ОБ УРОВНЯХ И ОЦЕНКЕ ТЕКСТОВОЙ СМЫСЛОВОЙ НЕПОЛНОТЫ

ON THE LEVELS AND EVALUATION OF SEMANTIC INCOMPLETENESS OF THE TEXTS

Леонтьева Н.Н. *(leont-nn@yandex.ru)*
Научно-исследовательский вычислительный центр МГУ им. М.В. Ломоносова

Семантическое представление (СемП) эксплицирует связность и избыточность как глобальные свойства текста, а также локальную смысловую неполноту. Все они влияют на оценку информационных качеств текста и объясняют его компрессию. Сжатое СемП (Знания текста) включает и не снятые вопросы (его Незнания).

1. Общие установки

Проблему **смысловой неполноты** и ее роли в оценке информационных свойств текста мы обсуждаем пока умозрительно, но в ожидании относительно «полного» или стремящегося к нему семантического анализа (СемАн) текста. Полный анализ нужен прикладной системе для обеспечения одной из ее важнейших функций – возможности *сравнивать* разные *тексты по содержанию*. Без этого трудно оценить *информационную значимость* каждого текста и выйти к актуальной задаче *накопления знаний* из самих текстов. С точки зрения этих безусловно семантических задач мы рассмотрим способы представления содержания всего текста и роль неполноты. Результаты разных типов анализа в этом направлении будем по традиции именовать **семантическими представлениями** (СемП).

2. О неполноте текстовых структур

Из отечественных подходов, стремящихся к автоматическому построению СемП, самую полную структуру моделирует теория «Смысл-Текст» и ее реализации, в частности система ЭТАП, строящая СемП на основе пофразных синтаксических структур. В ней максимально учитываются все лингвистические нюансы, но структура целого текста как последовательность СемП отдельных предложений неполна, пока не установлены межфразовые отношения и кореферентные связи по всему тексту. К тому же детальное и специфическое для каждого текста деление на фразы и описание их внутренних связей делает сравнение СемП целых текстов очень сложной, если не невозможной задачей. Накопление концептуальных знаний из лингвистических структур – проблема, которую придется решать, видимо, отдельно для каждой предметной области (ПО).

В системах экстрагирования знаний (Information extraction и др.) в роли СемП выступает частичная база данных (БД) с заданной структурой. Из текста или текстовых корпусов извлекаются описания немногих типов объектов (географические имена, персоны, организации и т.п.), а к сборке нужных единиц привлекаются достаточно простые лингвистические знания, в основном графематического и морфологического уровней. В простых случаях, если задан соответствующий массив, подобные методы эффективны, см. [1]: в целом и сжатом СемП текста оставлена только «полезная» информация, ее можно формально сравнивать с другими БД по составу и содержанию полей. Но если в документе есть сведения, не ложащиеся в поля заданной анкеты, они будут утрачены. На произвольном массиве текстов потери существенно превысят полезную часть.

Системы традиционного информационного индексирования опираются на тезаурус терминов и иногда доходят до построения СемП в виде иерархии терминов и тематических линий в целом тексте (как, например, в системе РОССИЯ). Такие данные хороши для сравнения и первичной сортировки текстов, для поиска информации в больших массивах. Но в таких структурах не отражены синтагматические связи между единицами, поэтому пользователь должен вчитываться в сами тексты, чтобы понять их главное утверждение и оценить информационную значимость.

Множество разных по составу и разноплановых структур для одного и того же текста характеризуют его содержание гораздо полнее, чем каждая из них, взятая в отдельности. К этому множеству отнесем и часть СемП, отображающую внешние и композиционные характеристики текста, в нашей терминологии это «внешняя дескрипция» (в корпусной лингвистике она описывается очень подробно), а также результат так называемого «сквозного» анализа [3], устанавливающего логические связи между фразами или более крупными фрагментами

текста. Но содержание каждого фрагмента нужно анализировать отдельно.

