Поиск референциальных отношений между информационными объектами в процессе автоматического анализа документов

© А. С. Серый © Е. А. Сидорова
Институт систем информатики им. А.П. Ершова СО РАН, Новосибирск
Alexey.Seryj@iis.nsk.su lena@iis.nsk.su

Аннотация

Предлагается подход к установлению референциальных связей между информационными объектами, получаемыми в результате автоматической обработки текстов абстрактным анализатором. Рассматриваются меры сходства, зависимые от класса объектов, набора определенных ключевых и второстепенных атрибутов, связей с другими объектами и расстоянием между объектами в тексте.

Работа выполняется при финансовой поддержке Президиума РАН (интеграционный проект СО РАН № 15/10 «Математические и методологические аспекты интеллектуальных информационных систем») и РФФИ (грант №12-07-31216).

1 Введение

Одной из актуальных задач, стоящих перед компьютерной лингвистикой, является выделение в текстовых документах упоминаний о различных сущностях: персонах, организациях, событиях, местах и пр., а также существующих между ними связей. Перечень таких сущностей, информация о которых извлекается из текста, зависит от предметной области (ПО). Извлекаемые данные унифицируются в виде сети формальных описаний, так называемых информационных объектов (ИО), с целью дальнейшего хранения в базе данных (БД). Каждый информационный объект соответствует некоторому понятию/отношению предметной области и имеет заданную структуру. В дальнейшем будем полагать, что обработка текста производится в рамках некоторой информационной системы, предметная область которой ограничена и явно описана на определенном формальном языке. Важными элементами автоматической обработки текста (АОТ) являются установление анафорических связей и отождествление различных наименований одного и того же объекта, например многократных упоминаний какой-либо персоны в том или ином контексте.

Разрешение анафоры – довольно серьезная задача, в решение которой вовлечено множество исследователей, придерживающихся различных точек зрения на проблему и использующих различные подходы: как традиционные (синтаксические и семантические), так и альтернативные (статистические), дающие лишь приблизительный результат [1,3]. Задача отождествления различных наименований одного и того же объекта является более общей, поскольку подобные наименования могут не ссылаться друг на друга, как в случае анафоры, но, тем не менее, также являться кореферентными[1].

Для достижения больших полноты и точности результата разработчики систем АОТ стараются использовать дополнительные источники информации о терминах, такие как словари и базы знаний. На сегодняшний день существует множество подобных ресурсов, большую часть которых составляют англоязычные ресурсы. Так, подход к разрешению кореференции, разработанный исследовательской группой Стэнфордского университета, предполагает использование Википедии для выявления этнохоронимов [4]. Сам подход основан на совместном применении нескольких простых фильтров. Система, разработанная на основе этого подхода, на данный момент уже расширена новыми фильтрами [2]. Два из пяти новых фильтров, предложенных в [2], используют внешние ресурсы, такие как WordNet [8], Wikipedia и Freebase [5]. Проекты подобные WordNet и Freebase лучше всего развиты для английского языка, что существенно влияет на исследования в области обработки англоязычных текстов.

Тем не менее, несмотря на декларируемую важность и актуальность упомянутых выше задач для процесса АОТ, в силу своей сложности вообще и для русскоязычных текстов в частности, они не всегда решаются целиком. Так, например, схема, описанная в [6] охватывает упоминания персон и организаций, а задача разрешения анафоры,

Труды 14-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2012, Переславль-Залесский, Россия, 15-18 октября 2012 г.

Рис. 1 *Схема процедуры идентификации объектов*

решаемая в [9] и [11], ограничивается местоимениями.

Не охваченные в процессе АОТ случаи могут послужить причиной появления информационных объектов, собранных на основе кореферентных выражений. Такие информационные объекты мы по аналогии будем называть кореферентными или тождественными; в данном случае это будет означать, что два (или более) объекта содержат различные части информации о некоей внеязыковой сущности ПО.

В статье предлагается подход к установлению кореферентности уже не языковых выражений, а информационных объектов, собранных на их основе. Информационные объекты воспринимаются как гипотезы о реальном объекте заданной ПО.

