КОРРЕКЦИЯ СМЫСЛОВЫХ ОТНОШЕНИЙ КАК ЭТАП СЕМАНТИЧЕСКОГО АНАЛИЗА (НА МАТЕРИАЛЕ КРИМИНАЛЬНЫХ СВОДОК)

CORRECTION OF SEMANTIC RELATIONS AS A STAGE OF SEMANTIC ANALYSIS

Ермаков М.В. *(mermakov@gmail.com)*
Российский государственный гуманитарный университет

Одной из важных проблем автоматического анализа текста является переход от его семантического представления к концептуальным структурам, имитирующим знания. Предлагается коррекция семантических отношений как способ такого преобразования. Рассматриваются возможные правила этого этапа анализа.

1. Постановка задачи

Данная работа посвящена одному из важнейших с точки зрения информационно-лингвистической модели (см. [Леонтьева 2006]) этапу автоматического анализа текста, а именно, переходу от лингвистических структур к концептуальным. С точки зрения этой принятой нами модели такой переход необходим, поскольку в отличие от собственно лингвистических структур концептуальные структуры не привязаны к схеме развёртывания текста, а собирают в виде специальных единиц и объектов только информацию, важную для пользователя.

Мы также используем сопутствующие информационно-лингвистической модели информационный язык-посредник и русский словарь анализа «Руслан» (см. [Леонтьева 2001]).

Лингвистической структурой, с которой нужно работать, должно быть уже некоторое семантическое представление. В настоящий момент существуют системы, автоматически строящие семантическое представление для отдельного предложения: например, алгоритм, изложенный в работе [Сокирко 2001] и функционирующий на сайте www.aot.ru. Далее эту систему будем называть процессором АОТ.

Необходимо отметить, что идея анализа, представленная в модели «Смысл-Текст» [Мельчук 2000], оказавшая огромное влияние на отечественные системы автоматической обработки текста и положившее начало самым известным из них, при практической реализации претерпела серьёзные изменения, особенно в области семантического этапа. Одной из таких практических систем стала система ПОЛИТЕКСТ, предшественник процессора АОТ, см. [Леонтьева 1995].

В результате работы процессора АОТ строится первичное семантическое представление (СемП I), которое пока далеко от идеала, но находится в рамках информационно-лингвистической модели и соответствующего метаязыка, поэтому и используется нами как исходный объект для дальнейшей обработки.

Первичный семантический анализ даёт довольно общее «понимание» текста, которое нужно уточнять, поэтому дальнейшая обработка представляется как ряд коррекций. Мы полагаем, что в этом есть преимущества, так как не стоит сразу чрезвычайно подробно разбивать значения каждой лексемы и устанавливать связи, которые могут проясниться лишь при прочтении всего текста. Более того, например, слишком тонкое разбиение значений лексем не подходит для обработки естественного текста, в котором не всегда можно выделить такие значения.

Итак, наша общая задача – показать, какие коррекции первичной семантической структуры понадобятся, какой переход необходимо сделать, чтобы приблизиться к концептуальным структурам, которыми владеет человек-пользователь. С этой точки зрения нами было рассмотрено около 180 криминальных сводок МВД России на сайте http://old.cry.ru/ (в основном 2000-2004 гг.), а также использовались некоторые другие материалы, например, примеры из газет или из новостных сайтов. Все они были проанализированы с помощью процессора АОТ.

Лексемы предметной области «Преступления» описывались нами в словаре русского анализа, также был составлен вариант её концептуальной структуры. Таким образом, наша задача конкретизирована для текстов данной предметной области (ПО).

Надо сказать, эта задача сходна с задачей систем извлечения частной информации (Information Extraction, далее ИЧИ), которые при обработке текстов конкретной ПО пытаются отбросить лишние сведения и оставить только информацию о важных (для пользователей систем) событиях и их участниках. Системы ИЧИ обычно использовали шаблоны для распознавания нужных им объектов, однако при поиске событий всё больше исполь-

зуют упрощённые виды анализа, в основном синтаксического, работают с размеченными корпусами текстов – см. [Grishman 2003]. В отличие от них, мы используем семантический процессор, и наш подход немного ближе к методу анализа, основанного на правилах. Мы считаем, что без семантического анализа, может быть, и возможно извлечение событий, но велика вероятность неправильного их понимания.

