ISSN 2079-3316 ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ № 1(15), 2013, с. 15–30

УДК 004.652.8:004.056.2:515.1

Е. А. Сулейманова, И. В. Трофимов

О подходе к отождествлению сущностей в рамках задачи извлечения информации из текстов

Аннотация. В статье описан метод отождествления текстовых упоминаний лиц, опирающийся на предметные знания в форме онтологии. В основе лежит понятие референциальной совместимости, определяемой как концептуальная совместимость при условии непротиворечивости индивидных признаков. Приводятся результаты экспериментальных исследований метода.

Ключевые слова и фразы: извлечение информации, разрешение кореферентности, формализованные предметные знания.

Введение

Методы извлечения информации из текстов, основанные на правилах [1], типично оперируют локальным контекстом. Правила выявления целевой информации последовательно применяются к небольшим текстовым фрагментам — обычно область поиска ограничена предложением текста. Сведения о той или иной сущности, извлекаемые из отдельно взятого предложения связного текста, часто имеют фрагментарный характер.

Чтобы преодолеть этот недостаток подхода на базе правил и получить целостные представления об упоминающихся в тексте объектах и фактах, необходимы механизмы, которые позволят объединить разрозненные фрагменты структурированных, но неполных их описаний. В статье предложен подход к отождествлению извлеченных сущностей, опирающийся на концептуальные и фактические знания.

Работа выполнена при поддержке РФФИ (проект **12-07-31101** «Исследование и разработка методов восстановления целостной информации о сущностях на основе предметных знаний в задаче извлечения фактов из текста».

Условия референциальной совместимости сущностей (т.е. возможности того, что они соотносятся с одним и тем же внетекстовым объектом) описываются на специальном формальном языке. Эти условия определяются в терминах атрибутивной и релятивной информации, ассоциированной как с самими извлеченными сущностями, так и с концептами и/или экземплярами, представляющими их в ресурсе знаний. Такой поход не претендует на универсальность, но позволяет получать приемлемые результаты для отдельных категорий сущностей в ограниченных задачах.

1. Постановка задачи

Прежде чем описывать подход, оговорим ограничения, в рамках которых мы проводили его экспериментальную оценку.

Не будет преувеличением утверждать, что успешное разрешение кореферентности (референциального тождества) текстовых единиц представляет собой задачу уровня понимания текста. В полной мере решение такой задачи требует, как нам кажется, моделирования интегрального контекста, основанного на:

(1) знании законов построения связного текста (уровень дискурса);
(2) знаниях об устройстве мира (концептуальный, или онтологический, уровень);
(3) энциклопедических знаниях (уровень фактов, или знаний о конкретных экземплярах);
(4) знаниях, выводимых из прагматического контекста (уровень метаданных текста).

На данном этапе исследований мы поставили перед собой задачу оценить одну из составляющих такого интегрального метода разрешения кореферентности — а именно ту его часть, которая апеллирует к знаниям (онтологическим и фактическим).

Задача принципиально решалась не как разметка кореферентных текстовых фрагментов, а как установление отношения референциального тождества на множестве элементов извлеченной и структурированной информации. Сознательный отрыв от текста и игнорирование структуры дискурса ни в коей мере не провозглашаются нами как достоинства подхода. Более того, нам представляется, что ведущую, управляющую роль в алгоритме разрешения кореферентности

должен играть именно дискурсивный компонент. Это, вероятно, следующая задача, которую необходимо решать в рамках интегрального подхода. Естественные ограничения, вытекающие из постановки задачи на данном этапе, — это исключение из рассмотрения местоименной анафоры и невозможность учитывать такие дискурсивные факторы, как порядок появления упоминаний в тексте, расстояние (линейное или какое-либо другое).

Отождествление и его оценка осуществлялись только в рамках документа. Хотя подход в целом позволяет описывать условия референциальной совместимости без привязки к тексту и потому может использоваться для кроссдокументального отождествления, ограниченность рамками документа позволяет описывать более слабые условия референциальной совместимости, которые обычно выполняются в рамках текста, но не обязаны выполняться в коллекции. Возможности подхода к кроссдокументальному отождествлению являются предметом отдельного исследования и здесь не рассматриваются.

В своих экспериментах мы ограничились оценкой разрешения кореферентности для упоминаний объектов только одного класса — лиц. Объекты этой категории выступают в качестве целевой информации во многих прикладных задачах. Для лиц характерно большое разнообразие способов номинации в тексте (включая открытое множество личных имен). Наибольшую сложность представляет разрешение референции так называемых актуальных имен (ситуационно-ролевых именований лиц), поскольку при выборе номинаций такого рода автор текста не только оперирует статическими знаниями о мире, но и, актуализируя в повествовании разные сценарии, соответственно этому динамически наделяет референта той или иной актуальной ролью. Еще одной проблемой является алгоритмическое различение наличия или отсутствия у именной группы со значением аспекта лица референции к конкретному лицу. Ср., например:

(1) ***Король Испании*** *устроил прием в королевском дворце.*
(2) *Хорошо быть* ***королем Испании****.*

Отметим, что жанровый состав документов экспериментальной коллекции ограничен новостными сообщениями.

