РЕФЕРЕНЦИЯ ОБОЗНАЧЕНИЙ ПЕРСОН И ОРГАНИЗАЦИЙ В РУССКОЯЗЫЧНЫХ ТЕКСТАХ СМИ: ЭМПИРИЧЕСКИЕ ЗАКОНОМЕРНОСТИ ДЛЯ КОМПЬЮТЕРНОГО АНАЛИЗА

Ермаков А.Е.

ermakov@metric.ru

ООО "Гарант-Парк-Интернет", Москва

Доклад посвящен проблемам установления кореферентности обозначений персон и организаций при компьютерном анализе текста. Рассматриваются особенности употребления таких обозначений в тексте СМИ и эмпирические закономерности, с опорой на которые на практике удается провести достаточно достоверный анализ кореферентности. Описывается схема принятия решения при определении референтов, реализованная в программных продукта RCO.

Введение

Одной из актуальных прикладных задач, вставших перед компьютерной лингвистикой, является выделение в тексте упоминаний о персонах и организациях (далее – объектах). В полной постановке эта задача включает в себя две подзадачи:

a) распознавание и разбор наименований объектов с выделением всех элементов наименования (ФИО, организационно-правовая форма, форма хозяйственной деятельности, название, географические атрибуты и т.д.), что позволяет использовать результаты разбора для последующего отождествления наименований и приведения их к заданному каноническому виду;
b) отождествление различных вариантов наименования одного и того же объекта в тексте, в том числе косвенных обозначений, не содержащих в себе имени собственного, например, *президент*, *предприятие*, *которое*, *он*, *ее*.

Как уже было показано в работе [8], задача (a) является очень сложной и требует привлечения различных технических средств – словарного и бессловарного морфологического анализа с генерацией множества гипотез о лексико-грамматических разрядах и способах словоизменения известных и неизвестных слов, учета формальных правил возможного написания сложных наименований и весьма изощренной системы принятия решений, учитывающей предисторию появления имен в тексте для снятия омонимии. Однако и этих средств оказывается не достаточно и наше последнее технической решение задачи А в дополнение к описанному использует результаты синтаксического анализа, без которого просто не возможно распознать такие, к примеру, обозначения организаций, как *Ковдорский, Костомукшинский и Оленегорский ГОКи* – три объекта в одной синтагме.

Задача (b) в лингвистической терминологии носит название задачи установления кореферентности слов в тексте, т.е. тождества референтов слов. Лингвистической теории референции посвящено множество научных работ, как за рубежом [5], так и у нас в России [1-4]. Существуют конструктивные модели, теоретические обоснованные и работающие на практике, в частности, модель референциального выбора, реализованная с применением нейронной сети для принятия решения о выборе подходящего антецедента слова [6]. К сожалению, все известные автору практические результаты этих работ связаны с компьютерной обработкой текста на иностранных языках. Анализ общих критериев установления кореферентности в русском языке, в том числе позиционируемых как якобы пригодных для компьютерной обработки (например, в [7]), дал автору основания считать, что многие из них либо вообще не смогут устойчиво работать на практике, либо не будут полезны при обработке обозначений персон и организаций.

Настоящий доклад описывает эмпирически сформированную систему правил и общую алгоритмическую схему их применения, позволяющую на практике во многих случаях определить кореферентность обозначений персон и организаций в текстах такого «свободного» жанра, как публикации СМИ. Описанные закономерности не претендуют на теоретическую новизну, вследствие чего доклад позиционируется лишь как отчет о положительном опыте использования в практической задаче некоторых аспектов теории референции, достаточно простых для того, чтобы уже сейчас быть внедренными в коммерческую программу.

Особенности референции обозначений в тексте

Основные проблемы установления кореферентности возникают при обозначении объектов именами нарицательными и местоимениями, усугубляясь еще и тем, что слово может употребляться как референтно, так и нет. При нереферентном употреблении оно используется либо в предикатном значении (*Иванов поработал директором; Лукойл – известная компания*), либо в качестве обозначения всего класса возможных референтов, т.е. денотата (*предприятие(я) не должны заниматься теневой деятельностью*). Местоимение также референтно не всегда, так как может замещать слово в нереферентом употреблении. Только имена собственные референтны всегда. Соотнесенность словоупотребления с референтом характеризуется так называемым референтным (денотативным) статусом. Если при первом употреблении имени нарицательного в тексте его денотативный статус выражается явно (*компания, принадлежащая Ходорковскому*; *российский президент*), то при последующих употреблениях слова во избежание неоднозначности или повтора могут быть использованы специальные служебные слова – актуализаторы денотативного статуса, к которым относится ряд прилагательных и местоимений - *другой*, *любой*, *этот*, *его*, *один из*, и др. В большинстве случаев авторы опускают актуализаторы в тексте, предоставляя возможность читателю самому определить денотативный статус слова на основании экстралингвистических знаний и контекста. Как показывает анализ языкового материала, задача определения кореферентности на практике принципиально не разрешима в полном объеме, так как требует привлечения прагматической модели предметной области, которая фактически безгранична для текстов СМИ.

