ISSN 2079-3316 ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ № 4(22), 2014, с. 41–66

УДК 004.89:004.912

Е. А. Сулейманова

О комплексном подходе к разрешению реляционно-аппозитивных неоднозначностей

Аннотация. В статье предложен метод разрешения одного частотного типа синтактико-семантической неоднозначности, предназначенный для использования в рамках задачи извлечения информации. Подход комбинирует различные способы снятия неоднозначностей в пределах предложения с выходом в дискурсивный контекст.

Ключевые слова и фразы: извлечение информации, синтактико-семантическая неоднозначность, дискурсивный контекст.

Введение

Проблема неоднозначности в языке носит универсальный характер и «является одной из самых сложных среди тех, что стоят перед любым текстовым анализатором, если вообще не самой сложной» [1]. Сталкиваются с ней и системы извлечения информации из текстов. Под термином «извлечение информации» (перевод англ. information extraction) понимают извлечение из неструктурированного текста (как правило, речь идет о больших объемах текста) фактической информации на заданную тему — например, поиск фактов о назначениях и отставках в текстах электронных СМИ. Задача извлечения информации состоит в том, чтобы обнаружить в тексте релевантные фрагменты и извлечь из них данные для заполнения целевых структур (фреймов события) [2], [3].

Большинство систем анализа текста реализуют поэтапную обработку: анализ текста на некотором уровне опирается на результаты, полученные на предшествующих уровнях. Практически любой текст содержит элементы, которые на том или ином уровне анализа

Работа выполнена в рамках НИР «Разрешение синтактико-семантической неоднозначности в рамках задачи извлечения информации из текстов, основанное на использовании кроссмодального контекста» (№ гос. регистрации 01201354592).

допускают более одной интерпретации. Чем ниже уровень, тем более формальные модели и методы лежат в основе машинного анализа и тем больше число альтернатив, возникающих в результате «беспристрастного» перебора всех формально допустимых возможностей. При этом неоднозначность на одном языковом уровне может разрешаться либо на основе собственных данных этого уровня, либо путем привлечения данных следующего уровня.

Семантико-семантическая неоднозначность — это возможность построения по одному отрезку текста разных синтаксических и, как следствие, семантических структур.

Для борьбы с синтаксической неоднозначностью при автоматическом анализе текста применяются статистические [4], [5], [6], инженерные (основанные на правилах) [7], [1] и комбинированные [8] методы. Все эти методы предполагают полный синтаксический анализ предложения, но ни один из них не учитывает экстрасентенциальный (т.е. лежащий за пределами предложения) контекст. Между тем, далеко не все случаи неоднозначности — не только собственно синтаксической, но и более низких уровней — в принципе поддаются разрешению в отдельно взятом предложении. Более того, не всегда даже человек, будучи ограничен рамками отдельного (взятого вне контекста) предложения, может выбрать правильную интерпретацию. Всего несколько примеров: *Эти типы стали есть в нашем цехе* (пример заимствован у ABBYY [9]); *Адвокат Петренко выступит на суде*; [...] *и назначил его заместителем Иванова.* Показательно, что при чтении связного текста трудностей с выбором единственно возможной альтернативы в тех же ситуациях у человека не возникает.

Несмотря на значительный прогресс в области машинной обработки естественного языка, «проблема разрешения неоднозначности остается камнем преткновения, особенно в тех случаях, когда в задачу системы входит извлечение смысла» [8]. Дальнейшее усовершенствование алгоритмов разрешения неоднозначности, опирающихся на информацию, «обеспечиваемую лексическими и грамматическими ресурсами систем обработки ЕЯ», с одной стороны, требует огромных затрат времени и труда, с другой — «многие случаи неоднозначности нельзя разрешить автоматически в принципе, так как для них существенны экстралингвистические знания, не извлекаемые непосредственно из текста» [там же].

1. Реляционно-аппозитивные неоднозначности

При извлечении информации о лицах значительная доля синтактико-семантических неоднозначностей приходится на конструкции с участием имен собственных и нарицательных имен существительных особого типа — реляционных имен. Реляционное имя обозначает первый член некоторого отношения и имеет облигаторную валентность, соответствующую второму члену отношения [**10**]. Примеры неоднозначностей такого рода: *адвоката Браудера, представитель Кэтрин Джексон, заместителя губернатора Ольги Васильевой, пресс-секретаря президента Израиля Шимона Переса* и т.п. В связи с рассматриваемой проблемой нас будут интересовать только те реляционные имена, которые обозначают лицо по отношению к другому лицу. Далее в тексте по умолчанию под реляционными именами мы будем понимать именно этот вид существительных.

Во всех приведенных примерах разрешение неоднозначности требует выбора единственной из нескольких возможных комбинаций связей двух типов. Связь первого типа — аппозитивная (связь между нарицательным именем существительным и именем собственным, заполняющим его валентность на имя). Связь второго типа — назовем ее *реляцией* — это отношение между реляционным именем существительным и заполнителем его валентности на отношение. Возникающие в таких конструкциях неоднозначности будем называть *реляционно-аппозитивными*.

Особенностью реляционно-аппозитивных неоднозначностей является то, что для их разрешения в общем случае недостаточно средств, ограниченных рамками лингвистического анализа отдельного предложения. Причем речь идет не только о компьютерном анализе: человеку при выборе интерпретации в подобных случаях часто требуется обращение к тому, что в психолингвистике принято называть дискурсивным, или референциальным, контекстом.