Предлагаемый подход позволяет абстрагироваться от технологии обработки текста, налагая некоторые требования лишь на формат самих ИО, определяемые способом описания онтологии ПО [7] (такие как разделение атрибутов на ключевые и второстепенные; наличие только бинарных отношений между объектами и др.)

2 Поиск и идентификация информационных объектов

Задача установления референциальных отношений между ИО рассматривается здесь в контексте другой более объемной задачи идентификации объектов – разрешения контекстной омонимии, являющейся одним из побочных эффектов АОТ. Контекстная омонимия проявляется в наличии двух и более вариантов отождествления полученных из текста информационных объектов с объектами базы данных информационной системы. Проблема идентификации объектов и метод ее решения описаны в [10]. На рис. 1 приведена общая схема метода.

Кратко, процесс идентификации ИО состоит из следующих этапов:

- Первичный анализ. Извлеченные из документа ИО попадают в компонент первичного анализа, где проходят проверки на наличие кореференции и на совпадение по кортежу ключевых атрибутов с объектами БД. Те ИО, для которых удалось установить совпадение с единственным объектом БД или для которых набор ключевых атрибутов полностью определен, считаются идентифицированными.
- Идентификация. Оставшиеся ИО попадают в компонент идентификации, где коллекции наиболее близких к ним объектов БД, при необходимости расширяемые по иерархии классов онтологии и/или другими отношениями онтологии, подвергаются фильтрации.
- Расчет достоверности. Разрешение противоречий при наполнении БД между старыми и новыми данными, посредством вычисления специального параметра, количественно выражающего достоверность того или иного атрибута или связи.

3 Разрешение кореференции на уровне объектов

Процесс разрешение кореференции является частью процедуры идентификации объектов, сосредоточенной в компоненте первичного анализа. Алгоритм установления кореференции или референциального тождества объектов включает в себя установление степени сходства объектов, построение множества гипотетических эквивалентов для каждого объекта и объединение действительно кореферентных объектов.

3.1 Степень сходства информационных объектов

Чтобы сделать выводы о наличии или отсутствии референциального тождества между теми или иными объектами, необходимо каким-то образом

сопоставить их друг другу, сравнить их атрибуты и связи. Необходима мера, выражающая степень сходства двух объектов. Введем такую меру и назовем ее коэффициентом сходства информационных объектов $SI(q^1, q^2)$ (similarity index), где q^1 и q^2 – объекты, которые нужно сравнить.

Величина коэффициента сходства зависит от аргументов и параметра $0 \le k \le 1$ и вычисляется по следующей формуле:

(1) $$SI(q^1, q^2) = \begin{cases} k \cdot SI_C + (1-k) \cdot SI_L;\ SI_c \neq 0 \\ 0;\ SI_c = 0 \end{cases}$$

Согласно принятой ранее договоренности, предметная область описывается некоторой онтологией $\mathfrak{O}$, а информационные объекты и их отношения являются экземплярами классов ее понятий. Одно из ограничений, налагаемых на ИО, заключается в том, что объект (отношение) может быть экземпляром единственного класса. Значение подвыражения $SI_C = SI_C(q^1, q^2)$ характеризует зависимость величины $SI(q^1, q^2)$ от онтологии, а именно – взаимного расположения классов понятий, экземплярами которых являются объекты q^1 и q^2, в ее иерархическом древе. Можно сказать, что SI_C – это степень сходства онтологических классов объектов q^1 и q^2: если $C(q^1)$ и $C(q^2)$ – классы объектов q^1 и q^2 соответственно, то $SI_C = SI_C(C(q^1), C(q^2))$. Аналогично, $SI_L = SI_L(q^1, q^2)$ можно назвать степенью сходства кортежей атрибутов и связей q^1 и q^2. Коэффициент k регулирует уровень влияния онтологических и атрибутивно-реляционных факторов на итоговую величину $SI(q^1, q^2)$. Его значение определяется экспериментальным путем и может изменяться в зависимости от задачи.

Рассмотрим подробнее каждое из подвыражений формулы **(1)**.