2. Описание подхода

2.1. Модель знаний

Устройство ресурса знаний, используемого системой извлечения знаний, описано, в частности, в [2].

Здесь, в связи с рассматриваемой проблемой, хотелось бы остановиться на одном из содержательных аспектов концептуализации. Значительная часть лексики, используемой для референции к лицам, имеет двойственную природу: с одной стороны, она обозначает нечто абстрактное — должность, род деятельности, профессию, звание и т.п., а с другой — человека, характеризуемого этой абстрактной сущностью. Эта полисемия настолько регулярна, что последовательная реализация в онтологии принципа «одно значение — один концепт» привела бы к удваиванию численности всех концептов такого типа и усугубила бы проблему неоднозначности анализа на низших уровнях.

Поэтому в нашей модели принято решение концептуализировать такие понятия односторонне — как концепты категории *аспект лица*. Категория *лицо* в нашей онтологии не представлена никакими словарными концептами (т.е. концептами, имеющими лексические соответствия в словаре). А вот экземпляры категории *лицо* могут иметь словарные соответствия. На этапе извлечения информации по личным именам строятся текстовые экземпляры категории *лицо*, а по всем прочим упоминаниям лиц (исключая местоимения) — текстовые экземпляры категории *аспект лица*, независимо от их референциальной природы. В результате отождествления каждый аспект, имеющий референцию к лицу, должен быть связан со своим референтом.

2.2. Описание алгоритма отождествления

Уровень представления входного текста, на котором выполняется отождествление, — это семантический уровень, полученный в результате работы подсистемы извлечения [2]. Единицы этого уровня мы называем текстовыми (промежуточными) экземплярами концептов и отношений. Текстовый экземпляр (ТЭ) концепта — это семантическое представление выделенной из текста именной группы, принадлежащей к одной из предопределенного множества категорий: лицо, аспект лица, организация, геополитическая единица (ГПЕ). Между текстовыми экземплярами концептов могут быть установлены экземпляры предметных отношений.

В общем виде задача отождествления упоминаний лиц на данном этапе формулируется следующим образом:

(1) построить (выделить) множество лиц, упоминаемых в тексте, — множество референтов;
(2) каждое упоминание некоторого референта в тексте связать с этим референтом.

Установление кореферентности номинаций лиц может опираться на отождествление текстовых упоминаний объектов других категорий — организаций и ГПЕ.

Для референции к лицу в тексте могут использоваться: (1) имя собственное, (2) личное местоимение, (3) именная группа (дескрипция), (4) дескрипция в сочетании с именем собственным в синтаксической роли приложения. В нашей модели номинациям первого типа соответствуют ТЭ категории *лицо*, номинациям третьего типа (назовем их автономными дескрипциями) — ТЭ категории *аспект лица*, а по четвертым строятся экземпляры отношений *$роль–лицо* между ТЭ — аспектом лица и ТЭ-лицом. Дескрипции четвертого типа будем называть собственными дескрипциями лица. Установление референтов местоимений (второй тип номинаций), как уже говорилось, выходит за рамки задачи данного этапа (хотя в тестовой коллекции случаи местоименной анафоры мы размечаем).

Кроме того, заметим, что из рассмотрения (и разметки) исключаются именные группы с семантикой лица и аспекта лица в следующих употреблениях:

(1) имена собственные в автонимном употреблении, т.е. обозначающие сами себя (*его сына зовут* ***Иван***);
(2) различные случаи, в которых ИГ с семантикой аспекта лица не имеет своим референтом конкретное лицо, например: *X назначен* ***губернатором*** *края Y; X,* ***генеральный директор*** *компании Y; выборы* ***главы*** *региона* — референция к должности; ***руководитель*** *такого уровня должен...* — родовая референция.

Поскольку собственная дескрипция уже связана с ТЭ-лицом, являющимся ее референтом, то конечную задачу — установление кореферентности упоминаний лиц — можно переформулировать как **установление референтов автономных дескрипций**. Собственную дескрипцию референта при этом естественно было бы считать

опорной. Однако следует сказать, что результаты, получаемые подсистемой извлечения, часто бывают неоднозначны. Нередки и случаи неоднозначного или просто неверного определения носителя собственной дескрипции (например, *адвоката предпринимателя NN* — в результатах NN будет и адвокатом, и предпринимателем). Алгоритму отождествления приходится по возможности исправлять и подобные ошибки извлечения.