2. Психолингвистика о роли контекста в разрешении неоднозначностей человеком

Тому, как человек в процессе понимания текста справляется с присущей языку неоднозначностью, посвящено множество экспериментальных психолингвистических исследований. Разные психолингвистические модели анализа предложения (human sentence processing) различаются между собой по двум основным взаимосвязанным измерениям: (1) последовательность (рассматривается одна

начальная гипотеза) или параллельность (строятся и оцениваются несколько альтернатив) и (2) модульность (анализ на каждом уровне выполняется независимо) или интерактивность (вся доступная информация может использоваться в любой момент времени). Все теории признают как очевидное важную роль контекста в разрешении неоднозначности, но расходятся во взглядах относительно того, на каком этапе контекстные факторы вовлекаются в процесс анализа [11], [12], [13]. На одном полюсе (синтаксические теории слабого взаимодействия [14]) придерживаются мнения, что при выборе первоначальной гипотезы человеческий анализатор руководствуется некоторыми универсальными синтаксическими стратегиями — принципами «минимального вложения» и «позднего закрытия», а все прочие факторы, в том числе и контекстные, привлекаются исключительно для оценки (и при необходимости — пересмотра) синтаксической гипотезы. В противовес этому, сторонники референциальной теории [15] склонны считать, что дискурсивные факторы оказывают влияние на анализ уже на начальных его стадиях, опережая действие других факторов — в том числе и синтаксических.

Важно, что в понятие дискурсивного контекста включается не только лингвистический (текстуальный) контекст, а «сущности и свойства в окружающем мире», доступные («accessible») участникам коммуникации, и общие для последних пресуппозиции [12] — т.е., по существу, общие фоновые знания коммуникантов. Контекст, в котором происходит анализ предложения человеком, может быть представлен в некоторой ментальной модели, которая опирается на его знания об устройстве мира и которую он достраивает в процессе восприятия текста [15], [16].

Очевидно, что контекстом в таком понимании чрезвычайно сложно манипулировать при проведении экспериментов, и это существенно затрудняет экспериментальное подтверждение или опровержение положений противоборствующих теорий сильного и слабого взаимодействия. В своих попытках выяснить место и роль дискурсивного контекста в процессе анализа предложения и разрешения неоднозначностей исследователи вынужденно ограничиваются изучением референциального контекста в «узком» смысле — как множества референтов, вводимых в предшествующий текст. В качестве объекта исследования чаще всего используются конструкции, допускающие вре́менную (локальную) неоднозначность при интерпретации придаточного предложения с that (как дополнения при глаголе или относительного предложения, подчиненного именной группе) [17],

а также конструкции с неоднозначностью типа «раннее-позднее закрытие» (например, с неоднозначной вершиной относительного придаточного) [11], [18], [19], [20].

Общие положения, раскрывающие роль дискурсивного контекста в разрешении структурных неоднозначностей в предложении, сформулированы в рамках упомянутой выше референциальной теории [15].

1. *Принцип априорного правдоподобия* (The principle of a priory plausibility) гласит, что предпочтение будет отдано гипотезе, которая в большей степени соответствует общему знанию о мире или специфическому знанию об универсуме дискурса. При этом частные знания имеют приоритет над общими.
2. *Принцип референциальной успешности* (The principle of referential success). Предпочтение отдается интерпретации, которую адресат может соотнести с некоторой сущностью в своей ментальной модели универсума дискурса. Впоследствии этот принцип был заменен *Принципом референциальной неудачи* (referential failure) [13]: если имеется интерпретация, которая не соотносится ни с одной сущностью в ментальной модели адресата, то эта интерпретация будет отвергнута.
3. *Принцип парсимонии* (The principle of parsimony) применяется при наличии нескольких интерпретаций, ни одна из которых не соотносится с ментальной моделью адресата. Если среди интерпретаций есть одна, при которой оказывается не выполнено меньшее число пресуппозиций или импликаций, чем при других, то, при прочих равных условиях, именно она будет выбрана адресатом, а недостающие пресуппозиции будут включены в его (адресата) ментальную модель.

Эти принципы должны обеспечивать выбор той альтернативы, которая в большей степени соответствует текущей модели дискурса (которая требует меньшей ревизии этой модели). Сложность формализации этих принципов для алгоритмического разрешения неоднозначности очевидна.

3. Проблема неоднозначности и задача извлечения информации

Извлечение информации, относящееся к т. наз. «поверхностным» (shallow) методам анализа текста, не включает в цикл

обработки полный синтаксис предложения. В силу фрагментарного характера синтаксиса и ограниченности (узкой направленности) семантического анализа в системе извлечения информации ИСИДА-Т [21], у анализатора гораздо меньше возможностей для снятия синтактико-семантических неоднозначностей с помощью древесных и семантических критериев, чем у систем с развитыми синтаксическим и семантическим компонентами. С другой стороны, принятый разработчиками системы ИСИДА подход к извлечению информации отличается рядом особенностей, благодаря которым к задаче автоматического разрешения синтактико-семантической неоднозначности можно подойти с качественно иных позиций.

(1) Извлечение информации опирается на ресурс знаний, включающий, наряду с лингвистическими, концептуальные знания о предметной области в форме иерархий и некоторое количество фактических знаний. Это открывает возможности для моделирования общих знаний о мире.
(2) В процессе решения основной задачи (извлечения фактов) строится специальный референциальный уровень представления текста — уровень участников извлекаемых фактов и отношений между ними. Единицы этого уровня — модели референтов текстовых упоминаний, получаемые в результате объединения фрагментов кореферентных описаний, — и являются теми объектами, которые подлежат включению в базу фактов [22].

Совокупность этих обстоятельств подтолкнула нас к идее компенсировать свойственные подходу ограничения лингвистического анализа в пределах предложения возможностью моделирования некоторого интегрального контекста, понимаемого как совокупность представления текста на референциальном уровне и знаний разной природы:

(1) знаний общих и свойственных жанру закономерностей построения связного текста и, в частности, номинации;
(2) знаний об устройстве мира (концептуальные, или онтологические, знания);
(3) знаний о конкретных экземплярах (фактические, или энциклопедические, знания);
(4) знаний прагматического контекста (уровень метаданных текста — место, время создания).

Мы предполагаем, что использование такого контекста позволит моделировать ограниченное, но достаточное для решаемой задачи понимание текста, включая успешное разрешение присущих языку неоднозначностей.

