого опыта является тот факт, что сочетаемость слов в осмысленном тексте на естественном языке не случайна и имеет закономерности. Большинство слов обладает так называемой семантической валентностью. Например, слово *письмо* в значении «отправляемое почтой письменное сообщение» имеет валентности автора, адресата, содержания, а также ряд типичных определений — свойств материального предмета: на какой бумаге, каким почерком написано, в каком конверте оно отправлено и т. п. В реальном тексте выражающие эти валентности и типичные определения слова имеют высокую вероятность встретиться в том же предложении и быть синтаксически связанными со словом *письмо*. Их совместная встречаемость и синтаксическая связь со словом *письмо* позволяет с некоторой вероятностью заключить, что оно употреблено в данном значении, и отделить от случаев употребления в других значениях, например «система письменности языка».

1.1. Поставленная задача

Наш экспериментальный проект призван определить типичные контексты многозначных слов в их различных значениях и описать их таким образом, чтобы на основе этого описания стало возможным формулировать правила для разрешения лексической и синтаксической неоднозначности в рамках системы ЭТАП-3. При этом используются семантические свойства составляющих контекст слов, которые кодируются ссылкой на класс понятий онтологии SUMO [2]. На данной стадии работы рассматривается только наиболее простой и ограниченный тип контекста — одно слово, которое синтаксически связано с рассматриваемым неоднозначным словом. Предшествующая работа над системой построения семантических графов UNL [6] показала, что такого минимального контекста хватает для разрешения лексической неоднозначности в достаточном числе случаев, чтобы это положительно сказалось на работе системы анализа текста.

Итак, задачей эксперимента является получение набора правил для разрешения неоднозначности в рамках системы ЭТАП-3 путем выбора правильного омонима и повышения приоритетности гипотез синтаксических связей.

1.2. Предшествующие проекты

Следует отметить, что ранее в лаборатории компьютерной лингвистики ИППИ РАН Вадимом Петроченковым была проведена другая работа на основе данных словаря Кросслексика [4]. Ее задачей была статистическая оценка вероятности разных кандидатных синтаксических связей в процессе построения синтаксической структуры. Для этого использовались только наиболее общие синтаксические свойства зависимых слов словосочетаний из Кросслексики. В результате делался выбор наиболее правдоподобного зависимого из числа слов соседствующих в предложении с описанным в Кросслексике словом. Одним из результатов стал файл автоматически построенных синтаксических структур словосочетаний из Кросслексики, при получении которых велся контроль правильности выбора омонима для иллюстрируемого этим словарем главного слова словосочетания из числа лексем комбинаторного словаря системы ЭТАП-3 (КС). Необходимые для этого соответствия между лексемами КС и Кросслексики были установлены вручную Л.Л. Цинманом. Эти соответствия и структуры использованы для проведения нашего опыта.

1.3. Исходный материал

Как уже было сказано выше, иллюстрирующие употребление слов словосочетания из словаря Кросслексика были преобразованы в синтаксические структуры ЭТАП-3. Такие структуры состоят из двух лексем, связанных друг с другом синтаксической связью непосредственно или через предлог. Иллюстрируемая словосочетанием лексема далее именуется ключевой или ключевым словом, а вторая, связанная с ключевой, лексема именуется диагностической или диагностическим словом. Полученный из Кросслексики набор структур может быть дополнен аналогичными фрагментами структур предложений из синтаксического корпуса Синтагрус [1], которые включают в себя те же лексемы, которые описаны Кросслексикой. Собранные таким образом расширенные списки синтаксически разобранных словосочетаний пригодны для дальнейшего анализа, позволяющего выявить общие черты у групп словосочетаний. Искомыми общими чертами в данном случае являются общие для диагностических слов семантические свойства.

Для проведения опыта было составлен список из 1428 наиболее интересных для опыта ключевых лексем. В него вошли лексемы КС обладающие зафиксированным в КС омонимом в пределах той же части речи и связанные с лексемами Кросслексики.