Каждое из упомянутых СемП отражает лишь часть, обычно один аспект, содержания нормального связного текста. Смысловая неполнота каждого СемП обусловлена уже самой постановкой задачи: интерпретировать только синтаксические связи в предложении, найти всю терминологию и только ее, выбрать только упоминания о таких-то объектах и их признаках, и т.п. «Полная» структура текста ***многомерна***, даже если не учитывать множество следующих интерпретаций текста пользователями.

3. Ещё одно «измерение» текста

К множеству СемП, отражающих какой-то аспект позитивного содержания (или знания) текста, можно добавить еще одну структуру, которая фиксирует и то, что сам текст «не знает». Ее назначение – облегчить связь лингвистических структур с концептуальными. (Концептуальными назовем те единицы и выражения, которые можно встретить как имена полей разных БД и заполнителей этих полей, устоявшиеся термины разных ПО, а также принятые в специальных текстах именования объектов и процессов). Считаю, что отображать в СемП в явном виде внутреннюю смысловую неполноту («незнание») полезно с разных точек зрения.

Рассмотрим сначала неполноту поверхностного уровня. Локальная смысловая неполнота имеется практически в любом предложении связного текста. Понятия *связный текст* и *локальная неполнота* неразделимы: именно неполнота разных видов (обычно игнорируемая при автоматической обработке текста) есть тот краеугольный камень, без которого не существует нормальный связный текст. Здесь нам придется остановиться на особенностях того семантического метаязыка, который позволяет записывать неполное знание и строить единицы, близкие или совпадающие с единицами внетекстовых знаний. Основу этого метаязыка составляет список двухместных семантических отношений вида P (A,B).

СемАн должен находить связи (P), или **семантические отношения** (СемО), между единицами A и B по всему тексту. Единицами A и B могут быть «лексемы» или более сложные **семантические узлы** (СемУ): термины, свободные словосочетания; имена, номера и символы формул, целых ситуаций, отсылок к внетекстовым единицам. Тем самым СемП оказывается сложной, многомерной **семантической сетью, пространством формул-триад**, которые могут соединять разносортные и разноуровневые понятия**.** Семантическая категория словарной единицы зависит от того, переходит ли она в узел, в отношение или в какой-то признак/параметр того или другого. Базовая составляющая такой сети формула P(A,B) выражает и минимальную порцию Информации, и элементарную ситуацию (ЭСит). Некоторые комбинации ЭСит могут образовать единицу «ситуация» (**СИТ**). Нетерминальный символ СИТ используется также для представления любого простого предложения текста. Наличие запроса может перевести СИТ в единицу **СОБ** («событие»). Главное же содержание текста представит единица, которую мы называем **текстовый факт (ТФ)**. Каждая из трех крупных единиц (СИТ, СОБ и ТФ) имеет лексическое ядро (**ЛЯ**), определяемое лингвистическим анализом, см. [2] .

Подобно «законным» лексемам или иным построенным в ходе анализа единицам, предсказанные текстом, но отсутствующие в нем единицы тоже получают знаковое отображение в структуре (в виде вопроса при узле или отношении и т.п.) – это **экспликация неполноты**. Локальная смысловая неполнота наследуется от предшествующего уровня анализа (не найденное в словаре слово или цепочка, синтаксический эллипсис, не получивший интерпретации знак препинания и др.) или появляется по правилам СемАн (не заполненная в пределах предложения смысловая валентность, нарушенная Грамматика смысловых отношений, парцелляция и др.) Иногда она снимается или гасится ближайшим контекстом. Не погашенную в пределах всего текста неполноту желательно суммировать и сделать вывод о свойствах самого текста. Восстановление опущенных частей текстовой структуры обычно сопровождается сжатием (устранением дублирования) и усилением параметра связности. Рассмотрим на небольшом примере, какие виды и уровни неполноты можно выделить уже имеющимися средствами на начальном этапе СемАн.

4. Неполнота первого уровня (краткий пример анализа)

30.08.06 12:08. NewsInfo (текст взят из Интернета).