4. О методе разрешения синтактико-семантической неоднозначности

С точки зрения модульного анализатора текста, используемого в системе ИСИДА-Т, реляционно-аппозитивные неоднозначности:

(1) могут быть следствием неустраненной неоднозначности результатов на уровнях анализа, предшествующих синтактико-семантическому. Примеры: *представитель Кэтрин Джексон* (морфологическая неоднозначность — именительный или родительный падеж имени собственного); *дочь композитора Александра Афанасьева* (лексико-грамматическая омонимия имени собственного — мужское или женское); *отца Михаила* (лексическая омонимия имени существительного);
(2) могут быть никак не связаны с омонимичностью результатов нижележащих уровней. Пример: *по словам пресс-секретаря Емельяненко* (*Емельяненко* — фамилия пресс-секретаря или его работодателя).

Естественно поэтому для автоматического устранения синтактико-семантических неоднозначностей использовать две взаимодополняющие стратегии:

«превентивная» — минимизировать возможность возникновения синтактико-семантической неоднозначности, устраняя всеми доступными средствами досинтаксическую омонимию (грамматическую, лексико-грамматическую, лексическую);
«постфактумная» — разрешать неоднозначности синтактико-семантического уровня, отбирая (исключая) варианты уже построенных связей.

Из сказанного не следует, что «превентивная» стратегия сможет полностью решить проблему неоднозначностей первой группы, оставив на долю «постфактумной» стратегии лишь неоднозначности второй группы. Часть случаев досинтаксической омонимии сохраняется до синтактико-семантического этапа, порождая неоднозначность, не всегда разрешимую в пределах отдельного предложения, например:

Иванов назначил его (личное или притяжательное местоимение?) *заместителем Петрова* (родительный или винительный падеж?).

Предлагаемый метод разрешения реляционно-аппозитивных неоднозначностей включает в себя следующие этапы:

(1) Разрешение досинтаксических видов неоднозначностей. Применительно к некоторой реляционно-аппозитивной конструкции, задача первого этапа состоит в том, чтобы ранжировать по вероятности варианты лексико-морфологического анализа ее компонентов. На основе наиболее вероятных результатов будут построены все возможные варианты синтактико-семантических связей.
(2) Разрешение реляционно-аппозитивных неоднозначностей в пределах предложения. На этом этапе часть связей отсеивается как а) невозможные — не отвечающие дополнительным ограничениям на компоненты; б) теоретически возможные, но маловероятные; в) теоретически невозможные в силу макросинтаксических ограничений — если таковые удалось проверить.
(3) Выход за пределы предложения и разрешение реляционно-аппозитивных неоднозначностей с помощью дискурсивных правил (эвристик) в процессе построения референциального представления текста. Дискурсивные эвристики выбирают наиболее вероятное решение в некоторых типовых ситуациях.

Таким образом, в случае успешного разрешения неоднозначности для некоторой конструкции результирующая интерпретация представляет собой наиболее вероятную гипотезу.

Правильность выбранной альтернативы оценивается на этапе верификации референциальной гипотезы — референциального представления, построенного по тексту модулем референциального анализа. Разработка методов верификации, опирающихся на моделирование дискурсивного контекста как некоторого фрагмента ментальной модели адресата текста, представляет собой предмет отдельного исследования.

5. Досинтаксические виды неоднозначности и методы их разрешения

К досинтаксическим видам неоднозначности относят грамматическую, лексико-грамматическую и лексическую омонимию. Применительно к задаче извлечения информации проблема неоднозначности на этих уровнях усугубляется тем, что системе приходится

иметь дело с принципиально открытыми классами лексики (личные имена и фамилии, названия организаций, географических и геополитических единиц). Для снятия досинтаксических неоднозначностей система извлечения информации использует весь арсенал доступных ей средств:

(1) для снятия частеречной омонимии — статистический метод на основе машинного обучения;
(2) для разрешения[1] внутрилексемной омоформии — микросинтаксический контекст (адъективы, предлоги);
(3) для разрешения межлексемной омоформии — микросинтаксический и ближайший линейный контекст (грамматическая форма и лексико-семантический состав); анализ графематики и позиции в предложении (для случайных совпадений собственных и нарицательных имен); понижение в ранге редких личных имен, имеющих более частотные омоформы (*Светлан, Марин*);
(4) для снятия лексической омонимии — контекстные эвристики, тематическое тегирование текста;
(5) для уменьшения неоднозначности при определении границ именованных сущностей предполагается использовать прецедентные данные (*мэр Рима Иньяцио Марино* — вся выделенная цепочка может быть принята за имя лица, т.к. в словаре имеется личное имя *Рима*).

5.1. Лексическая неоднозначность имен лица

Остановимся подробнее на двух аспектах лексики, используемой для именования лица. Речь пойдет о тех ее особенностях, которые при автоматическом анализе служат постоянным источником реляционно-аппозитивных неоднозначностей.

5.1.1. *Лицо или не лицо*

Первая особенность связана с двойственной референциальной природой, присущей существительным со значением аспекта лица. Такие существительные могут быть употреблены с референцией к конкретному лицу, и в таком употреблении они обладают валентностью на имя (способны подчинять себе имя собственное по аппозитивной связи).

[1] Начиная с лексического уровня, выбор и удаление вариантов при разрешении неоднозначности технически моделируются как вероятностное ранжирование гипотез (т.е. отвергнутые варианты не удаляются, а понижаются в ранге)

С другой стороны, они теряют такую способность, будучи употреблены предикатно (в значении свойства лица), с референцией к собственно аспекту (например, для обозначения должности как таковой) или нереферентно (родовые, универсальные употребления). Очевидно, что распознавание употреблений такого типа уменьшит число неоднозначностей за счет запрета на построение или устранения построенных ложных аппозитивных связей. Ситуации случайного соседства нереферентной именной группы — аспекта лица с именем собственным настолько редки, что ими можно пренебречь. А вот должность в значении ‘должность’, а не ‘лицо’, часто оказывается рядом с именем собственным (*Он останется работать в ЦБ в качестве советника Набиуллиной*). Отличить должность от лица с большой вероятностью можно на основании типового линейного контекста — и тогда существительное получит специальную помету, которая запрещает построение аппозитивной связи с именем. Довольно часто встречаются и случаи предикатного употребления реляционно-аппозитивных конструкций. Чтобы диагностировать их (отличить от референтных употреблений в составе сочинительных конструкций или в конструкциях идентификации), требуется более широкий макросинтаксический контекст. В качестве такого контекста могут выступить результаты фрагментационного анализа (см. раздел «Устранение аппозитивных связей от вершин обособленных приложений при именах собственных»).