Возвращаясь к используемому нами процессору АОТ, конечно, нельзя сказать, что он учитывает все необходимые термины и устойчивые словосочетания для текстов нашей предметной области «Преступление», а также всегда правильно сегментирует предложение и собирает группы в кавычках, в скобках и т.д. Но все эти недостатки известны, и большинство из них могут быть устранены еще до уровня семантического анализа, как показывают работы Т.Ю.Кобзаревой, результатами которой мы тоже пользовались – см. [Кобзарева 2004].

Результаты работы процессора АОТ, на наш взгляд, показывают необходимость использования при дальнейшем семантическом анализе разного рода коррекций, затрагивающих семантические отношения (СемО) и учитывающих их особенности. Таким образом, нужна своего рода грамматика СемО. Грамматика СемО – не единственное, что требуется для достижения хороших результатов при работе лингвистического процессора, но это центральный компонент, отвечающий за преобразования СемП I в нужном направлении.

Чтобы перейти к описанию правил постобработки СемП I, нам надо сначала остановиться на нашем понимании самого СемО.

2. О нашем понимании семантического отношения

Семантическое отношение записывается как R в формуле вида R(A,B), где A и B – узлы, вступающие в данное отношение. Любое семантическое отношение направленное, главным узлом считается B (из него выходит стрелка). Формула читается как «A находится в отношении R к B» или «A является R для B», например: Функция (рисование, художник) – «Рисование является функцией художника», Инструмент (топор, рубить) – «Топор является инструментом действия *рубить*» (мы будем писать R с большой буквы).

Семантических отношений ограниченное количество (около 80), многие из них общепризнанны как семантические роли, или падежи, например, Агент, Объект, Время и т.п. Узлов, которые могут вставать на место A или B – открытое множество, неограниченное, от любого смыслового элемента до составных семантических групп. Например, Содержанием претензии налоговых органов к человеку может быть несколько ситуаций, которые в свою очередь имеют много участников: Содержание ({Ситуация$_1$, Ситуация$_2$, …, Ситуация$_n$}, претензия). Как мы видим, семантические отношения могут включать в качестве участников множественные узлы.

Два семантических отношения могут быть близкими по смыслу друг к другу, одно может являться своего рода гиперонимом для другого (например, отношение Инструмент является разновидностью отношения С-помощью). Назначение и Цель отличаются друг от друга тем, что назначение присуще предмету, а цель – ситуации; и то, и другое описывает будущие события. Также возможны конверсивы: Причина является конверсивом Следствия.

Любой семантический узел – A или B – имеет собственные смысловые характеристики (СХ), которые нужны в частности при обработке СемО (далее в формулах мы будем писать СХ с большой буквы, как и СемО, различая элементы языка-посредника и лексемы русского языка).

Формулу вида R(A,B), где места R, A и B заполнены конкретными единицами, мы называем элементарной ситуацией, или ЭСит. Действительно, перед нами минимальное, бинарное отношение между двумя участниками, простейшая ситуация.

При анализе текста важно учитывать самые разные сведения о семантических отношениях, поэтому в грамматике СемО даются все синонимические, гипо- и гиперонимические связи конкретного СемО, информация о конверсивах и т.д. Приводятся сведения об A и B, входящих в отношение – их возможные смысловые характеристики. Также с помощью ЭСит записываются логические выводы, если их можно сделать для довольно общих R, A и B. Благодаря разнообразию семантических полей в словаре Руслан мы можем описывать СемО как единицу данного словаря.

Основными результатами данной работы по описанию СемО стали сегменты уточненной грамматики семантических отношений и правила постобработки первичной семантической структуры.

В приложении мы привели фразу, её СемП I, построенную процессором АОТ, и структуру, получившуюся в результате применения предложенных правил коррекции к СемП I.