Тем не менее, как показал наш опыт, установление кореферентности возможно с высокой степенью достоверности для ряда случаев:

- Полное или краткое наименование, которое содержит в себе имя собственное (*ОАО "АКБ "Московский Деловой Мир"*, *Иванов И.П.*).
- Имя нарицательное при соблюдении ряда условий. Это существительное-классификатор, которое отражает определенные признаки референта, например, должность или род занятий персоны, организационно-правовую форму или форму хозяйственной деятельности предприятия, и может (к сожалению) выступать в роли признакового предиката или приложения при имени собственном, т.е. употребляться вообще не референтно, особенно часто во множественном числе, творительном падеже и в роли приложения-уточнения (см. примеры выше). С точки зрения нормы, принятой при косвенном обозначении в тексте персон и организации, между этими типами объектов существуют значительные отличия. Именование персон по должностям в общем случае встречается в текстах крайне редко, лишь для определенных категорий VIP – *президент*, *королева*, *министр*, причем должности первых лиц государства могут употребляться референтно вообще без упоминания имени собственного в тексте. Референтность такого упоминания обычно не обусловлена тем, была ли ранее в тексте введена должность персоны, и определяется лишь предположением о наличии у читателя экстралингвистических знаний. В итоге, косвенные обозначения персон именами нарицательными стоит отождествлять только с известными объектами, заданными словарным способом, что описано далее. Косвенное именование организаций (*компания*, *банк*) является нормой. При отождествлении таких обозначений следует учитывать возможность сложных синонимических замен. Так, например, любая организация, названная вначале *компанией*, *фирмой* или *заводом*, может в дальнейшем именоваться *предприятием*, однако, будучи изначально названа *заводом*, в дальнейшем не именуется *компанией* (и наоборот). Организация, введенная в текст без указания формы хозяйственной деятельности, в дальнейшем обозначается как *компания*, *фирма* или *предприятие*, но не как *завод* или *банк*, за исключением случаев, когда в самом имени собственном содержится указание на род деятельности (*Метпромбанк*).
- Относительное местоимение (*котор-ый,-ая,-ое,-ые*) в любой грамматической форме. Эти местоимения не имеют анафорических референтов и в норме кореферентны последней ближайшей именной группе из того же предложения, согласованной по роду-числу, и отделенной запятой, так как вводятся в придаточном предложении.
- Личное местоимение третьего лица (*он*, *она*, *оно*) в именительном падеже. Антропоцентрический характер языка и мышления отражается в том, что имена одушевленных предметов стремятся занять в предложении позицию субъекта – подлежащего. В результате на практике референтов личных местоимений в номинативе можно с высокой вероятностью отождествлять с одушевленными предметами (чаще персонами и реже организациями), упомянутыми в двух предшествующих предложениях и согласующихся по роду-числу. Референта местоимения в косвенном падеже достоверно установить нельзя, так как такое местоимение может быть кореферентно любому существительному из предыдущих предложений, и даже применение синтаксического анализа не позволяет снять

омонимию и установить род местоимения (*его*, *ему* и т.д. – и мужской, и средний род в одинаковых падежах).

