Автоматическое разрешение референции в русскоязычных новостных текстах

Ерин А.Н. 425 группа

Понятия и определения

Референ́ция — отнесённость актуализованных (включённых в речь) имён, именных групп или их эквивалентов к объектам внеязыковой действительности (референтам, денотатам).

Референ́т — объект внеязыковой действительности, который имеет в виду говорящий в контексте конкретной языковой ситуации; предмет референции.

Ана́фора, анафори́ческое отноше́ние — отношение между языковыми выражениями (словами или словосочетаниями), при котором в смысл одного выражения входит отсылка к другому, ранее упомянутому языковому выражению

Кореферéнтность — отношение между именами, имеющими один референт; то есть отношение между компонентами высказывания, которые обозначают один и тот же объект внеязыковой действительности.

Постановка задачи

По сегодняшний день существует актуальная проблема обработки естественно-языковых текстов.
При проведении семантического анализа текста одной из проблем является задача разрешения референции, т.е. определения реальных объектов по каким-либо словам-указателям.
В отличие от обработки англоязычных текстов, для русскоязычных данная проблема развита слабо.

Например:

*«**Иванов** разбил очки **Петрову**, за это **его** наказали.»*

Семантическая задача: определить кого наказали?

Постановка задачи

Два типа анафор

1) Представленные существительным или группой существительных:

*«**Президент Медведев** за дальнейшее сокращение часовых поясов.*
***Дмитрий Медведев** сегодня заявил, что считает возможным дальнейшее сокращение часовых поясов в России.*
***Президент** напомнил, что уже принят ряд решений по переводу пяти субъектов России в новые для них часовые пояса.»*

Для данного типа характерной проблемой является определение наличия анафорического отношения

Постановка задачи

Отношение присутствует:

*«**Президент Медведев** за дальнейшее сокращение часовых поясов.*
***Дмитрий Медведев** сегодня заявил, что считает возможным дальнейшее сокращение часовых поясов в России.*
***Президент** напомнил, что уже принят ряд решений по переводу пяти субъектов России в новые для них часовые пояса.»*

Отношение отсутствует (абстрактное обозначение объектов или типов объектов):

*«**Президент** — выборная должность главы государства»*
*«Перед вступлением на должность **президент** обязан принять присягу государству»*

Постановка задачи

2) Представленные местоимением:

А) Личные местоимения (я, ты, Вы, он, она, …):

*«**Я** категорически против вступления России в ВТО» сказал **глава КПРФ***

Б) Возвратное местоимения (себя, себе, собой, собою):

*«**Я** купил **себе** машину»*

В) Притяжательные местоимения (мой, твой, наш, Ваш, его, …)

*«**Ваш** автомобиль превысил скоростной режим» сказал инспектор **водителю**.*

Постановка задачи

Г) Вопросительные местоимения (какой, каков, чей, который)

«***Какая планета*** *третья от Солнца?*»

Д) Указательные местоимения (этот, это, тот, такой, таков)

«**Этот пример** не самый подходящий»

Для остальных типов местоимений (определительные, отрицательные , неопределённые) согласованность с существительным (или выражением) может быть опущена или отсутствовать вовсе.

Проблемы и сложности

Основные проблемы и сложности обработки русскоязычных текстов (в том числе и разрешения референций) возникают из-за возможной многогранности семантических форм для единственной синтаксической конструкции отдельного предложения или фразы.

*«Простой(**прил**.) солдат(**ед. ч., им. п.**)»*
*«Простой(**сущ**.) солдат(**мн. ч., род. п.**)»*

Сложности установления кореферентности возникают при обозначении объектов именами нарицательными и местоимениями, усугубляясь еще и тем, что слово может употребляться как референтно, так и нет (как было показано в примерах выше).
Имена собственные референтны всегда.

Методы и подходы

Первым шагом, который присутствует во всех принципах и подходах разрешения референции, является определения кандидатов-референтов по номинационным свойствам:

- число и род
- одушевленность/неодушевленность
- и.т.д.;

То есть эти свойства у антецедента(референта) и его анафоры должны совпадать или по крайней мере не различаться.

Кандидатами могут быть только слова и фразы из данного или предшествующих предложений текста.

Данный шаг является чисто техническим и не использует каких-либо эвристик, поэтому число кандидатов может быть очень велико.

Методы и подходы

Эвристические подходы

Общим подходом является оценка по расстоянию и местоположению: выбирается ближайший объект выше по тексту.