(2) $$SI_C = SI_C(q^1, q^2) = \frac{|\mathcal{C}^1 \cap \mathcal{C}^2|}{|\mathcal{C}^1 \cup \mathcal{C}^2|}$$

$$\mathcal{C}^1 = \{C(q) | H(C(q), C(q^1)) \vee C(q) = C(q^1)\};$$

$$\mathcal{C}^2 = \{C(q) | H(C(q), C(q^2)) \vee C(q) = C(q^2)\}$$

Здесь $C(q)$ – класс онтологии $\mathfrak{O}$, экземпляром которого является объект q, H – бинарное отношение на множестве классов, такое что $C_1 H C_2 \Leftrightarrow C_1$ является предком C_2 и, таким образом, $\mathcal{C}^i$ – это множество классов, лежащих в иерархическом древе онтологии выше класса $C(q^i)$, плюс сам класс $C(q^i)$. Последнее гарантирует непустоту множеств $\mathcal{C}^i$ и, как следствие, ненулевое значение знаменателя в формуле **(2)**. Выражение SI_L, в свою очередь, раскладывается на два подвыражения SI_{LA} и SI_{LR}, характеризующие зависимость соответственно от атрибутов и связей объектов.

(3) $$SI_L = SI_L(q^1, q^2) = \frac{1}{2}(SI_{LA} + SI_{LR})$$

Из формулы **(3)** можно видеть, что атрибуты и связи объектов в одинаковой степени влияют на значение SI_L.

(4) $$SI_{LR} = SI_{LR}(q^1, q^2) = \begin{cases} \frac{|R_{EQU}|}{|R^2|},\ |R^2| > 0 \\ 0, |R^1| > 0 \text{ и } |R^2| = 0 \\ 1, |R^1| = 0 \text{ и } |R^2| = 0 \end{cases}$$

Здесь R^1 и R^2 – множества связей объектов q^1 и q^2 соответственно, $R_{EQU} = \{(r^1, r^2) | r^1 \in R^1 \& r^2 \in R^2 \& C(r^1) = C(r^2) \& \exists q: ((r^1(q^1, q) \& r^2(q^2, q)) \vee (r^1(q, q^1) \& r^2(q, q^2)))\}$. Другими словами, R_{EQU} – это множество пар отношений из R^1 и R^2, связывающих q^1 и q^2 с одним и тем же объектом q, онтологические классы которых тождественны.

(5) $$SI_{LA} = SI_{LA}(q^1, q^2) = f \cdot SI_{LA}^K + (1-f) \cdot SI_{LA}^A,$$
$$0 \le f \le 1$$

Атрибуты объектов поделены на ключевые и второстепенные. Кортеж ключевых атрибутов однозначно идентифицирует объект в информационном пространстве системы. Значение ключевого атрибута не может быть неопределенным или множественным (это тоже одно из ограничений на формат ИО). На второстепенные атрибуты это не распространяется. Будет естественным предположить, что влияние ключевых атрибутов на величину коэффициента близости должно отличаться от влияния второстепенных. Поэтому выражение SI_{LA} можно разложить еще на подвыражения SI_{LA}^K и SI_{LA}^A, для ключевых и второстепенных атрибутов соответственно. Коэффициент *f*, аналогично коэффициенту *k* из формулы **(1)**, получен из эксперимента и регулирует степень участия различных типов атрибутов.

(6) $$SI_{LA}^K = SI_{LA}^K(q^1, q^2) = \begin{cases} \frac{|K_{EQU}|}{|K^2|},\ |K^2| > 0 \\ 0, |K^1| > 0 \text{ и } |K^2| = 0 \\ 1, |K^1| = 0 \text{ и } |K^2| = 0 \end{cases}$$

(7) $$SI_{LA}^A = SI_{LA}^A(q^1, q^2) = \begin{cases} \frac{|A_{EQU}|}{|A^2|},\ |A^2| > 0 \\ 0, |A^1| > 0 \text{ и } |A^2| = 0 \\ 1, |A^1| = 0 \text{ и } |A^2| = 0 \end{cases}$$

Формулы **(6)** и **(7)** аналогичны **(4)**: K^1, K^2, K_{EQU} – соответственно множества ключевых атрибутов объектов q^1, q^2 и множество атрибутов из K^1, K^2, значения и типы которых совпадают. A^1, A^2, A_{EQU} – аналогично для второстепенных атрибутов.