Эмпирически были обнаружены следующие формальные правила, которым должно (бы) подчиняться построение связного текста для упрощения его восприятия. Данные правила легко проверяются при машинном анализе текста:

1) Референт может употребляться дважды в одном предложении только в составе двух разных пропозиций – базовой и осложняющей. В противном случае налицо семантическое противоречие - референт участвует в одной ситуации в различных ролях. Обычно это означает, что между упоминаниями одного референта в предложении должна стоять хотя бы одна запятая (не считая запятых между однородными членами).
2) Возможный референт слова при своем последнем упоминании не должен входить в состав группы однородных членов предложения (*Сидоров* столкнулся с Ивановым и Петровым в дверях, после чего *ему* не удалось избежать разговора). Это относится к референту слова, стоящего в единственном числе. Слово во множественном числе, напротив, может иметь несколько референтов в единственном числе в составе группы однородных (В дверях школьницы столкнулись с *Васей и Петей*, *которых* знали еще с детства).
3) При наличии нескольких потенциальных референтов слову более естественно иметь того референта, который употреблялся в теме предшествующего предложения, нежели в реме. Это связано с тем, что для автора наиболее естественно именовать кратко тот предмет, который уже находится в фокусе внимания – теме. (*Иванов* познакомился с Петровым в прошлом году. Тогда *он* впервые участвовал в выставке). При переводе же предмета из ремы в тему – выведении в фокус - естественно называть его более полным именем, чтобы избежать неоднозначности. Это отражает следующее правило.
4) Референт слова не должен упоминаться после него в том же предложении, будучи обозначен более полным наименованием (*Компания обанкротилась, после чего акционеры МММ тщетно пытались вернуть свои деньги* – если *компания* обозначает *МММ*, то фраза воспринимается анормально). То есть, денотативный статус обозначения не должен уточняться при повторном упоминании референтого объекта в предложении, так как референт уже находится в фокусе внимания.

К сожалению, на практике эти правила часто безболезненно нарушаются даже при построении стилистически грамотного текста ввиду того, что основной опорой при его восприятии полагаются все же экстралингвистические знания. Тем не менее, проверка данных правил позволяет принять решение в случае неоднозначности при поиске референта, как описано далее.

Обобщенный алгоритм анализа референции

Общая схема поиска референта слова в тексте, который анализируется последовательно по предложениям, следующая:

1) Определение всех атрибутов возможного референта, указанных при слове в предложении – фамилия, имя, отчество для персоны; имя и тип для организации (*компания "Мобильные телесистемы"*, *ООО "Орловский сталепрокатный завод"*); дополнительные атрибуты организации или персоны (*нефтяная компания*, *биржа металлов*, *американский президент*, *президент компании*), актуализаторы денотативного статуса. На основании словарной информации к некоторым атрибутам референта могут быть приписаны дополнительные значения – слова-синонимы, при помощи которых он может именоваться в других местах текста (*авиазавод = авиационный завод = предприятие*, *компания = фирма = предприятие*).
2) Определение денотативного статуса слова. При этом учитываются лексико-семантический разряд слова, найденные на этапе (1) атрибуты возможного референта, грамматические характеристики слова. Так, в рамках решаемых здесь задач ищутся референты только имен собственных, относительных и личных местоимений, нарицательных существительных из заданного словаря – возможных обозначений организаций и известных персон. Имена нарицательные во множественном числе, творительном падеже и в роли приложения-уточнения считаются нереферентными. Наличие актуализатора при слове может относить его к одной из трех категорий – референтом является подходящий объект, упоминавшийся ближайшим по тексту (*этот*, *вышеуказанный*); референт отсутствует (*другой*, *всякий*, *такой*) или референт есть, но практически не может быть установлен (*его*, *чей-то*, *некий*, *один из*, *тот*).
3) Поиск возможных референтов слова, ранее упоминавшихся в тексте, или известных словарных объектов. При этом проверяются необходимые и достаточные условия тождественности референтов, суть которых в том, что значения атрибутов определенного типа у них должны присутствовать и совпадать, а значения атрибутов других типов должны либо отсутствовать у одного из объектов-референтов, либо совпадать. В итоге, например, допускается то, что референт словосочетания *нефтяная компания* именуется дальше по тексту либо как *компания Юкос*, либо как *компания*,

либо как *российская нефтяная компания*, но не как *немецкая компания*. При этом, в зависимости от типа анализируемого обозначения объекта, допустимым считается тот референт, последнее упоминание которого отстоит не более, чем на заданное число предложений от текущего анализируемого упоминания. Так, для имен собственных референт ищется во всем тексте, для личных местоимений – в текущем предложении и в двух предложениях позади него, для относительных местоимений – только в текущем предложении. Существуют и другие особые случаи, интервалы для которых определены эмпирически.