В основе алгоритма отождествления лежит идея попарного сопоставления ТЭ и проверки их **референциальной совместимости** (т.е. возможности их отождествления с одним и тем же референтом). В большинстве случаев мы считаем необходимым и достаточным условием референциальной совместимости двух ТЭ **семантическую совместимость** их родительских концептов (концептуальную совместимость) при **непротиворечивости индивидных признаков** самих ТЭ. ТЭ, успешно прошедшие проверку на референциальную совместимость, связываются с одной и той же **референциальной вершиной**. Для определения референтов дескрипций могут использоваться и специальные, более сложные правила.

2.3. Концептуальная совместимость

Считаем, что текстовые экземпляры ТЭ-1 и ТЭ-2 концептуально совместимы, если выполнено какое-либо из следующих условий:

(1) **Совпадение.** Родительские концепты экземпляров ТЭ-1 и ТЭ-2 совпадают (*мать — мать, рядовой — рядовой*).

(2) **Концептуальная эквивалентность.** Родительский концепт экземпляра ТЭ-1 может быть получен из родительского концепта экземпляра ТЭ-2 применением правила установления концептуальной эквивалентности: *премьер-министр — глава правительства, руководитель правительства, председатель правительства, глава кабинета министров* и т.п.

(3) **Род-вид.** Между родительскими концептами экземпляров ТЭ-1 и ТЭ-2 в онтологии существует родо-видовая связь (отношение $UP): *министерство — ведомство; Приморский край — регион; погранзастава — застава.*

(4) **Колокализация.** Между родительскими концептами экземпляров ТЭ-1 и ТЭ-2 в онтологии имеется отношение *$колокализация* (сходное по значению с языковой синонимией, только на уровне концептов): *рядовой — солдат; адвокат — защитник-юридич.*

(5) **Род-вид (специальное)**. Между родительскими концептами экземпляров ТЭ-1 и ТЭ-2 в онтологии существует определенная комбинация специальных отношений. Например, родительские концепты имеют общего родителя в иерархии, но при этом характеризуют разные аспекты этого родителя — *рядовой, пограничник*, общий родитель — *военнослужащий*.

(6) **Ассоциация.** Родительские концепты ТЭ-1 и ТЭ-2 связаны ассоциативной связью через посредника — родительский концепт экземпляра ТЭ-3, связанного с ТЭ-1 или ТЭ-2 в тексте предметным отношением. Например, *начальник погранзаставы — пограничник*.

2.4. Индивидные признаки

Типы индивидных признаков для ТЭ различаются в зависимости от семантической категории.

Выделяются категории, в которых экземпляры концептов индивидуализируются посредством собственных имен — это лица, ГПЕ, некоторые виды организаций. Для таких категорий имя собственное и считается индивидным признаком. Для удобства такие категории назовем «именуемыми».

Экземпляры концептов таких категорий, как аспекты лица (звания, должности, имена по роду деятельности, роли-отношения и т.п.), органы управления и власти, организации при других организациях, не имеют (явных) собственных имен и индивидуализируются через отношения с другими экземплярами и значения особых строковых атрибутов. Например:

- *начальник погранзаставы «N»* — экземпляр должности индивидуализируется через экземпляр организации, подчиненный ему по отношению *$должность/роль_по_отн_к*;
- *министр транспорта и дорожного хозяйства Республики Татарстан* — экземпляр должности индивидуализируется через значение строкового атрибута «Ограничение» (*транспорта и дорожного хозяйства*) и экземпляр ГПЕ, подчиненный должности по отношению *$должность_в_ГПЕ*;
- *правительство Российской Федерации* — экземпляр организации индивидуализируется через экземпляр ГПЕ, подчиненный ему по отношению *$адм-тер_принадлежность*;

- *Комиссия ООН по миростроительству* — экземпляр организации индивидуализируется через экземпляр другой организации по отношению *$орг_в/при_организации* и через значение строкового атрибута «Ограничение» (*по миростроительству*).

Условия непротиворечивости индивидных признаков для двух концептуально совместимых ТЭ также различаются в зависимости от категории. Общий принцип установления непротиворечивости состоит в том, что если значения некоторого индивидного признака явно присутствуют у обоих ТЭ (по результатам анализа текста), то эти значения должны быть одинаковы. Если речь идет о строковых значениях, то они должны совпадать. Что касается признаков–отношений, то для непротиворечивости достаточно, чтобы текстовые экземпляры, «индивидуализирующие» по некоторому отношению каждый из двух рассматриваемых ТЭ, были к этому моменту уже объединены под одной референциальной вершиной (например, индивидные признаки-отношения для ТЭ *президент* (*республики*) и *глава* (*Татарстана*) будут непротиворечивы, если ТЭ *республика* и *Татарстан* будут отнесены к одной референциальной вершине). Отсутствие значения признака у одного ТЭ при наличии его у другого, равно как и отсутствие значений признака у обоих ТЭ, удовлетворяет условию непротиворечивости.