5.1.2. *Реляционное или нереляционное*

Вторая особенность именований лица связана со свойством реляционности — понимаемым как наличие у имени облигаторной валентности на отношение к другому лицу. Это свойство порождает неоднозначные реляционно-аппозитивные интерпретации в силу нескольких обстоятельств.

Во-первых, наличие облигаторной (семантической) валентности на реляцию не означает, что она непременно должна быть заполнена в тексте явно (бóльшая часть реляционных имен допускает её конситуативное заполнение [**10**]).

Во-вторых, реляционные имена могут иметь нереляционные омонимы (реляционные *отец, мать* как термины родства и нереляционные — как именование лиц духовного звания).

В-третьих, есть целый класс нереляционных имен со значением профессии, рода деятельности (*адвокат, врач, продюсер* и т.п.), которые при необходимости легко превращаются в реляционные

(*адвокат Виктора Бута, врач Майкла Джексона*); оба значения (или употребления в рамках одного значения) могут сосуществовать в одном тексте и применительно к одному и тому же референту. Например: *Адвокату* (реляционное) *Петрова стало известно, что* [...]. *Адвокат* (реляционное или нереляционное?) *утверждает* [...] — ср. вполне равноценную замену повторной номинации *адвокат* на нереляционное *юрист*. Очевидно, во втором упоминании можно говорить не о лексической неоднозначности, а скорее о лексической двусмысленности, недоопределенности.

В-четвертых, реляционные имена могут терять реляционность в составе собственных имен (*Дед Хасан, тетя Маша* — А. Д. Шмелёв считает такое употребление близким к «цитатному» [**23**]).

На этом фоне то обстоятельство, что реляционные имена могут превращаться в нереляционные, употребляясь абсолютивно (*хороший отец*; *вдова* в значении ‘женщина, потерявшая мужа’), можно считать наименьшим злом — во всяком случае, в новостных текстах вклад таких употреблений в проблему реляционно-аппозитивных неоднозначностей незаметен.

Подробнее реляционные свойства имен обсуждаются в разделе «Разрешение синтаксических неоднозначностей в процессе построения референциального представления».

6. Разрешение синтаксических неоднозначностей в пределах предложения

6.1. Виды неоднозначных реляционно-аппозитивных конструкций

Формальными признаками синтаксической неоднозначности являются следующие:

(1) у слова больше одной исходящей связи одного типа;
(2) у слова больше одной входящей связи;
(3) аппозитивная связь вложена по границам в другую связь при том же хозяине.

В принципе, любая реляция или аппозитивная связь, построенная фрагментарным синтаксическим анализом, в том числе и в цепочках упомянутых видов, может оказаться фиктивной (*через своего представителя Христенко заявил*). Ошибки такого рода особенно коварны в условиях поверхностного подхода, поскольку не имеют явных «локальных» признаков.

На практике в большинстве случаев реляционно-аппозитивные неоднозначности представлены двух- или трехкомпонентными конструкциями следующего вида.

(1) Двухкомпонентная конструкция, состоящая из дескрипции Д и имени собственного ИС (часто в омонимичной форме), между которыми возможны два взаимоисключающих варианта синтаксической связи — реляция и аппозитивная (*отца Михаила, адвокат Гарри Погоняйло*).
(2) Трехкомпонентная конструкция, состоящая из дескрипций Д1 и Д2 и имени собственного ИС (возможно, в омонимичной форме). Между компонентами могут быть построены следующие связи: реляция от Д1 к Д2 и две взаимоисключающих аппозитивных связи — от Д1 к ИС и от Д2 к ИС (*друга журналиста Гленна Гринвальда*).

Следующие два вида цепочек являются потенциальным источником реляционно-аппозитивных неоднозначностей, связанных с разными вариантами членения предложения на сегменты-поддеревья.

(3) Цепочка, состоящая из омонимичной формы местоимения, дескрипции Д и имени собственного ИС. Для фрагментов возможны следующие взаимоисключающие комбинации связей: (1) реляция от Д к местоимению и аппозитивная от Д к ИС, (2) при отсутствии связи между Д и местоимением возможна реляция от Д к ИС (*его брата Арсена*).
(4) Цепочка, состоящая из дескрипции Д и двух имен собственных ИС1 и ИС2 (возможно, в омонимичной форме). Возможны следующие взаимоисключающие комбинации связей: (1) реляция от Д к ИС1, аппозитивная от Д к ИС2; (2) аппозитивная от Д к ИС1, а ИС2 не входит в сегмент-поддерево (*пресс-секретаря Бейнера Брендана Бака*).

6.2. Проверка дополнительных признаков у компонентов

Проверка рода подчиненного имени собственного позволяет устранить реляции типа *жены Евгении*.

6.3. Устранение аппозитивных связей от вершин обособленных приложений при именах собственных

Этот метод опирается на результаты фрагментации и на следующее предположение: реляционное имя, являющееся вершиной обособленного фрагмента, подчиненного имени собственному (или именной группе, подчиняющей имя собственное по аппозитивной связи), не может подчинять себе имя собственное по аппозитивной связи. Примеры: *Александра Афанасьева-Шевчук — дочь композитора Александра Афанасьева; оставили его с Кэрол Миддлтон, матерью Кэтрин; полиция задержала Давида Миранду, друга журналиста Гленна Гринвальда.*

Реляционно-аппозитивные неоднозначности в такой позиции довольно распространены. Успешность их разрешения зависит от надежного определения типа фрагмента и его вершины, что в рамках поверхностного подхода обеспечить довольно сложно (в связи с проблемой «синтаксической омонимии» при сегментации предложения [**24**]).