3.2 Вычисление множества гипотетических эквивалентов

Из формул **(4)**, **(6)** и **(7)** очевидно следует, что операция вычисления $SI(q^1, q^2)$ в общем виде коммутативной не является, поэтому правильнее будет говорить, что $SI(q^1, q^2)$ вычисляет степень сходства объекта q^2 с объектом q^1. Объект q^1 при этом называется эталоном, а q^2 – кандидатом. Таким образом, выражение **(1)** сопоставляет объект-кандидат объекту-эталону и вычисляет степень их сходства.

Перейдем непосредственно к описанию процесса установления референциального тождества инфор-

мационных объектов. Каждый объект необходимо проверить на наличие эквивалента – ближайшего кореферентного ему объекта. Объект, соответствующий самому первому упоминанию, будем называть G-эквивалентом (от *global*, т.к. G-эквивалент является вершиной референциальной цепочки объектов).

Чтобы найти референциальные связи объекта, либо убедиться в том, что их не существует, следует построить и проанализировать множество гипотетических эквивалентов. Это множество $Pr(q)$ определяется следующим образом:

(8) $Pr(q) = \{q' \in Ctx(q) | SI(q', q) > \alpha > 0\}$,

где $Ctx(q)$ – это контекст объекта q, а α – положительное число, задающее нижнюю границу значений коэффициента близости, при которых q' может считаться вероятным эквивалентом объекта q. Значение параметра α зависит от характеристик q. Множество $Pr(q)$ содержит все объекты из некоторого контекста объекта q, степень сходства с которыми у объекта q больше некоторого положительного числа. Размер контекста зависит от того, словарные единицы какого типа участвовали в сборке объекта q: имена собственные, имена нарицательные, личные местоимения и т.п.

Разрешая кореференцию «извне», мы не имеем доступа к источнику порождения того или иного объекта, однако можем судить о нем по ряду косвенных признаков, определенных в результате экспериментальных исследований. Например, если можно считать, что объект был порожден упоминанием имени собственного, то в качестве контекста необходимо охватить всю предшествующую этому упоминанию часть документа, представленную коллекцией извлеченных из нее фактов. Когда источником послужило имя нарицательное либо местоимение, контекст ограничивается несколькими ближайшими объектами, такими, что их онтологические классы расположены в той же ветви иерархического древа, что и $C(q)$. В тех случаях, когда контекст q включает объекты по обе стороны от него, гипотетический эквивалент может как линейно предшествовать q, так и следовать за ним. В лингвистике отношение между выражениями, аналогичное второму случаю, называется катафорой и его исследование выходит за рамки данной работы. Следовательно, для установления эквивалента объекта q необходимо убедиться, что не существует эквивалента справа. Эквивалентом объекта q считается ближайший к нему объект q' из множества $Pr(q)$ с максимальным либо близким к максимальному значением $SI(q', q)$. Если таковой отсутствует или не является предшествующим объекту q, то говорим, что объект q был упомянут в тексте впервые. В противном случае, существует объект q', предшествующий q и кореферентный ему. Объект q помечается как эквивалентный объекту q'. Если в дальнейшем будет обнаружен некий объект q'', для которого эквивалентом является q, он должен быть помечен как эквивалентный объекту q. Подобная разметка организует цепочку объектов, где каждый следующий элемент референциально тождественен предыдущему – референциальную цепочку.

3.3 Объединение информационных объектов

В результате список объектов O_Q размечается в соответствии с выявленными связями как показано на рис. 2.