2. Эксперимент

2.1. Порождение шаблонов

Лексемы комбинаторного словаря связаны с классами онтологии SUMO (и некоторыми дополнительно введенными нами онтологическими классами). Классы онтологии объединяют множества лексических значений[1] лексем с общими существенными компонентами определений. Замена диагностического слова в структурах словосочетаний с известной ключевой лексемой на имя соответствующего ему класса онтологии порождает семантический шаблон, которому соответствует не только исходное словосочетание, но целый ряд подобных ему других словосочетаний. Например, заменив в сочетании *партия телевизоров* диагностическое слово *телевизор* на онтологический класс *ElectricDevice* «устройство работающее на электрической энергии», мы получаем шаблон *ПАРТИЯ1 («набор предметов»)* — $\overset{\text{квазиагент}}{\rightarrow}$ *ElectricDevice*, которому также соответствуют ранее не встреченные словосочетания *партия холодильников, партия кондиционеров, партия радиоприёмников* и другие подобные. Аналогичным образом словосочетание *партия коммунистов* порождает шаблон *ПАРТИЯ2 («политическая организация»)* — $\overset{\text{квазиагент}}{\rightarrow}$ *believes*, которому соответствуют сочетания *партия демократов*, *партия националистов* и т.п.

В дополнение к этому есть процедура обобщения шаблонов. Она использует онтологию для поиска таких групп из двух и более шаблонов, где классы диагностических слов

[1] В целях оптимизации синтаксического разбора используемые в комбинаторном словаре системы ЭТАП-3 лексемы часто соединяют в себе несколько лексических значений с сильно отличающимися определениями, но синтаксически трудноразличимых.

непосредственно входят в один и тот же вышестоящий класс онтологии. Если находится такая группа, то в дополнение к составляющим ее шаблонам порождается более общий шаблон, их семантически объединяющий. Например, если на основе примеров *партия телевизоров* и *партия часов* были получены шаблоны *ПАРТИЯ1 («набор предметов») — квазиагент → ElectricDevice* и *ПАРТИЯ1 («набор предметов») — квазиагент → MeasuringDevice*, то процедура обобщения породит шаблон *ПАРТИЯ1 («набор предметов») — квазиагент → Device* «изготовленный разумными существами предмет, который служит необходимым инструментом для чего-либо». Этот новый шаблон будет выявлять много дополнительных словосочетаний, чьи диагностические слова попадают в другие подклассы класса *Device* например *партия автомобилей* (*TransportationDevice*), *партия карандашей* (*WritingDevice*) и т.п.

Этот процесс реализован в специально написанной программе, которая читает файлы с синтаксическими структурами примеров из Кросслексики и/или корпуса Синтагрус и порождает наборы шаблонов по найденным словосочетаниям.

2.2. Возможное применение шаблонов

Полученные шаблоны представляют собой ничто иное как готовые наборы аргументов для трафаретных правил, которые могли бы осуществлять разрешение неоднозначности. Подобные правила уже существуют в системе ЭТАП-3 [5] и используются для построения семантических графов UNL и перевода. Наиболее близким примером может служить трафаретное правило UNL-CONV1.15. Логику его работы можно записать в виде следующего псевдокода:

ПРОВЕРИТЬ

1. Если от рассматриваемой лексемы X исходит синтаксическая связь T1 к другой лексеме Z, и
2. Лексема Z обладает признаком T2, то

ВЫПОЛНИТЬ

1. Заменить лексему X на лексему L.

Здесь T1, T2 и L — переменные аргументы правила, которые определяются в статье словаря КС. Одно и то же правило может вызываться много раз с разными аргументами, что соответствует проверке нескольких шаблонов. Сейчас в качестве признака T2 используются так называемые семантические дескрипторы, которые соответствуют некоторым очень широким классам онтологии, или синтаксические признаки слова Z. Использование гораздо более детальной системы классов онтологии позволяет добиться более высокого результата. Сейчас система ЭТАП имеет встроенное средство для обращения к онтологии с запросом, входит ли лексема в указанный онтологический класс. Система наследования OWL-онтологии позволяет получать положительный ответ, даже если лексема отнесена не непосредственно к запрошенному классу, а к одному из входящих в него подклассов. Аналогичные правила могут использоваться не только для перевода, но и при синтаксическом анализе на этапе, когда существует много конкурирующих гипотез связей. В этом случае вместо замены одной лексемы на другую правила могли бы повышать приоритет связи, и приоритеты ключевой и диагностической лексем.