7 человек погибли в ДТП в Калужской области

В среду утром семь человек стали жертвами ДТП, произошедшего в Калужской области. Об этом сообщает Управление информации МЧС России. Авария произошла в 7:50 утра на трассе Калуга-Медынь. Автомобиль Тойота вылетел с дороги и врезался в дерево неподалеку от нас. пункта Лев Толстой.

Следуя схеме анализа, принятой в информационно-лингвистической модели (ИЛМ) [2], сначала создается Внешняя дескрипция, которая фиксирует все имеющиеся в корпусе атрибуты данного текста, а сам текст получает статус ДОКумента. В нашем примере заполнятся четыре поля:

ЗГЛ(**7 человек погибли в ДТП в Калужской области,** ДОК)
ИСТОЧник(NewsInfo, ДОК)
ДАТА(30.08.06, ДОК)
ВРЕМЯ(12:08, ДОК).

К этому прибавятся формальные параметры количественного свойства (размер массива/корпуса, количество знаков, слов, предложений и т.п.), имеющиеся в любой системе. Не все параметры можно считать окончательными: так, уточнится первичное деление на «слова» и предложения. Критерий деления на предложения «по точке» установит сначала, что очередное предложение данного текста кончается словами *неподалеку от нас.* Синтаксический анализ сочтет такое предложение правильным; и как высказывание (при интерпретации они часто не совпадают) оно тоже формально правильно. Лексический анализ восстановит для слова *нас* лемму *мы,* а семантический может предложить гипотезу, что это авторское *мы*. Анализ следующего отрезка (*пункта Лев Толстой*) не признает его предложением (начинается со строчной буквы, не назывное), а квалифицирует его как «*обрывок* предложения». Структурная неполнота заставит пересмотреть принятые решения. Первой гипотезой должно быть предположение о том, что само конечное слово (*нас.*) является сокращением, а слова, разделенные точкой, образуют некоторый связный узел. Новая проверка на вхождение возможного термина (*нас. пункт* с леммой *населенный пункт*) в какой-либо из специализированных словарей или в общий словарь терминов даст скорее всего положительный ответ, достаточный для восстановления связности.

Но если не был найден термин *населенный пункт*, на вход синтаксического анализа поступит обрывок: *пункта Лев Толстой.* Правила семантической интерпретации не признают его отдельным Высказыванием (отсутствует предикат и это не назывное предложение), но внесут несколько неполных формул, фиксирующих возможные роли в составе отсутствующей Ситуации (в виде ?Сит). Интерпретация косвенного падежа оторванной именной группы создает гипотезу «Актант/Объект(пункт,?Сит)», а гипотеза Лок(пункт,?Сит) идет от семантической характеристики этой именной группы. Сама лексема *пункт* требует уточнения признаком, заполняющим валентность Уточн(?,*пункт*), при этом словосочетание *Лев Толстой* опознается как Имя(,) то ли этого неопределенного пункта: Имя(*Лев_Толстой,пункт)*, то ли актанта отсутствующей Сит (как в примере *Приказывает начальник этого пункта Лев Толстой*).

Теперь рассмотрим первичное (упрощенное) СемП заголовка (*7* ***человек погибли в ДТП в Калужской области***):

Пациенс (*7 человек*, *погибать*) // 1.
Модальн (РЕАЛ/Прош, *погибать*) // 2.
СемО? (*ДТП*, Сит1) // 3.
Лок (*Калуж.обл.*, *погибать*) // 4.
Причина (?, *погибать*) // 5.
Время (?, *погибать*) // 6.

В нем лексическое ядро строящейся Сит1 определяется однозначно: ЛЯ (*погибать, Сит1*); другой способ написания этой формулы: ЛЯ(Сит1)=*погибать.* Не заполнившиеся валентности лексемы становятся неполной формулой, с вопросом на месте отсутствующего актанта (здесь это валентности Причины и Времени у лексемы *погибать*). Вопрос на месте отношения в формуле 3 вызван отсутствием в словаре лексемы *ДТП*, поэтому вся предложная группа в ДТП с неясной семантической характеристикой присоединилась к нетерминальному символу Сит1 неопределенным отношением *СемО?*

Рассмотрим следующую порцию текста, - это высказывание, соответствующее первому предложению (***В среду утром семь человек стали жертвами ДТП, произошедшего в Калужской области***). Его СемП (тоже упрощенное) продолжает нумерацию формул предыдущего отрезка:
Время (*среда*, Сит2) // 7.
Время (*утром*, Сит2) // 8.
Равно (*семь человек*, *жертвы*) // 9.
Связан? (*жертвы*, *ДТП*) // 10.
Лок (*Калуж.обл.*, *ДТП*) // 11.