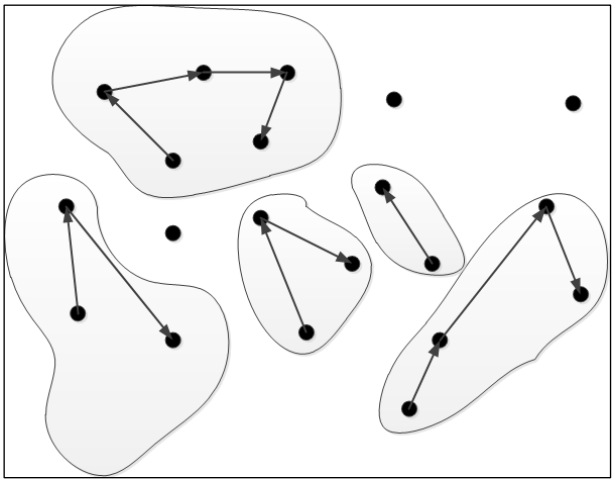

Рис. 2 *Разметка множества объектов*

Пусть $\mathfrak{R}$ – бинарное отношение на множестве информационных объектов, q^1 и q^2 – объекты, и пусть $q^1 \mathfrak{R} q^2 \Leftrightarrow q^1$ и q^2 признаны референциально тождественными. Очевидно, отношение $\mathfrak{R}$ является отношением эквивалентности и, следовательно, оно разбивает множество O_Q на непересекающиеся подмножества-кластеры. Классы эквивалентности по отношению $\mathfrak{R}$ совпадают с компонентами связности графа на рис.2. Вся информация, содержащаяся в элементах кластера, объединяется в одном объекте, называемом узловым объектом или узлом. Узловым считается G-эквивалент цепочки объектов. Очевидно, что таковой всегда найдется (рис.3).

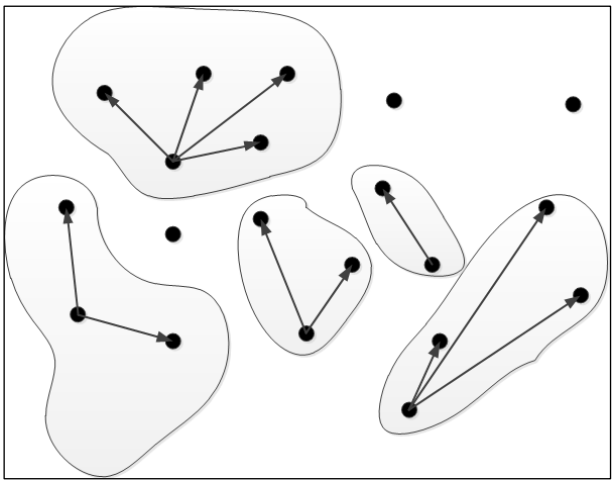

Рис. 3 *Узловые объекты*

Объединяя объекты в соответствии со связями, показанными на рис. 3, мы добиваемся того, что мощность множества объектов становится равной

мощности множества классов эквивалентности по отношению $\Re$. Ясно, что $|O_Q/\Re| \leq |O_Q|$.

3.4 Пример

Для иллюстрации рассмотрим предметную область компьютерной лингвистики, формально описанную с помощью онтологии, представленной в [7]. Для практических испытаний нашего метода и получения первых приблизительных значений k и f был разработан редактор объектов со встроенными механизмами вычисления коэффициента сходства любых двух выбранных объектов q^1, q^2 и разрешения референциальных связей в заданном множестве ИО. Для эксперимента были выбраны краткие новостные сообщения и информационные письма с конференций. Рассмотрим в качестве примера фрагмент информационного сообщения с сайта ИАиПУ ДВО РАН, посвященного конференции «Философия, математика, лингвистика: аспекты взаимодействия-2009» (http://www.iacp.dvo.ru/is/events.php?eid=226). Выбранный фрагмент, объемом 409 слов, содержит все референциально тождественные объекты. Объем всего текста составляет 632 слова, число извлеченных объектов – 40. Объекты, не вошедшие в данный фрагмент, не оказывают принципиального влияния на результат разрешения референциальных связей. Были извлечены экземпляры онтологических классов: **Географическое место**, **интернет-ресурс**, **Научное Мероприятие**, **Организация** и **Персона**. Всего из фрагмента получено 12 объектов, из них три объекта класса **Научное Мероприятие** и два объекта класса **Организация** референциально тождественны.