3. Правила постобработки первичного семантического представления

Первичное семантическое представление состоит из узлов-лексем, связанных семантическими отношениями, и представляет собой ориентированный граф (не всегда связный). Правила постобработки СемП I имеют вид

импликации. Их можно разделить на следующие виды:

- правила уточнения (коррекции) семантических отношений
- правила установления кореферентности между двумя элементами структуры (например, между двумя участниками разных ситуаций)
- правила преобразования СемП I (связанные с введением и удалением элементов структуры)
- правила логического вывода
- другие правила (основывающиеся на синтаксических и иных принципах)

Здесь мы более подробно остановимся только на первых двух типах правил, имеющих наибольшее отношение к СемО.

Все импликации достаточно сложны для программной реализации, так как зачастую используют не только грамматику СемО, но и элементарные ситуации, связанные с конкретными лексемами, т.е. ЭСит, зафиксированные в словаре.

На настоящий момент больше всего правил, связанные с конкретными лексемами – их можно записать в словаре, и почти у каждого полнозначного слова, описанного нами, были выделены такие правила (всего около 300). Количество правил уточнения СемО тоже прямо пропорционально числу самих СемО, так как почти каждое из них имеет гипонимы или гиперонимы, в которые может переходить при определённых обстоятельствах. Таких правил сейчас около 60. Меньше всего общих правил установления кореферентности и правил логического вывода, не зависящих от конкретных слов (около десятка в общей сложности). Также существует сценарий-последовательность совершения преступления и проведения следственных и других действий, связанных с ним. К сожалению, многие правила остаются пока недостаточно точно формализованными, что мешает привести их здесь.

Итак, приведём примеры. Ниже даны некоторые правила – сначала в текстовом виде, а потом в виде формул (когда это возможно). Также приведены примеры предложений, в которых можно использовать данные импликации.

Примеры правил уточнения СемО (эти правила входят в грамматику СемО):

1. Если коммуникативное действие (сообщение и т.п.) В происходит в организации А и Автор не указан, то Автором может быть объявлена эта организация.

   Локализация (А, В) & Автор (А1?, В) => Автор (А, В) / СХ(А) ∋ Организация и СХ(В) ∋ Действие, Коммуникативное.

   Здесь и далее вопросом отмечены неизвестные участники ситуаций, за знаком «/» следует перечисление условий применения правила, знак «∈» обозначает включение смысловых элементов в семантические характеристики какого-л. участника.

   (1а). *О поимке преступников сообщили в пресс-службе МВД.* (= МВД сообщил…)

   (1б). *В МИДе никак не отреагировали на подобные заявления* (= МИД не отреагировал…)

2. Если А = Время события В и А является интервалом времени, то В происходит В-течение времени А.

   Время (А,В) => В-течение (А, В) / СХ(А) ∋ Интервал, Время

   (2). *Он искал следы 3 часа*. (= Он искал следы в течение трёх часов)

3. Если А есть ситуация и А занимает валентность Агента при предикате, то Агент заменяется Причиной.

   Агент (А, В) => Причина (А, В) / СХ(А) ∋ Ситуативное

   (3а). *Неизвестная болезнь убила уже пять человек*. (= Причина (болезнь, смерть))

   (3б). *Ветер сдвинул тележку с места*. (= Причина (ветер,сдвигать))

4. Если А = Конечная точка перемещения некоторой информации, и А – одушевлено, то Конечная точка уточняется как Адресат информации.

   Конечная-точка (А, В) => Адресат (А, В) / СХ(В) ∋ Информация & СХ(А) ∋ Одушевлённое.

   (4). *Информация дошла до оперативного штаба*. (= Адресат (оперативный штаб, информация))

5. А постоянно «Локализован» в организации В, значит, Локализация переходит в отношение В-составе.

   Локализация (А, В) ∧ Конечная-точка (А, В) => В-составе (А,В) / СХ(В) = Организация

   (5). *Этот преступник входит в N-скую группировку.* (= В-составе (преступник, группировка))

Примеры правил установления кореферентности между двумя элементами СемП I:

Установление кореферентности между двумя элементами семантического представления зачастую основы-

вается на связях между актантами одного или нескольких лексем, на свойствах развития текста, например:
1. В рамках предложения две ситуации, связанные причинно-следственными связями, часто сохраняют одних и тех же участников. Например, если ситуация 2 есть результат ситуации 1, и у одной ситуации отсутствует актант по валентности с семантической ролью X, а вторая ситуация обладает такой же валентностью X, то реализации этих валентностей совпадают. Ещё одно необходимое условие – то, чтобы участники любой из ситуаций не имели выражения в предыдущих предложениях. В следующем примере ситуации «операция» и «изъятие» имеют одинаковых деятелей – сотрудники Госнаркоконтроля.