В этой фразе тоже нет сложностей для синтаксического анализа, а при семантической интерпретации соответствующего высказывания будет введен нетерминальный символ Сит2, с ядром ЛЯ (*быть жертвой*, Сит2). В отличие от любого нормального синтаксиса, который присоединит два начальных обстоятельства (***В среду*** и ***утром***) к сказуемому ***стали***, семантический интерпретатор свяжет каждый из них (на основании их начальной позиции) с верхним символом Сит2 и уточнит оба отношения как Время (, Сит2), потому что обе значимые лексемы имеют в словаре СХ=’отрезок времени’.

Несмотря на то, что СемП, полученное такой прямой интерпретацией синтаксических связей, формально

является полным, оно неявно содержит ряд неполных формул. Во-первых, остается неизвестной хотя бы семантическая характеристика слова *ДТП*, не найденного в словаре: СХ(ДТП)=?, и поэтому нельзя уточнить СемО в формуле 10. А оно требует уточнения: в иерархии СемО «Связан(,)» – самое высшее и ставится тогда, когда смысловая связь двух синтаксически связанных слов неясна. Во-вторых, СемАн требует устанавливать связь каждого следующего высказывания с предыдущим, а это выражается пока в неопределенном СемО между главными единицами: СемО? (Сит2,Сит1). Перечислим кратко возможные результаты сравнения двух фрагментов СемП, приводящего к взаимному насыщению незаполненных и неопределенных формул вместе с их сжатием (устранением избыточных формул). Отождествление узлов ***7 человек*** из первого (это ЗГЛ) и ***семь человек*** из второго высказывания текста плюс их отождествление с узлом ***жертвы*** (формула 9 Сит2), который является Пациенсом действия ***погибать*** (словарь), свидетельствует о том, что вторая Сит практически совпадает с первой, дополняя ее. Неполная формула 6 Сит1 (Время) может быть формально насыщена формулами 7 и 8 Сит2 (это еще и сжатие). Ненайденный узел *ДТП*, связанный с *жертвами*, а значит, с действием ***погибать,*** может быть (предположительно) назван Событием с отрицательной оценкой: ?СХ (*ДТП*)=Сит3,нехор. Его можно восстановить как Причину в формуле 5 Сит1, а неясное отношение в формуле 3 между ДТП и Сит1 уточнить тоже как Причина (ДТП, Сит1), причем эти две формулы становятся одной: Причина (*ДТП*, *погибать*), так как при сложении формул предпочтение отдается более конкретной, тем более что это валентность конкретной лексемы. Формулу 10 на тех же основаниях можно переписать как СемО Следствие (*жертвы*, *ДТП*) или его конверсив: Причина (*ДТП*, *жертвы*).

Заметим, что во всех преобразованиях мы опирались в основном на лингвистические знания. Из семантического словаря берутся валентности, сведения о синонимии, лексический вывод и др. Грамматика метаязыка задает иерархии СемО и СХ и правила сложения формул. Грамматика лингвистической Ситуации, представляющей содержание каждого простого предложения в стандартной форме, позволяет присоединять к символу Сит все лексемы и фрагменты текста, не затребованные семантическими валентностями лексем, да и сам символ Сит имеет свои валентности: Лок (,), Время (,) и др. Сжатие повторов – это следование законам построения текста.