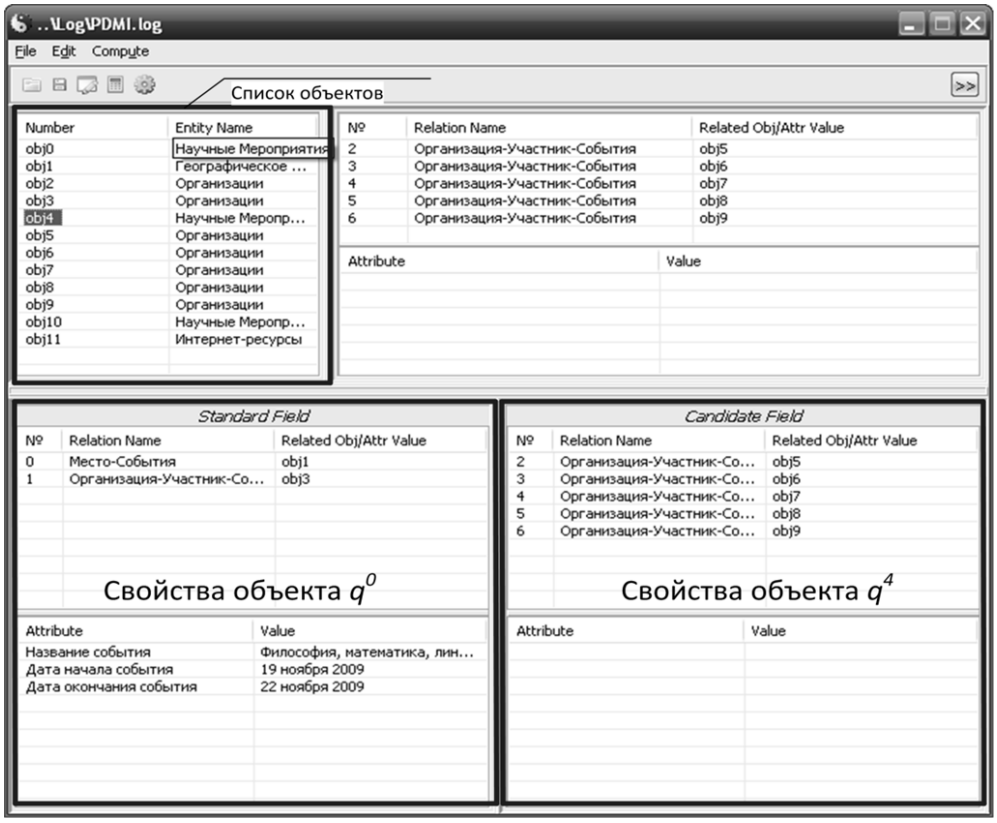

Рис. 4 *Главное окно редактора объектов*

На (Рис. 4) представлен начальный список объектов в окне редактора в порядке их встречаемости в тексте. Известно, что референциально тождественными являются семейства (0, 4, 10: научные мероприятия) и (3, 5: организации). Класс **Научное Мероприятие** имеет 4 атрибута: *дата основания*, *статус*, *частота проведения* и *язык* и еще 4 атрибута наследуются от родительского класса **Событие:** *дата начала события*, *дата окончания события*, *название события* и *описание события*. Класс **Организация** имеет 9 атрибутов: *e-mail*, *аббревиатура*, *адрес*, *дата основания*, *название организации*, *описание организации*, *телефон*, *тип организации*, *факс*. Как видно из (Рис. 4) у объекта q^4 не определен ни один из атрибутов, при этом имеются 5 связей. Согласно пункту **3.2** из этого следует, что эквивалент данного объекта существует и его следует искать среди нескольких ближайших объектов того же типа. Этому условию в полной мере отвечает объект q^0. Очевидно, что $SI_C(q^0, q^4) = 1$, так как оба объекта являются экземплярами класса **Научное Мероприятие**, следовательно объект q^0 является эквивалентом объекта q^4. Рассмотрим теперь объекты q^3 и q^5. Из (Рис. 5) можно видеть, что у них обоих определен ключевой

атрибут *название организации* и значения совпадают. Также, объекты q^3 и q^5 связаны отношением **Организация-Участник-События** с объектами q^0 и q^4 соответственно. Так как список объектов анализируется сверху вниз на момент встречи объекта q^5мы уже знаем, что q^0 и q^4 референциально тождественны. Таким образом $SI(q^3, q^5) = 1$ и очевидно, что они референциально тождественны.