Здесь следует еще раз отметить, что приведенное правило является частным случаем, так как существуют и другие варианты связи между словами в предложении помимо прямой синтаксической зависимости. Такая связь может быть опосредована предлогами, вспомогательными или модальными глаголами, членами сочинительной цепочки, куда входит одно из слов и пр. синтаксическими конструкциями. Тем не менее, рассматриваемый простейший случай является наиболее массовым.

2.3. Оценка шаблонов

Не все порождаемые компьютером шаблоны правильны и полезны для разрешения неоднозначности. Некоторые из них могут быть результатом ошибок в исходных данных, например неправильного синтаксического разбора примера или неверной семантической классификации. Качество порождаемого набора шаблонов зависит от нескольких ключевых факторов:

1) собственно наличия достаточного числа необходимых примеров и контрпримеров в списке словосочетаний;
   Так, например, в используемом корпусе есть словосочетание *делать бочку* в значении «выполнять фигуру высшего пилотажа `Бочка`», но нет того же словосочетания в значении «изготавливать большой сосуд для жидкостей». В результате, алгоритм вырабатывает шаблон *Making — 1-компл → БОЧКА2 «фигура высшего пилотажа»*, но не

может получить альтернативный шаблон *Making* — *1-компл*→ *БОЧКА1 «сосуд»*, который опроверг бы первый шаблон. После такого обучения на недостаточно репрезентативном корпусе примеров алгоритм всегда будет ошибочно толковать слово *БОЧКА* в сочетаниях *делать бочку, изготавливать бочки, производить бочки* и т. п. как фигуру пилотажа.

2) правильности синтаксического разбора примеров;
   Возникающие при автоматическом разборе примеров ошибки приводят к неверным исходным данным, которые порождают неправильные шаблоны.

3) правильности и оптимальности связей между лексемами диагностических слов и классами онтологии;
   Одной из известных проблем на данном этапе является использование слишком общего класса *SubjectiveAssessmentAttribute* «Субъективная оценка, которая может быть разной у разных субъектов или меняться со временем». Многочисленные оценочные атрибуты порождают много чрезмерно общих шаблонов с классом *SubjectiveAssessmentAttribute*, которые не позволяют надежно делать выбор между лексическими значениями. При этом многочисленность примеров вроде *радикальная партия, просроченная партия, лихая бочка, пузатая бочка* и т.п. дает этим шаблонам высокий рейтинг при ранжировании, принципы которого описаны далее в п. 2.4. В дальнейшем планируется использовать более глубокие деления внутри этого класса.

4) детальности кодирования многозначности слов в самом наборе лексем.
   В ходе этой работы проявляется проблема фактической многозначности лексем КС в ЭТАП-3. Все используемые в ходе опыта исходные данные для порождения и оценки шаблонов представляют собой наборы синтаксических структур, где многозначность кодируется лексемами КС. Однако, многие из них фактически объединяют разные лексические значения. Например, лексема ДОМ может быть использована для передачи значений «жилое здание» - *кирпичный дом*, «влиятельное в обществе или богатое семейство» - *дом Романовых*, «организация» - *торговый дом, дом моды Зайцева*... Эти значения соответствуют совершенно разным и почти несвязанным друг с другом классам онтологии *House/FamilyGroup/Corporation*. Однако, все эти классы оказываются связанными с лексемой ДОМ. При формировании шаблонов с диагностической лексемой ДОМ, например, *дом из кирпича* применяемый алгоритм породит ошибочные шаблоны **Corporation* — *атриб (из)*→ *КИРПИЧ* и **FamilyGroup* — *атриб (из)*→ *КИРПИЧ*. В дальнейшем этим шаблонам будут соответствовать явно неверные связи типа *семья из кирпича*, *родня из кирпича*, *фирма из кирпича* и т.п.