4) При наличии более одного возможного референта проверяются правила построения связного текста, перечисленные выше, и референтом считается ближайший подходящий объект, если для него не нарушается ни оно из этих правил. В противном случае референт считается не установленным, так как вероятность ошибки чересчур велика.
5) Отождествление референта слова (если установлено, что таковой существует) с одним из ранее упоминавшихся объектов (или впервые упомянутым словарным), либо фиксация упоминания о новом объекте. При этом для объекта сохраняется информация о его последнем упоминании – номер предложения и номер слова в нем, показатели темы-ремы и наличия однородных. Атрибуты объекта пополняются новыми значениями, если таковые появились в данном упоминании, что позволяет иногда восстановить полное наименование объекта по ходу анализа текста. Важным является то, что информация обо всех референтных упоминаниях с неустановленным референтом также сохраняется с тем, чтобы учесть возможную неоднозначность далее. Так, если в предложении говорилось о компании *"Альфа"*, а также упоминалась другая компания, то при последующем упоминании слова *компания* будет учтена неоднозначность, которая будет решаться на этапе (4).

Рисунок 1 иллюстрирует обычное качество, с которым на основе описанных принципов программа выделяет упоминания обо всех заранее неизвестных персонах и организациях. Конечно, ошибки случаются, но не часто.

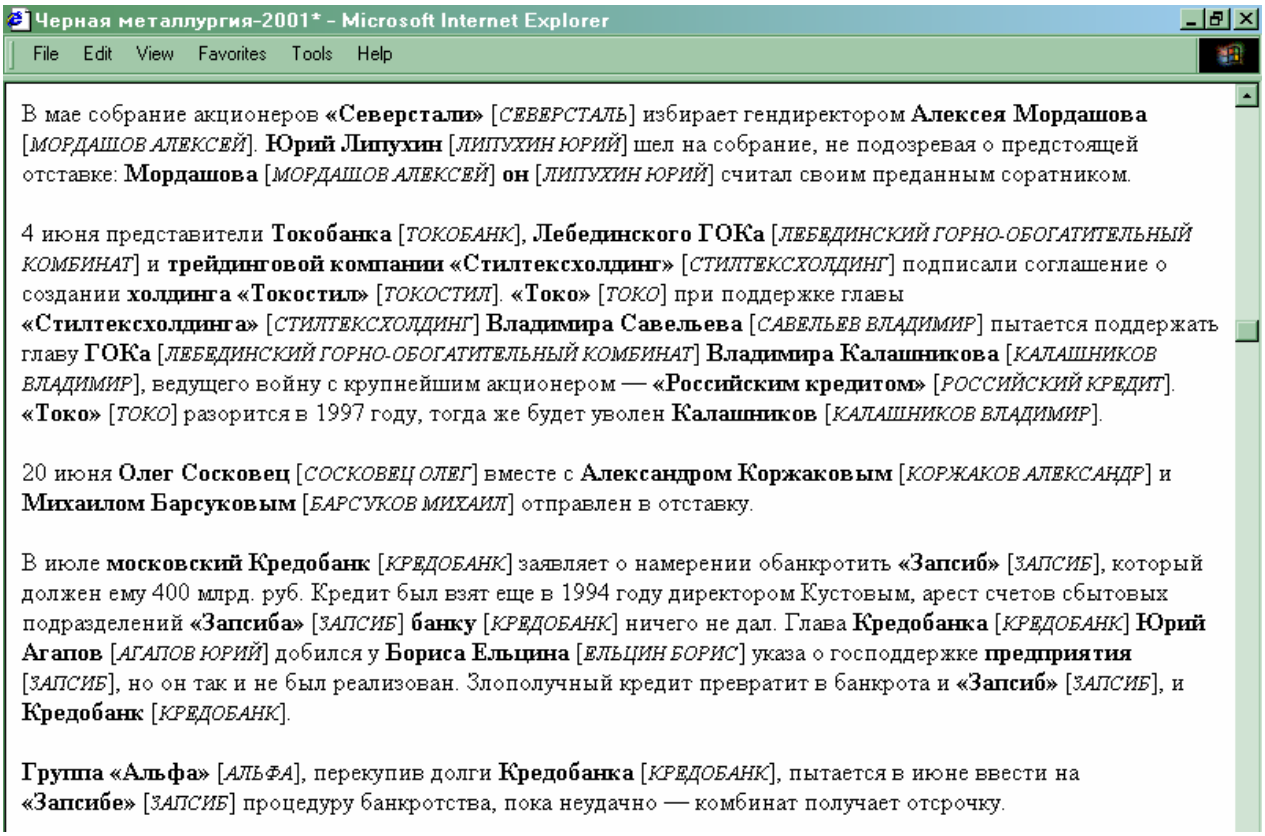

Рис. 1. Пример выделения программой всех упоминаний о заранее неизвестных объектах в тексте документа "Черная металлургия". После выделенного упоминания объекта в квадратных скобках программой указано имя соответствующей персоны или организации, приведенное к канонической форме. Как видно, программа не только выделила и привела к единообразному виду имена всех объектов, но и сумела отождествить местоимение "он" с "Липухиным Юрием", "ГОК" с "Лебединским горно-обогатительным комбинатом", "банк" с "Кредобанком", а "предприятие" – с "Запсибом"