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:

- для имен собственных ищется во всем тексте
- для личных местоимений - в текущем предложении и в двух предложениях позади него
- для относительных местоимений - только в текущем предложении.

Методы и подходы

Двойное употребление референта в одном предложение
- только в составе двух разных пропозиций (базовой и осложняющей), т.е. разделяются запятой
- иначе имеется семантическое противоречие (референт участвует в одной ситуации в различных ролях)

Референт в единственном числе при последнем своем 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»*

Методы и подходы

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»*.

Наиболее вероятное наличие референта в предшествующем предложение, нежели в реме:
*«**Иванов** познакомился с Петровым в прошлом году. Тогда **он** впервые участвовал в выставке»*.

Методы и подходы

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.

На практике все эти правила могут нарушаться, но тем не менее помогают в ряде случаев устранить неоднозначность выбора.

Методы и подходы

Подходы для конкретных ситуаций

Референт личного местоимения третьего лица
- два предыдущих предложения,
- одушевленные существительные,
- согласование по роду-числу

Употребление в косвенном падеже
- любое существительное

Методы и подходы

Относительное местоимение (*котор-ый,-ая,-ое,-ые*)

- не имеют анафорических референтов
- кореферентны последней ближайшей именной группе из того же предложения, согласованной по роду-числу, и отделенной запятой.

Имя нарицательное это существительное-классификатор

- отражает определенные признаки референта *(должность или род занятий персоны, организационно-правовую форму или форму хозяйственной деятельности предприятия)*
- может употребляться вообще не референтно *(во множественном числе, творительном падеже и в роли приложения-уточнения)*

Методы и подходы

Косвенное обозначение персон и организаций

Обозначение персон

- не именуются по должностям
- исключение для определенных категорий VIP

("президент", "королева", "министр" и даже без упоминания имени собственного в тексте)

Обозначение организаций

- является нормой

*(**Сбербанк** предупредил о возможных технических сбоях, теперь клиентам **банка** надо работать с банкоматами с особой осторожностью)*

Методы и подходы

Наличие актуализатора при слове определяет одну из трех его категорий

- референтом является подходящий объект, упоминавшийся ближайшим по тексту (*этот, вышеуказанный*)
- референт отсутствует (*другой*, *всякий, такой*)
- референт есть, но практически не может быть установлен (*его*, *чей-то*, *некий, один из, тот*).

Методы и подходы

Общий план поиска референтов

1)Определение грамматический атрибутов возможного референта *(род, число, и.т.д.)*
2)Определение семантических атрибутов указанных при слове в предложении
(фамилия, имя, отчество для персоны; имя и тип для организации (компания "Мобильные телесистемы", ООО "Орловский сталепрокатный завод")).
3) На основании словарной информации к некоторым атрибутам референта приписываются дополнительные значения - слова-синонимы
(авиазавод = авиационный завод = предприятие, компания = фирма = предприятие).

Методы и подходы

4) Определение денотативного статуса слова, учитывая:
- лексико-семантический разряд слова,
- найденные на этапе (3) атрибуты возможного референта
- грамматические характеристики слова.

Поиск происходит для
- имен собственных
- относительных и личных местоимений
- нарицательных

Имена нарицательные во множественном числе, творительном падеже и в роли приложения-уточнения считаются нереферентными.

Методы и подходы

5) Поиск возможных референтов слова, ранее упоминавшихся в тексте, или известных словарных объектов.

Проверяются необходимые и достаточные условия тождественности референтов

- значения атрибутов определенного типа у них должны присутствовать и совпадать;
- значения атрибутов других типов должны либо отсутствовать у одного из объектов-референтов, либо совпадать

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****.*

6) При наличии более одного возможного референта выбирается ближайший подходящий.

Заключение

Основные предложенные методы и подходы взяты из публикации Ермакова А.Е., которые были реализованы в коммерческих продуктах с закрытым кодом компании RCO. На данный момент единственной известной реализации методов разрешения референции для русскоязычных текстов.

http://www.rco.ru/
http://www.rco.ru/article.asp?ob_no=2339

В рамках собственной курсовой работы планируются опробовать предложенные методы разрешения референции для местоимений.
На данный момент опробован чисто технический метод (без эвристик), показавший необходимость применения дополнительных методов для сокращения количества возможных кандидатов-референтов.

Литература

Лебедев М.В., Черняк А.З. Онтологические проблемы референции. М., "Праксис", 2001

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

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 - С. 185-190.