2.4. Ранжирование и фильтрация шаблонов

Поскольку не все полученные шаблоны одинаково полезны, необходимо сначала отсеять наиболее очевидные ошибки, а затем определить, какие из порожденных шаблонов помогают получать правильный ответ на вопрос о значении своего ключевого слова, а какие — нет. Для этого использовались процедуры ранжирования и фильтрации порожденных шаблонов.

Ранжирование позволяет оценить потенциальное качество шаблона. Для этого используются статистическая и справочная информация, которая собирается в процессе порождения шаблонов. Имеется несколько критериев оценки:

1) частотность по отношению к ключевому слову; Программа порождения шаблонов ведет учет, сколько словосочетаний-образцов породило один и тот же шаблон при данном ключевом слове, и запоминает, какие диагностические слова встретились среди примеров. Чем больше примеров подкрепляют шаблон, тем более вероятно, что он окажется полезным. Этот критерий способен отсеивать случайные единичные ошибки, но для его надежного применения требуется очень большой корпус. В имеющемся корпусе примеров редкие, но хорошие примеры своей частотностью не выделяются на фоне ошибок.

2) частотность по отношению к диагностическим словам; Если оказывается, что одно диагностическое слово порождает несколько шаблонов, но один шаблон

подкрепляется только одним этим диагностическим словом, а другой еще несколькими другими диагностическими словами, то последний считается более предпочтительным.

3) семантическая близость шаблона к другим шаблонам. Для ее оценки используются предоставляемые онтологией связи между классами понятий, которые объединяют разные диагностические слова. Программа подсчитывает число онтологических классов, которые объединяют классы диагностических слов в двух и более шаблонах. В отличие от процедуры обобщения шаблонов учитываются классы, которые отстоят на два и более шагов от цитируемого в шаблонах, но не принадлежат к трем верхним уровням онтологической иерархии. Это позволяет учесть более слабое сходство, которого недостаточно для порождения обобщающего шаблона.

На основании этих критериев подсчитывается общий рейтинг шаблона, который в дальнейшем может быть использован для предпочтения одного шаблона другому, если их применение дает противоположные результаты.

Фильтрация шаблонов заключается в отбрасывании тех из них, которые приводят к чрезмерному количеству ошибок. Даже вполне корректные шаблоны оказываются общими для нескольких разделяемых в наборе лексем значений. Например, шаблон *РАБОТА1* $\leftarrow^{1\text{-компл}}$ *IntentionalProcess* соответствует примерам *завершать работу, спешить на работу, продолжение работы, прервавший работу, совмещавший работу, предлагать работу....* Некоторые из этих словосочетаний неоднозначны в рамках синтаксических структур ЭТАП, так как КС различает лексемы РАБОТА1 «деятельность» и РАБОТА2 «должность». Словосочетание *предлагать работу* может быть понято как «предлагать трудоустройство на должность X» или «предлагать включиться в деятельность X», а также «предлагать результат деятельности», как во фразе *Стыдиться предложить работу к торгам, потому что художник - творческое начало, духовное*. Последнее значение, а также значение «место, где происходит трудовая деятельность» *Пришел на работу в восемь утра*, вообще не выделяются в качестве отдельных лексем КС (как и в примере со словом ДОМ выше).

Чтобы выявить вредные шаблоны, весь список шаблонов применяется для разрешения неоднозначности в предварительно размеченном корпусе, где искомые ответы уже известны. Это позволяет подсчитать число правильных результатов и ошибок для каждого шаблона с их процентным соотношением — индексом надежности шаблона. Фильтрация выполняется с помощью отдельной программы. По умолчанию отбрасываются все шаблоны, которые дают свыше 20% ошибок. Те шаблоны, соответствий которым так и не было найдено в контрольном материале могут быть отброшены или оставлены в зависимости от рейтинга.

Поскольку число шаблонов прошедших фильтрацию относительно невелико, становится возможным ручной контроль. Его можно осуществлять как до формирования правил в словаре, просматривая список шаблонов, так и после, сравнивая результаты разбора и перевода корпуса текстов до и после внесения правил в словарь и прослеживая возможные ухудшения.