6.4. Устранение маловероятных связей

Речь идет о предпочтении некоторой синтаксической конструкции при анализе фрагментов, в принципе допускающих и другие, более редкие, синтаксические интерпретации. Заметим, что подобными методами не пренебрегают и системы с чрезвычайно развитым лингвистическим аппаратом — описан опыт использования эмпирических весов лингвистическим процессором ЭТАП-3 [**1**].

Выбор наиболее вероятной комбинации связей применяется, например, к цепочкам вида (3) (см. раздел «Виды неоднозначных реляционно-аппозитивных конструкций»). Для последовательности *его заместителя Петренко* (с омонимичной формой имени собственного) синтактико-семантический анализ построит три возможные связи:

- реляция от *заместителя* к *его*,
- реляция от *заместителя* к *Петренко*,
- аппозитивная от *заместителя* к *Петренко*.

В действительности такой фрагмент допускает четыре возможных интерпретации (комбинации связей):

- реляция от *заместителя* к *его*, аппозитивная от *заместителя* к *Петренко*;

- реляция от *заместителя* к *его*, а между *заместителя* и *Петренко* связь отсутствует;
- реляция от *заместителя* к *Петренко*, а между *его* и *заместителя* связь отсутствует;
- аппозитивная от *заместителя* к *Петренко*, а между *его* и *заместителя* связь отсутствует.

В качестве наиболее вероятной выбирается первая гипотеза, при которой цепочка представляет собой единую синтаксическую конструкцию и которая требует удаления (понижения в ранге) реляции от реляционного имени (*заместителя*) к имени собственному.

7. Разрешение синтаксических неоднозначностей в процессе построения референциального представления

Построением референциального представления текста занимается специальный модуль референциального анализа. Для разрешения в процессе этого реляционно-аппозитивных неоднозначностей служат эвристические дискурсивные правила.

Правила кодируют ту часть лингвистической компетенции анализатора, которая отвечает за знание некоторых закономерностей референции, как общих, так и свойственных текстам новостного жанра.

В качестве первой общей закономерности укажем основной прагматический принцип употребления имен собственных [**23**]. Он заключается в том, что в отсутствие специальных показателей интродуктивности имя собственное может быть употреблено конкретно-референтно только в том случае, если адресату (по мнению автора) носитель имени собственного уже известен (введен в дискурс или является носителем т. наз. «прецедентного» имени). Известность предполагает наличие в сознании адресата «мысленного досье» носителя имени — информации в виде совокупности дескрипций. Если известность референта адресату не предполагается, то интродуктивное употребление имени собственного предполагает наличие при нем некоторой дескрипции, которая и позволит локализовать носителя имени в релевантном денотативном пространстве.

Далее сформулируем еще несколько принципов, которыми дискурсивные эвристики руководствуются при разрешении неоднозначностей в реляционно-аппозитивных конструкциях.

7.1. Дискурсивные принципы употребления реляционных конструкций

Принцип референтного употребления нового имени собственного

(следствие из общего прагматического принципа)

Новое имя собственное может быть употреблено референтно, только если:

(1) это «прецедентное» (общеизвестное) имя или
(2) при ИС имеется синтаксически обособленная дескрипция или
(3) имеет место презумптивная референция в рамках акта имплицитного представления (сначала носитель имени вводится в текст посредством одной только дескрипции, а в последующем тексте он же упоминается при помощи имени собственного).

Замечание: случаями типа (3) для разрешения неоднозначности можно пренебречь, поскольку употребление имени собственного с презумптивной референцией в составе неоднозначной синтаксической конструкции маловероятно.

Принцип иерархии новизны для реляционных конструкций

Референт вложенного упоминания не может быть «новее», чем референт вершины.

Принцип иерархии именованности для реляционных конструкций

Имя собственное в составе реляционной конструкции с двумя дескрипциями может быть присоединено по аппозитивной связи к вершине только в том случае, если референт вложенного упоминания в нем не нуждается.

Принцип минимальной интродукции

В ненулевом референциальном контексте реляционная конструкция, включающая новое имя собственное, обычно вводит один новый референт через отношение к уже известному (упомянутому в тексте или прецедентному). Ситуация, когда такая конструкция вводит в дискурс два новых референта, менее типична.

Принцип приоритета частного над общим

Фактические данные (прецедентные или предоставленные текстом) имеют приоритет над общими принципами.

7.2. Классификация реляционных имен

Класс имен существительных, могущих выступать в качестве вершины реляционной конструкции, неоднороден. Разным именам свойственно различное поведение в составе реляционно-аппозитивных конструкций; естественно попытаться систематизировать эти различия и учитывать их при разрешении возникающих неоднозначностей.

7.2.1. *Признак «реляционность»*

В интересах дискурсивных правил разрешения неоднозначностей введем лексический признак «реляционность». Признак имеет следующие значения: «реляционное», «переменно-реляционное» и «нереляционное». Лексеме приписывается некоторое априорное значение этого признака, которое у текстовой словоформы может быть переопределено.

Будем считать, что имя в реляционном употреблении (т.е. словоформа с признаком «реляционное») подчиняется следующему принципу:

Принцип приоритета реляции

Имя в реляционном употреблении может присоединять к себе имя собственное по аппозитивной связи только при заполненной валентности на отношение.

Большинство реляционных имен в этой ситуации не допускают конситуативного заполнения реляции. Валентность на имя собственное при таком имени может быть заполнена только при явно заполненной валентности на отношение (допустим лишь эллипсис в конструкции сочинения — *его друга и <∅=его> коллеги Петрова*). Исключение составляют, пожалуй, лишь термины родства и некоторые другие имена.