5. Неполнота второго уровня

Итак, мы заполнили неполные формулы, уточнили насколько можно неопределенные узлы и отношения и устранили дублирование формул в СемП, опираясь на явные показатели неполноты. Но и такое СемП нельзя еще считать окончательным – необходимо проверить корректность каждой лингвистической единицы относительно «знаний» внешней информационной среды. Среди таких знаний будем различать общие и специальные. Легче сначала выделить специальные: они задаются каждой предметной областью (ПО) в том виде, который установила сложившаяся на настоящий день традиция. Лингвистам, создающим системы автоматического анализа текста, лексика ПО-знаний доступна в виде разного рода словников. Это общие классификаторы, отраслевые номенклатуры, поисковые списки, задаваемые самим пользователем, и Тезаурусы, которые снабжены грамматикой в виде иерархии терминов и минимума стандартных отношений (род-вид и т.д.). Самые конкретные единицы ПО – их «грамматика» и лексика – представлены многочисленными специальными базами данных: это вербальные названия полей и значения полей.

Как пример «общих» знаний можно привести описания единиц и логики Времени и Пространства, хорошо структурированные, доступные из многих источников и соответствующие здравому смыслу.

Возвратимся к нашему примеру с точки зрения неполноты второго уровня. Синтаксически и семантически полный и правильный узел, представляющий Время (А,Сит2), где А найдено как ***среда*** и ***утро***, являет собой **неполноценную информационную единицу**. На вопрос о времени события полным ответом было бы указание года, месяца и числа (или дня, ср. принятый в программировании шаблон ГГ.ММ.ДД) - такие сведения войдут и в базу данных. «Лок (***Калужская область***, Сит)» как место события тоже получит признак информационной неполноты (точнее, неполноценности): полная географическая единица должна отвечать шаблону «Страна, Область, Город». Мы назвали такую неполноту информационной, или неполнотой второго уровня, так как для ее обнаружения нужно выходить во внетекстовое пространство. Иногда восполнять недостающую информацию удается из богатого семантического словаря или из какой-либо структуры того же текста. В данном случае полный текстовый референт для Времени может быть восстановлен из даты Документа (*30.08.06*), а Место уточнится из продолжения текста (*Калуга-Медынь*). Но недостающее звено (страна = ***Россия***) может быть взято из Географической БД либо получено прагматическим выводом типа: «Если не указана страна, она совпадает со страной публикации». Исходный текст и его СемП могут содержать сколь угодно детальные подробности о месте и времени событий (*Авария произошла в **7:50** утра **на трассе** Калуга-Медынь)*, но СемАн должен по возможности восстанавливать полный вид единицы. Это нужно, чтобы на вопрос о времени или месте события система выдавала полноценный с информационной точки зрения ответ. Очень важно также доказательство связности тек-

ста, ведь продолжение текста может идти по опущенному в тексте, но очевидному элементу. Например, следующее предложение могло бы быть таким: «В ***России*** в этом году произошло рекордное количество аварий».

При формулировании конкретных целей анализа можно корректировать слишком жёсткие требования к полноте узла, задавая уточнение грамматики для данной ПО или для данного типа запроса. Каждая задача и привлекаемая БД могут усиливать или, наоборот, ослаблять свои требования к заполнению формул, соответственно текст будет получать оценку свойств по этому признаку. Так, понятие «ДТП» мы найдем обязательно в одном из Тезаурусов вместе с расшифровкой и указанием ассоциативных связей с лексемами, имеющимися и в нашем тексте, что поможет обоснованию связности текста. Но возможны и расхождения внутренней грамматики с грамматикой конкретной ПО. Например, лексема «человек» требует уточнений, какой-либо Идентификации, и это записано в лингвистическом словаре в виде валентности. В нашем тексте (*7* ***человек***) отсутствует идентификация. Семантика может обосновать это тем, что если именная группа обозначает «множество», валентность Идент (,) гасится и такой узел объявляется полным. В массиве «Криминальные новости» важно только количество людей, попавших в ДТП, и такие их идентификаторы как «потерпевший» или «жертва», так что и с точки зрения ПО-задачи этот узел можно признать полным. Но на самые важные для пользователя вопросы, например *Кто конкретно погиб?* и *Есть ли среди жертв дети?,* данный текст ответит «не знаю». Значит, этот текст неполон относительно вопросов «заинтересованных лиц».