(6). *Более 830 г героина было изъято сотрудниками управления Госнаркоконтроля России по г.Москве в результате успешной операции в одном из районов столицы*.

Прямо не относятся к грамматике СемО преобразования СемП I, зависящие от конкретной лексемы. Например, у конкретной лексемы могут быть нестандартные семантические отношения между актантами, либо она сама может быть преобразована в какое-либо СемО. Таких не общих, единичных правил очень много, они используют элементарные ситуации, связанные с конкретным словом, и описываются словарно.
2. Для слова *орден*: название ордена отражает свойство того, кто награждается данным орденом.

Если Принадлежность (орден, A1) и Идентификатор (A2, орден) и CX(A2) ∋ Хорошее, Параметр, то Параметр (A2, A1).

(7). *За спасение детей он был награждён орденом мужества*.

3. Для слова *сопряжённый*: если A сопряжено с B, то A – вместе с B (т.е. эта лексема переходит в семантическое отношение Вместе)[1].

Если Первый_актант (X, сопряжённый) и Второй_актант (Y, сопряжённый), то Вместе (X, Y)

(8). *Криминалистическая характеристика преступлений, сопряженных с вооружённым нападением*.

Примеры других правил:

Правила логического вывода для конкретных лексем могут быть примерно такими:
а) По умолчанию действие с указанным объектом совершает тот, кому указали.
б) Если некто – адресат какого-либо предмета, то в скором будущем этот предмет будет ему принадлежать.

Главное правило удаления узлов СемП I связано с их информационным весом: если узел имеет малый информационный вес или является отношением либо лексической функцией, в его словарной статье должно быть указано необходимое правило преобразования структуры.

4. Заключение

Наша работа показала необходимость дальнейшего исследования и использования свойств семантических отношений при коррекции семантического представления. Были приведены примеры, дающие возможность считать, что при наполнении семантического словаря и развитии грамматики СемО такие коррекции достаточно точны, их можно формализовать. В результате достигается более глубокое «автоматическое» понимание текста. Семантическое представление становится более близким к концептуальным структурам.

В ходе работы в очередной раз стало ясно, что при анализе необходимо опираться на свойства развития текста. «Неизвестных» участников ситуаций в отдельном предложении обычно можно найти раньше в тексте, и, как правило, актуальные герои из предыдущих фраз с большим приоритетом занимают это «вакантное» место, чем словарные значения по умолчанию. Нельзя изменить тему текста (в том числе объекты и ситуации, о которых идёт речь), явно этого не показав, поэтому очень важно отслеживать такие преобразования, сохраняя эту информацию. Для дальнейшего улучшения семантического анализа необходимо, чтобы его сферой действия стал весь связный текст.

Список литературы

1. Кобзарева Т.Ю. Принципы сегментационного анализа русского предложения // Московский Лингвистический Журнал. М.: РГГУ, 2004. Т.8, №1, С. 31-80.
2. Леонтьева Н.Н. К теории автоматического понимания естественных текстов. Часть 2. Семантические словари: состав, структура, методика создания. М.: Изд-во МГУ, 2001.
3. Леонтьева Н.Н. Автоматическое понимание текстов: системы, модели, ресурсы. М.: Издательский центр “Академия”, 2006.
4. Леонтьева Н.Н. Политекст: информационный анализ политических текстов // НТИ. Сер. 2, № 4. М.,1995
5. Мельчук И.А. Опыт теории моделей «Смысл - Текст». Изд.2. М.: Языки рус. культуры, 2000

6. Сокирко А.В. Семантические словари в автоматической обработке текста (по материалам системы ДИАЛИНГ). Диссертация на соискание уч. степени к.ф.н. М., 2001. http://www.aot.ru/docs/sokirko/sokirko-candid-1.html
7. Grishman R. Information Extraction // The oxford handbook of computational linguistics / Ed. by Ruslan Mitkov. Oxford etc: Oxford university press, 2003, P. 545–559.