Априори считаем «реляционными» большинство имен категории «термин родства», некоторые должности (*заместитель, помощник*), а также ряд имен разных семантических категорий: *бизнес-партнер, друг, знакомый, коллега, любимец, одноклассник, однокурсник, партнер, подзащитный, подруга, последователь, представитель, предшественник, преемник, приятель, противник, родственник, собеседник, соотечественник, соратник, сторонник*, а также результативные имена с сильной валентностью на объект (субъект) действия — *убийца, жертва*.

Значение «переменно-реляционное» имеют имена разных категорий, которым свойственны как нереляционные, так и реляционные значения (употребления): *агент, адвокат, врач, клиент, продюсер, пресс-секретарь, секретарь, советник, ученик, учитель*. Сюда же можно отнести существительные-должности типа *начальник, командир, шеф*, которые имеют валентность на объект управления, но могут терять ее и превращаться в реляционные в принятом нами смысле (*начальник* (чей) *Петрова*).

Лексемы, которым не приписано ни одно из двух указанных значений признака, по умолчанию считаются «нереляционными».

Априорное значение признака реляционности, кроме значения «нереляционное», может быть переопределено при анализе в контексте — в тех случаях, когда реляционность или нереляционность употребления маркируется явными (легко обнаруживаемыми автоматически) лексико-синтаксическими средствами. Так, в примерах *второй/прежний адвокат* переменно-реляционное имя *адвокат* употреблено скорее реляционно (ср. *бывший адвокат* — здесь сохраняется априорное значение признака). *Известный/хороший адвокат/врач, известный продюсер, известный/заслуженный учитель* — примеры нереляционных употреблений. Опознание имени как нереляционного позволяет устранить синтаксические неоднозначности, связанные с построением ложных реляций. Реляционность употребления, подтвержденная контекстом, сама по себе не снимает возможной неоднозначности, а позволяет применить для ее разрешения соотвествующие дискурсивные эвристики.

7.2.2. *Признак «семантическая категория»*

Признак приписывается лексеме и имеет следующие значения: «должность», «род деятельности», «термин родства», «личные отношения» и др.

7.2.3. *Признак «этикетное»*

Признак характеризует некоторые априорно реляционные имена как способные (склонные) к нереляционному цитатному или этикетному употреблению, например *дед, дедушка, тетя, дядя*.

7.2.4. *Признак «наличие нереляционного омонима»*

Признак полезен в тех случаях, когда до работы дискурсивных правил для словоформы не удалось снять лексическую омонимию, сопряженную с реляционностью или нереляционностью имени.

7.3. Дискурсивные правила разрешения неоднозначностей

Дискурсивные правила запускаются модулем референциального анализа. Наряду с признаками реляционных имен, правила используют признак для имен собственных, имеющий два значения: «новое» и «неновое». Новым имя собственное считается в том случае, если ни это имя, ни его варианты (предположительно относящиеся к одному и тому же референту) в тексте еще не встречались. Соответственно, неновым имя собственное считается в том случае, если оно или его вариант уже встречались в тексте. Значение признака определяет модуль референциального анализа.

Кроме того, модуль референциального анализа выполняет процедуры и проверки, необходимые дискурсивным правилам для принятия решения:

- в каком референциальном контексте — нулевом или ненулевом — встретилась данная неоднозначная конструкция;
- проверка имени собственного на «прецедентность» (общеизвестность) — поиск его в базе прецедентов. База прецедентов — это база данных, содержащая относительно достоверные результаты извлечения информации из текстов (т.е. те, которые получены по «хорошим» фрагментам, не допускающим разночтений);
- проверка данной дескрипции на совпадение с интродуктивной дескрипцией носителя имени собственного;
- проверка совместимости по тексту данной дескрипции с неновым именем собственным;
- проверка совместимости данной дескрипции с прецедентным именем собственным;
- проверка, является ли дескрипция повторным упоминанием некоторого ранее введенного в дискурс референта;
- проверка, может ли данная дескрипция быть упоминанием некоторого носителя прецедентного имени (т.е. принадлежит ли она списку лексем, типично обозначающих общеизвестных деятелей).

Порядок применения дискурсивных правил не детерминирован. Выполняется то правило, условия которого выполнены.

Дискурсивные эвристики применяются только для случаев явно выраженной неоднозначности (см. «Виды неоднозначных реляционно-аппозитивных конструкций»). Распознавание случаев

ложной связи (единственной ложной связи между двумя элементами) является одной из задач этапа верификации референциальной гипотезы.

7.3.1. *Примеры дискурсивных эвристик*

Правило 1.

(1) Последовательность вида «дескрипция (Д) + имя собственное (ИС)»;
(2) от Д к ИС построены две связи — реляция и аппозитивная;
(3) Д (главное слово Д) имеет признак «термин родства»;
(4) ИС — фамилия без имени.

удалить аппозитивную связь (пример 1).

Пример 1.

20 февраля дочь Тимошенко отпраздновала день рождения в итальянской столице.

Правило 2.

- Последовательность вида «дескрипция (Д) + имя собственное (ИС)»;
- от Д к ИС построены две связи — реляция и аппозитивная;
- Д (главное слово Д) имеет признаки «переменно-реляционное» и «род деятельности»;
- ИС — новое;
- ИС не найдено в базе прецедентов.

удалить реляцию (пример 2).

Пример 2.

Администрация краснокаменской колонии изъяла адвокатское удостоверение у адвоката Ирины Хруновой [...].

Правило 3.

(1) Последовательность вида «дескрипция (Д) + имя собственное (ИС)»;
(2) от Д к ИС построены две связи — реляция и аппозитивная;
(3) Д (главное слово Д) имеет признак «реляционное»;
(4) ИС — не новое.

удалить аппозитивную связь (примеры 3а, 3б).

Пример 3а.

Мать короля поп-музыки Майкла Джексона Кэтрин, которая две недели назад «пропала» [...] Ранее, 22 июля, представитель Кэтрин Джексон подала в полицию Калифорнии заявление о пропаже ее клиента.