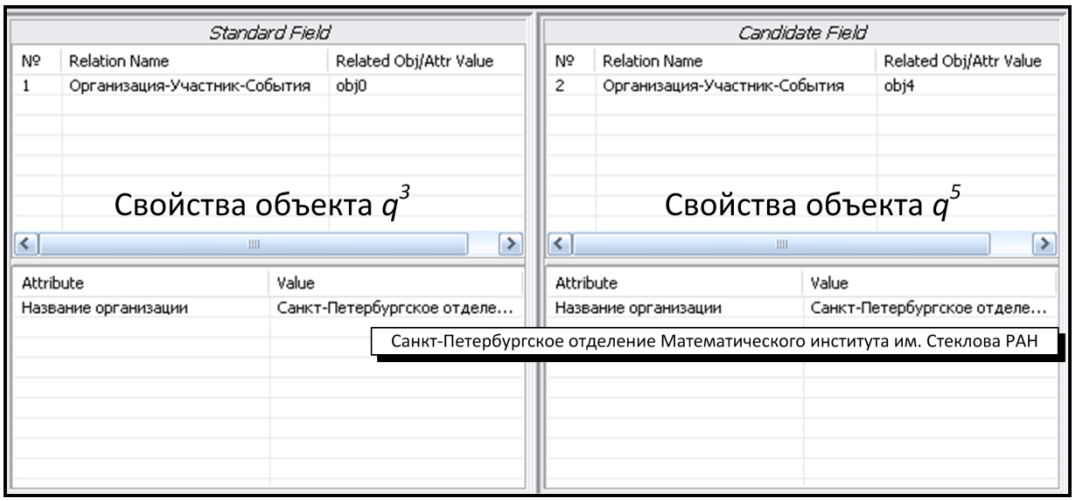

Рис. 5 *Сравнение свойств объектов q^3 и q^5*

В заключение заметим, что референциальное тождество объектов q^0 и q^{10} так и не было установлено. При текущих значениях коэффициентов k и f $SI(q^0, q^{10}) \approx 0.406$. В то же время у объекта q^{10} определены два значения атрибута *язык*, что не позволяет сделать однозначного вывода о языковых выражениях, явившихся его источником. В таких условиях имеющегося значения $SI(q^0, q^{10})$ недостаточно для того, чтобы считать объект q^0 эквивалентом q^{10} (объект q^4 не может считаться эквивалентом, так как он не имеет атрибутов). В итоге информация о том, что официальные языки конференции русский и английский, оказалась недоступна рис. 6.