Интернет-источники

1. Процессор семантического анализа АОТ: http://www.aot.ru/demo/graph.html
2. Сводки МВД России: http://old.cry.ru/; http://www.cry.ru
http://old.cry.ru/theme.shtml?Theme=%D1%E2%EE%E4%EA%E8%20%CC%C2%C4

Приложение. Первичное семантическое представление фразы и корректированная структура.

(9). *Милиция принимает срочные меры к розыску и задержанию преступников, решается вопрос о возбуждении уголовного дела*.

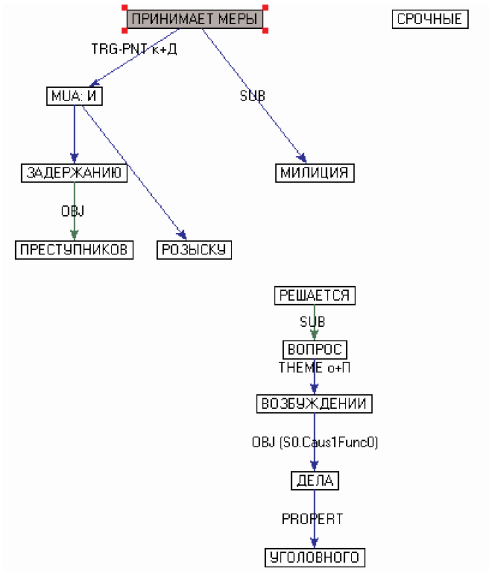

Рис. 1. СемП I примера (9а), полученное с помощью процессора АОТ

В результате применения правил постобработки к СемП I получаем структуру, которую можно прочитать так: ‘Милиция сейчас разыскивает преступников, а потом задержит их, одновременно возбуждается уголовное дело’.

Данная структура довольно сильно отличается от предыдущей, поскольку разбита на ситуации, для каждой из которых отдельно указаны её валентности и модификаторы. В ней предикации получили статус ситуаций, выраженных конкретными лексемами. Вопросом выделены неизвестные участники ситуаций. Жирным шрифтом выделены сведения, полученные в результате вывода. Одинаковым подчёркиванием отмечены одинаковые участники.

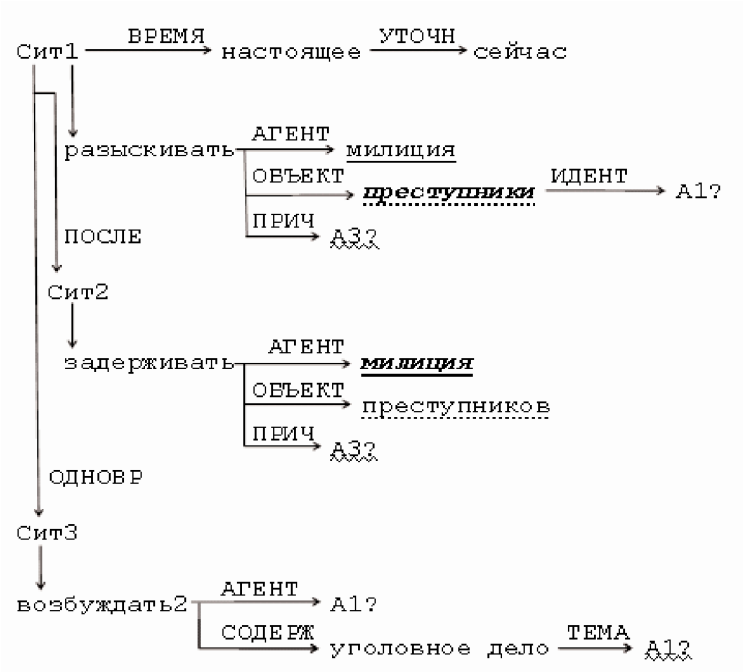

Рис.2. Результат постсемантической обработки