Пример 3б.

В профиле автора говорится, что он [упомянутый ранее сын Асада — прим. автора статьи] является выпускником Оксфордского университета и игроком футбольного клуба «Барселона», что не может быть правдой, учитывая юный возраст сына Башара Асада и тот факт, что он проживает в Дамаске.

Правило 4.

(1) Последовательность вида «дескрипция Д1 + дескрипция Д2 + имя собственное (ИС)»;
(2) от Д1 к Д2 построена реляция, от Д1 к ИС и от Д2 к ИС построены аппозитивные связи;
(3) ИС найдено в базе прецедентов.

Проверить Д2 на совместимость с прецедентным ИС.
Если совместима, то удалить аппозитивную связь от Д1 к ИС (пример 4а).
Если не совместима, то удалить аппозитивную от Д2 к ИС (пример 4б).

Пример 4а.

Брат знаменитого модельера Джанни Версаче [...].

Пример 4б.

Окончательная правовая оценка действиям супруги экс-мэра Москвы Елены Батуриной будет дана по его завершении.

Правило 5.

(1) Последовательность вида «дескрипция Д1 + дескрипция Д2 + имя собственное (ИС)»;
(2) от Д1 к Д2 построена реляция, от Д1 к ИС и от Д2 к ИС построены аппозитивные связи;
(3) Д1 (главное слово Д1) имеет признак «должность»;
(4) Д2 (главное слово Д2) имеет признак «должность»;
(5) ИС — новое;
(6) ИС не найдено в базе прецедентов.

удалить аппозитивную от Д2 к ИС (предполагаем, что имеем дело с «псевдореляционной» конструкцией, в которой зависимый по реляции член представляет собой часть названия должности: примеры 5а, 5б, 5в).

Пример 5а.

В ходе выставки состоялись двухэтапные переговоры представителей ГК «Укрспецэкспорт» и заместителя министра обороны Брунея Падука Хаджи Мустафа бин Хаджи Сирата.

Пример 5б.

Помимо этого состоялись переговоры заместителя начальника управления Россельхознадзора Александра Пономарева с главным ветинспектором Польши Янушем Звионзеком.

Пример 5в.

Об этом, как сообщили корреспонденту ИА REGNUM в прессслужбе администрации региона, было заявлено 26 марта на заседании регионального Совета по проблемам инвалидов и граждан пожилого возраста, которое прошло под председательством заместителя губернатора Ольги Васильевой.

Правило 6.

(1) Последовательность вида «дескрипция Д1 + дескрипция Д2 + имя собственное (ИС)»;
(2) от Д1 к Д2 построена реляция, от Д1 к ИС и от Д2 к ИС построены аппозитивные связи;
(3) референциальный контекст — не нулевой;
(4) либо Д1 (главное слово Д1), либо Д2 (главное слово Д2) не имеет признака «должность»;
(5) ИС — новое;
(6) ИС не найдено в базе прецедентов.

Проверить, может ли Д2 быть повторным упоминанием некоторого референта, введенного ранее в дискурс. Если да, то удалить аппозитивную связь от Д2 к ИС (примеры 6а, 6б).

Пример 6а.

Между тем сегодня истекает срок пребывания Гончара под стражей, однако в Минске уже слышны заявления о том, что этого не произойдет. Адвокат задержанного Гарри Погоняйло полагает, что [...].

Пример 6б.

[...] по законодательству Белоруссии адвокатом находящегося в СИЗО КГБ российского бизнесмена Владислава Баумгертнера не может быть юрист «Уралкалия». По словам белорусского адвоката российского бизнесмена Алексея Басистова [...].

Правило 7.

(1) Последовательность вида «дескрипция (Д) + имя собственное ИС1 + имя собственное ИС2»;
(2) от Д к ИС1 построены реляция и аппозитивная связь, от Д к ИС2 — аппозитивная;
(3) ИС1 — не новое, ИС2 — новое.

Удалить аппозитивную связь от Д к ИС1 (пример 7).

Пример 7.

Изучив доводы сторон суд не нашел оснований для удовлетворения

требований адвоката. По мнению суда, права Браудера не были нарушены в ходе предварительного следствия.

Судебное заседание по делу отложено на 3 июля, в связи с отсутствием второго адвоката Браудера Владимира Бойко.

8. Заключение

В статье описан один тип синтактико-семантических неоднозначностей, который требует разрешения в задаче извлечения информации о лицах, — реляционно-аппозитивные неоднозначности. Особенность неоднозначностей этого типа состоит в том, что для выбора единственной альтернативы часто недостаточно лингвистической информации, ограниченной пределами отдельно взятого предложения. Суть предлагаемого метода разрешения реляционно-аппозитивных неоднозначностей заключается в последовательном отборе наиболее предпочтительных результатов анализа на основе критериев разной природы: от линейного и микросинтаксического контекста до дискурсивных ограничений на способы упоминания референтов в связном тексте и фактических (прецедентных) знаний.

Результаты комплексной оценки метода будут опубликованы отдельно.

Список литературы

[1] Иомдин Л. Л., Сизов В. Г., Цинман Л. Л., «Использование эмпирических весов при синтаксическом анализе», Труды международной конференции, Обработка текста и когнитивные технологии, **6**, Отечество, Казань, 2001, с. 64–72 ↑ 41, 42, 53.

[2] Poibeau T., Saggion H., Piskorski J., Yangarber R. (eds.), *Multisource, Multilingual Information Extraction and Summarization*, Theory and Applications of Natural Language Processing, Springer, Berlin–New York, 2013, 323 pp. ↑ 41.

[3] Кормалев Д. А., Куршев Е. П., Сулейманова Е. А., Трофимов И. В., «Технология извлечения информации из текстов, основанная на знаниях», *Программные продукты и системы*, 2009, № 2(86), с. 62–66 ↑ 41.