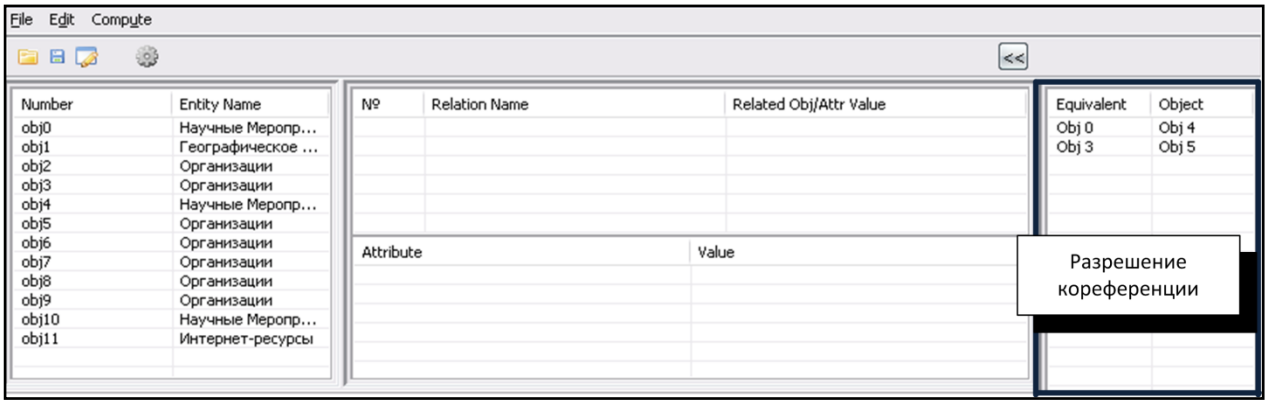

Рис. 6 *Результат разрешения референциальных связей*

5 Заключение

Основной целью поиска референциально тождественных объектов является сокращение числа ИО, представляющих одну сущность, в идеале до одного, что, в свою очередь, повышает вероятность их успешной идентификации. Однако ошибочное объединение объектов заметно снижает итоговую эффективность процедуры идентификации и даже может послужить причиной некорректного отождествления с объектом базы данных. По этой причине во главу угла выведена точность, и объединение объектов производится лишь в самых очевидных случаях зависимости. Имеющиеся на данный момент результаты позволяют считать, что подобный подход достаточно эффективен для документов рассматриваемой нами тематики (информационные письма, новостные сообщения о школах и конференциях, краткие статьи по компьютерной лингвистике). Более общие случаи нуждаются в дополнительных экспериментальных проверках. Заметим, что на полноту результата имеет возможность повлиять эксперт: для этого в описанном редакторе доступны функции корректировки объектов и их объединения «вручную». В дальнейшем планируется осуществить интеграцию модуля разрешения кореференции непосредственно в технологию анализа текста. Это позволило бы

иметь доступ к структуре предложений и абзацев, а также к конкретным лексическим единицам, участвующим в сборках ИО. Анализируя объекты на стадии сборки, можно повысить полноту за счет расширения списка рассматриваемых случаев ее применения. Однако, в общем виде подобное вряд ли возможно. Тем не менее, подобную интеграцию теоретически реально осуществить для конкретной системы АОТ.

Литература

[1] Caroline V. Gasperin Statistical anaphora resolution in biomedical texts. Technical report, University of Cambridge Computer Laboratory. 2009. ISSN 1476–2986

[2] Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task. In Proceedings of the CoNLL-2011 Shared Task.

[3] Mitkov, R. Anaphora resolution: the state of the art, Working paper, (Based on the COLING'98/ACL'98 tutorial on anaphora resolution), University of Wolverhampton, Wolverhampton, 1999.

[4] Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. 2010. A Multi-Pass Sieve for Coreference Resolution. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010).

[5] Главная страница проекта Freebase. [Электронный ресурс] – Режим доступа http://www.freebase.com/(дата последнего обращения: 16.08.2012)

[6] Ермаков А.Е. Референция обозначений персон и организаций в русскоязычных текстах СМИ: эмпирические закономерности для компьютерного анализа. // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2005». М.: Наука, 2005.

[7] Загорулько Ю.А., Боровикова О.И., Кононенко И.С., Сидорова Е.А. Подход к построению предметной онтологии для портала знаний по компьютерной лингвистике. // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2006». М.: РГГУ, 2006.

[8] Официальный сайт Принстонского университета. [Электронный ресурс]. Главная страница проекта WordNet. – Режим доступа http://wordnet.princeton.edu (дата последнего обращения 16.08.2012).

[9] Поцепня В.Н. Разрешение местоименной анафоры в многоязычных информационных системах. // Искусственный интеллект-2006 №4 С.619-626.

[10] Серый А.С., Сидорова Е.А. Идентификация объектов в задаче автоматической обработки документов. // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2011». М.: РГГУ, 2011. С. 580-591.

[11] Толпегин П.В., Ветров Д.П., Кропотов Д.А. Алгоритм автоматизированного разрешения анафоры местоимений третьего лица на основе методов машинного обучения. Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2006» // Под ред. Н.И. Лауфер, А.С. Нариньяни, В.П. Селегея. – М.:РГГУ, 2006

Searching referential relationships between the information objects during the automatic document processing

Alexey Seryj, Elena Sidorova

This article describes a way to establish a referential identity (or coreference) of the information objects extracted from natural-language documents. Information objects are taken as hypothesis about the real object that lies in a given subject area and put together with vocabulary units found in the text during the analysis. The proposed approach allows to abstract from the text processing technologies. There are several certain requirements imposed only on the information objects format and they are specified by the ontology description. Coreference establishing process consists of three stages: similarity degree calculation and analysis; construction of the set of hypothetical equivalents for each object; coreferential objects unification. We introduced a new quantity called similarity index that is used for estimating the objects similarity degree. Referentially identical objects are merged into one.

[1]Кореферентность (референциональное тождество) – отношение между компонентами высказывания (обычно именными группами), которые обозначают один и тот же внеязыковой объект или ситуацию, т.е. имеют один и тот же референт.