Поиск референтов известных объектов

Повысить полноту, а в случае неоднозначности еще и точность отождествления кореферентных обозначений помогает словарная информация об известных объектах. Обработка упоминаний всех слов, относящихся к таким объектам, производится в соответствии с описанной выше схемой, в результате чего для них определяется денотативный статус и выясняется, относятся ли они к словарному объекту, или к каким-то другим – так разрешается омонимия.

К примеру, используемое нами сегодня словарное описание персоны в дополнение к ФИО содержит следующую информацию:

- Синоним - возможное обозначение объекта в тексте, обычно должность или прозвище. Например, синонимами к Владимиру Путину являются слова: *президент России*, *российский президент*, *наш президент*, *российский лидер*, *кремлевская власть*, *Кремль*. При нахождении в тексте такого обозначения оно считается референтным данному объекту, если при нем отсутствуют прочие показатели денотативного статуса. Так, *президент России* будет считаться обозначением Путина, если в тексте не сказано: *президент России Ельцин*, *новый президент России* и т.п.
- Контекстный синоним - в отличие от обычного синонима такое обозначение считается референтным данному объекту лишь в том случае, если ранее в тексте явно это объект упоминался явно (по ФИО или по синониму). Например, для Аллы Пугачевой контекстными синонимами могут быть слова: *певица*, *примадонна*, *звезда эстрады*, *звезда российской эстрады*. Так же, как и для обычных синонимов, для них проверяется денотативный статус в тексте: *звезда эстрады* будет считаться обозначением Пугачевой, если в тексте не сказано: *звезда эстрады София Ротару*, *известная звезда эстрады*.
- Референтный контекст – слова, присутствие которых в тексте позволяет снять неоднозначность – установить референта в случае, если объект обозначен только именем-отчеством, или если его обозначение (ФИО, синоним) может относиться к другим объектам. Например, присутствие в тексте таких слов, как *президентская гонка*, *“Единая Россия”* позволяет отождествить неизвестный объект, названный *Владимир Владимирович*, с Владимиром Путиным, если в тексте не сказано *Владимир Владимирович Маяковский* или нечто подобное. А в случае, если используются описания двух объектов – Путина и Ельцина, присутствие этих слов позволяет отождествить не указанного в тексте российского президента с Путиным, а присутствие слов *расстрел Белого дома* – с Ельциным.

Заключение

Изложенные в докладе общие принципы определения кореферентности подтвердили свои работоспособность на практике и использованы в коммерческой аналитической программе RCO Fact Extractor, производящей поиск в тексте описаний фактов, фигурантами которых являются персоны и организации. Демонстрационную версию программы можно получить на сайте http://www.rco.ru/.

Список литературы:

1) Арутюнова Н.Д. Предложение и его смысл. Москва, Наука, 1976.
2) Арутюнова, Н.Д. Язык и мир человека. М., 1998.
3) Падучева, Е.В. Высказывание и его соотнесенность с действительностью. М., 1985.
4) Лебедев М.В., Черняк А.З. Онтологические проблемы референции. М., "Праксис", 2001.
5) Studies in Anaphora / ed. Barbara Fox. Amsterdam: Benjamins, 1996.
6) Кибрик А.А. Референция, рабочая память и нейронные сети: о взаимодействии лингвистики с психологией и когнитивной наукой // Материалы Первой Российской интернет-конференции по когнитивной науке (http://www.auditorium.ru/conf/conf_fulltext/cognitio/kibrik.pdf)
7) Кобзарева Т. Ю. Проблема кореференции в рамках поверхностно-синтаксического анализа русского текста // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог’2003. – Москва, Наука, 2003.
8) Ермаков А.Е., Плешко В.В. Компьютерная морфология в контексте анализа связного текста // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог’2004. – Москва, Наука, 2004.