3. Результаты

Для обучения и оценки результатов использовался стандартный метод разделения размеченного корпуса образцов на две части. Одна — обучающая использовалась для дополнения списка примеров из словаря Кросслексика и повышения надежности рейтинга шаблонов, а вторая — контрольная применялась для измерения результатов.

В результате автоматического порождения получено 22276 шаблонов для 1230 лексем КС. После фильтрации с отбрасыванием всех шаблонов, которые порождали свыше 20% ошибок, и тех, соответствий которым в корпусе найдено не было, удалось выделить 2802 действенных шаблона для 392 лексем КС. Отобранные шаблоны срабатывают на примерах, которые встречаются в корпусе и обеспечивают правильный выбор лексемы КС ключевого слова с не менее чем 80% успеха.

Применение этих шаблонов на контрольной части корпуса обеспечивает точность 96,2% правильного разрешения неоднозначности. Отклик оказывается небольшим и составляет 24,5% числа словосочетаний, содержащих одно из 392 ключевых слов отобранных шаблонов и 40,3%, если исключить словосочетания, диагностическое слово которых не было отнесено ни к одному классу онтологии, что делает

невозможным сравнение с шаблонами. Для сравнения, применение тех же самых шаблонов к обучающей части корпуса дает точность 98,36% и отклик до 55,51% онтологически классифицированных диагностических слов.

4. Заключение

Описанный метод не является новым сам по себе. Использование правил для разрешения неоднозначности стало исторически самым первым способом решения этой задачи и применялось уже в Джорджтаунском эксперименте, а применение онтологий для классификации лексических значений является широко известным и опробованным подходом в рамках «Онтологической семантики» [3] и систем машинного перевода на основе онтологий.

Тем не менее, проделанная работа имеет ценность так как позволяет решать задачу стандартными средствами действующей системы анализа текста ЭТАП-3. Для применения этого подхода достаточно внесения новой информации в словарь КС. Описанный подход позволяет обучать систему путем автоматизированного накопления данных с помощью внешних инструментов. Накопленная в результате обучения информация выражена в понятном человеку виде и может быть проконтролирована лингвистами как до, так и после помещения в базу знаний. Существуют неиспользованные на данный момент возможности увеличения отклика за счет анализа более сложных синтаксических конструкций и более сложных контекстов с несколькими диагностическими словами. Также возможно повышение детальности классификации значений внутри недостаточно глубоко проработанных классов, прежде всего SubjectiveAssessmentAttribute.

References

[1] I. Boguslavsky, L. Iomdin, S. Timoshenko, T. Frolova - *Development of the Russian Tagged Corpus with Lexical and Functional Annotation.* // Metalanguage and Encoding Scheme Design for Digital Lexicography. MONDILEX Third Open Workshop. Proceedings. Bratislava, Slovakia, 15-16 April, 2009. ISBN 978-80-7399-745-8. P. 83-90 (соавторы: S. Timoshenko, I. Boguslavsky, T. Frolova).

[2] A. Pease., (2011). *Ontology: A Practical Guide. Articulate Software Press, Angwin*, CA. ISBN 978-1-889455-10-5.

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

[4] И.А. Большаков, *КроссЛексика – большой электронный словарь сочетаний и смысловых связей русских слов*. // Комп. лингвистика и интеллект. технологии: Труды межд. Конф. «Диалог 2009». Вып. 8 (15) М.: РГГУ, 2009, с.. 45-50..

[5] Ю. Д. Апресян, И. М. Богуславский, Л. Л. Иомдин, А. В. Лазурский, Н. В. Перцов, В. З. Санников, Л. Л. Цинман. *Лингвистическое обеспечение системы ЭТАП-2*, Москва, Наука (1989).

[6] В.Г. Диконов, *Развитие системы построения семантического представления текста с использованием языка-посредника UNL на базе лингвистического процессора ЭТАП-3* // Информационные технологии и системы (ИТиС’08). Сборник трудов 31-ой Конференции молодых ученых и специалистов ИППИ РАН. Геленджик, 27 сентября – 4 октября 2008 г. М., 2008. С. 195-200. ISBN 978-5-901158-08-01.