Мы назвали два уровня неполноты единиц СемП. На самом деле более детальное рассмотрение собственно лингвистических шагов, в том числе выводов семантического анализа, дало бы еще не один промежуточный уровень неполноты. Ср. следующее высказывание «*Об этом сообщает Управление информации МЧС России»*, в котором референт содержания сообщения (*Об этом)* должен быть восстановлен как Сит1-2, а сложная неоднозначность «автор-адресат», возникшая дважды на синтаксическом уровне, должна разрешиться в результате нескольких обращений в административную ПО и проверкой на непротиворечивость выбранных интерпретаций. Неоднозначность нужно отнести к отдельному виду или уровню смысловой неполноты: ведь некоторые высказывания и рассчитаны на разные понимания. Еще сложнее анализ иронических и метафорических высказываний (например, *В нашей песочнице реконструкция совка идет полным ходом* - радио). СемАн подобных высказываний или целых текстов будет еще долго вне возможностей прикладных систем понимания с их «прямым» интеллектом. Но и в этом случае обнаруженная смысловая неполнота может сигнализировать, что здесь прямая интерпретация неприменима.

6. Заключение

Неполнота текста определяется ***относительно:*** описаний в словаре, длины анализируемого отрывка текста, полной структуры Ситуации, относительно Грамматики принятого метаязыка, относительно задачи и т.п. В докладе мы хотели только подчеркнуть, что неполнота скрепляет текст и приводит в действие собственно семантический анализ, который ставит вопросы и ищет на них ответы по всему текстовому пространству, сжимая первичное СемП и оставляя в нем только значимые единицы, в пределе лишь «текстовые факты», см. [2]. Вопросы, не получившие ответа в составе текста и не снятые как мало значимые, приводят в движение уже внетекстовые операции: важные вопросы должны быть переадресованы другим текстам или войти как «белые пятна» в общие либо специальные знания.

Полный семантический анализ текста в идеале включает в себя оценку качеств текста по разным признакам (связность, последовательность, логичность, полнота отображения содержания текста в его СемП). Наиболее важной нам представляется информационная значимость текста относительно внешней среды. Без общего решения проблемы Онтологии (привлечения прагматики и внетекстовых знаний к анализу текста) никакая система автоматического понимания текста не добьется ожидаемых от нее практических результатов. Недаром в США утвердилась новая компьютерная дисциплина под названием «Онтологическая семантика» [4]. Создать единую онтологию, объединяющую все ПО, представляется пока задачей мало реальной. Но проработать общие правила «перевода» с метаязыка лингвистических структур на языки представления знаний, используемые сейчас специалистами разных предметных областей – обязанность прикладных лингвистов. Можно остановиться пока на Тезаурусах как представителях ПО-знаний. Привлечение Тезаурусов, с одной стороны, повышает семантическую силу лингвистического анализа текстов, а с другой стороны, дает представление о тех сложностях взаимодействия лингвистических структур с внешней средой, которые встанут при обращении к «широкой» семантике. Обратный переход (от знаний к языковому воплощению) уже реализуется в нескольких системах генерации текстов – СГТ. Смысловая неполнота разных уровней, отображаемая в СемП, вполне «зримо» взывает к необходимости выхода в пространство целого текста и далее во внешнюю информационную среду.

Список литературы

1. Кузнецов И.П. Семантико-ориентированный лингвистический процессор для автоматической формализации автобиографических данных // Труды международной конференции «Диалог 2006». С. 317-322.
2. Леонтьева Н.Н. Корпусная лингвистика: не только вширь, но и вглубь // Труды международной конференции “Корпусная лингвистика-2006”. СПб.: Изд-во С.-Петерб. Ун-та, 2006. С. 234-241.
3. Севбо И.П. Сквозной анализ как шаг к структурированию текстовых знаний // НТИ. Серия 2. - 1989. - N 2. - С. 1-7.
4. Nirenburg S., Raskin V. Ontological Semantics // Cambridge, MA: MIT Press, 2004.