[4] Miyao Y., Tsujii J., “A model of syntactic disambiguation based on lexicalized grammars”, Proceedings of CoNLL-2003 (Edmonton, Canada, 2003), pp. 1–8 ↑ 42.

[5] Chiang D., “Statistical parsing with an automatically-extracted tree adjoining grammar”, Proceedings of ACL-2000, pp. 456–463 ↑ 42.

[6] Гельбух А., «Разрешение синтаксической неоднозначности и извлечение словаря моделей управления из корпуса текстов», Материалы VIII Международной конференции KDS-99 (Кацивели, 1999) ↑ 42.

[7] Ю. Д. Апресян, И. М. Богуславский, Л. Л. Иомдин и др., *Лингвистическое обеспечение системы ЭТАП-2*, Наука, М., 1989 ↑ 42.

[8] Богуславский И. М., Иомдин Л. Л., Лазурский А. В., Митюшин Л. Г., Бердичевский А. С., «Интерактивное разрешение внутренней и переводной неоднозначности в системе машинного перевода», *Компьютерная лингвистика и интеллектуальные технологии*, Труды конф., Диалог 2005 (Звенигород, 1–6 мая 2005 г.), Наука, М., 2005, с. 216–221 ↑ 42.

[9] *Лингвистические технологии ABBYY. От сложного — к совершенному*, http://www.3dnews.ru/software/624398/print (дата обращения 14.11.2014) ↑ 42.

[10] Шмелёв А. Д., «Типы «невыраженных валентностей»», Семиотика и информатика, **36**, М., 1998, с. 167–176 ↑ 43, 50.

[11] Desmet T., De Baecke C., Brysbaert M., "The influence of referential discourse context on modifier attachment in Dutch", *Memory & Cognition*, **30**:1 (2002), pp. 2002 ↑ 44, 45.

[12] Spivey M. J., Tanenhaus M. K., Eberhard K. M., Sedivy J. C., "Eye movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution", *Cognitive Psychology*, **45** (2002), pp. 447–481 ↑ 44.

[13] Altmann G. T. M., Steedman M., "Interaction with context during human sentence processing", *Cognition*, **30** (1988), pp. 191–238 ↑ 44, 45.

[14] Frazier L., "Sentence processing: A tutorial review", Attention and performance, **XII**, ed. M. Coltheart, Erlbaum, Hillsdale, 1987, pp. 559–586 ↑ 44.

[15] Crain S., Steedman M., "On not being led up the garden path: The use of context by the psychological syntax processor", *Natural language parsing: Psychological, computational, and theoretical perspectives*, eds. D. R. Dowty, L. Kartunnen, A. M. Zwicky, Cambridge University Press, Cambridge, 1985 ↑ 44, 45.

[16] Gillen K., *The comprehension of doubly quantified sentences*, Durham theses, Durham University, 1991, URL http://etheses.dur.ac.uk/1486/ ↑ 44.

[17] van Berkum J. J. A., Brown C. M., Hagoort P., "Early Referential Context Effects in Sentence Processing: Evidence from Event-Related Brain Potentials", *Journal of Memory and Language*, **41**:2, August (1999), pp. 147–182 ↑ 44.

[18] Юдина М. В., Федорова О. В., Янович И. С., «Синтаксическая неоднозначность в эксперименте и в жизни», *Компьютерная лингвистика и интеллектуальные технологии*, Труды международной конференции «Диалог 2007», Изд-во РГГУ, М., 2007 ↑ 45.

[19] Юдина М. В., «Что может помочь компьютеру понять, кто стоял на балконе», *Сборник трудов конференции «Диалог»*, М., 2010, с. 604–609 ↑ 45.

[20] Драгой О. В., *Разрешение синтаксической неоднозначности предложений с определительным придаточным в русском языке*, Дисс. ... канд. филол. наук, М., 2007 ↑ 45.

[21] Кормалев Д. А., Куршев Е. П., Сулейманова Е. А., Трофимов И. В., «Извлечение информации из текста в системе ИСИДА-Т», *Труды XI Всероссийской научной конференции RCDL'2009*, КарНЦ РАН, Петрозаводск, 2009, с. 247–253 ↑ 46.

[22] Сулейманова Е. А., Трофимов И. В., «О подходе к отождествлению сущностей в рамках задачи извлечения информации из текстов», *Программные системы: теория и приложения*, **4**:1(15) (2013), с. 15–30, URL http://psta.psiras.ru/read/psta2013_1_15-30.pdf ↑ 46.

[23] Шмелёв А. Д., *Русский язык и внеязыковая действительность*, Языки славянской культуры, М., 2002, 496 с. ↑ 51, 54.

[24] Кобзарева Т. Ю., Лахути Д. Г., Ножов И. М., «Модель сегментации русского предложения», *Труды международного семинара ДИАЛОГ'2001 по компьютерной лингвистике и ее приложениям*. т. 2, с. 185–194 ↑ 53.

Рекомендовал к публикации *к.т.н. Е. П. Куршев*

Об авторе:

Елена Анатольевна Сулейманова

Научный сотрудник Исследовательского центра искусственного интеллекта ИПС им. А. К. Айламазяна РАН, одна из разработчиков технологии построения систем извлечения информации.

e-mail: yes@helen.botik.ru

Образец ссылки на эту публикацию:

Е. А. Сулейманова. *О комплексном подходе к разрешению реляционно-аппозитивных неоднозначностей* // Программные системы: теория и приложения: электрон. научн. журн. 2014. Т. 5, № 4(22), с. 41–66.

URL http://psta.psiras.ru/read/psta2014_4_41-66.pdf

Elena Suleymanova. *An integrated approach to disambiguating relational-appositional constructions.*

Abstract. The paper suggests a method for solving a common type of syntactico-semantic ambiguity, to be used as part of information extraction. The approach combines various intrasentential disambiguation methods with those based on discourse constraints. (*In Russian*).

Key Words and Phrases: information extraction, syntactico-semantic ambiguity, discourse context.