2.5. Схема алгоритма установления референциального тождества ТЭ-лиц

Отбор и маркировка «хороших» собственных дескрипций при ФИО. ТЭ — аспект лица, связанный с ТЭ-лицом (построенным по ФИО) отношением *$роль-лицо*, помечаем как «хорошую» дескрипцию этого ТЭ-лица, если в этот ТЭ-лицо не входит другое отношение *$роль–лицо* и если между ТЭ-аспектом и ТЭ-лицом нет отношения $должность/роль_по_отн_к. Это позволит отделить бесспорные дескрипции от таких, которые могут оказаться ошибочными (как результат неоднозначной интерпретации конструкции на этапе извлечения), что очень важно для последующих шагов, поскольку собственные дескрипции референта служат алгоритму основанием для распознавания других дескрипций этого референта в тексте. В уже приведенном примере *адвоката предпринимателя NN* ни одна из дескрипций референта NN не будет признана «хорошей», т.к. не выполнено первое условие. Во фрагменте *адвоката NN* дескрипция не считается «хорошей» в силу неоднозначной интерпретации связи

между *адвокат* и *NN* в результатах извлечения (*NN* — имя самого адвоката или имя подзащитного безымянного адвоката).

Отождествление ТЭ-лиц (по ФИО). В результате работы подсистемы извлечения (т.е. до отождествления) к категории лица могут быть отнесены исключительно ТЭ, построенные по личным именам собственным. ТЭ-лица, чьи имена успешно прошли проверку специальной процедурой сопоставления, объединяются в одну референциальную вершину. Построенная в результате вершина в дальнейшем выполняет роль модели референта, с которой связываются прочие упоминания референта в тексте (дескрипции).

Отождествление ТЭ «именуемых» категорий — организаций и ГПЕ. Правила отождествления выполняют проверку концептуальной совместимости ТЭ-организаций и ТЭ-ГПЕ и непротиворечивости индивидных признаков. ТЭ, успешно прошедшие проверку, связываются с одной референциальной вершиной.

Отождествление ТЭ — «неименуемых» организаций. Правила отождествления выполняют проверку концептуальной совместимости ТЭ-организаций и непротиворечивости индивидных признаков (опираясь при необходимости на полученные к этому моменту результаты отождествления). ТЭ, успешно прошедшие проверку, связываются с одной референциальной вершиной.

Отбор и маркировка «хороших» дескрипций. Правила проверяют собственные дескрипции одного референта на референциальную совместимость. Непомеченные дескрипции, референциально совместимые с помеченными, тоже помечаются как «хорошие».

Удаление непомеченных дескрипций. При наличии у референта как помеченных, так и непомеченных дескрипций, удаляются экземпляры отношений *$роль–лицо*, идущие к ТЭ-лицу от непомеченных дескрипций. Т.е. сами ТЭ-дескрипции физически не удаляются, а только лишь перестают считаться дескрипциями данного референта.

Перенос экземпляров предметных отношений. Экземпляры предметных отношений, входящих в дескрипцию референта, переносятся на сам референт.

Определение референтов автономных дескрипций. Эту задачу выполняют очень разные по сложности и рискованности правила. Риск связан главным образом с отсутствием у алгоритма возможности оперировать данными о структуре связного текста, о чем уже упоминалось.

Самые простые правила находят автономные дескрипции, референциально совместимые с дескрипцией некоторого референта (имеющейся у него к моменту применения правила), и связывают их с этим референтом. Например, при наличии у референта NN дескрипции *министр обороны РФ* дескрипция *глава Минобороны* будет приписана этому же референту.

Приведем примеры другого рода — использование отношения конверсии между концептами в онтологии. Фрагмент текста: ***Мать рядового NN***, *обвиняемого в убийстве 14 пограничников и егеря, после свидания с* ***сыном*** *заметила следы побоев.* К моменту применения этого правила результаты имеют такой вид, что ТЭ *мать* связан с референтом NN отношением *$должность/роль по отн к.* На основании того, что в онтологии концепты *мать* и *сын* связаны отношением $CONVERSIVE, правило связывает ТЭ *сын* с референтом NN. Аналогично, в примере *адвокат X–а... его подзащитному...* дескрипция *подзащитный* будет отнесена к референту X.

2.6. Язык правил отождествления и его интерпретатор

В основе разработанного нами подхода к отождествлению извлеченной информации лежит декларативный язык, описывающий условия референциальной совместимости сущностей в терминах элементов модели знаний, таких как типы сущностей в той или иной онтологической таксономии, значения атрибутов сущностей, отношения, в которые вступают сущности с другими сущностями. Условия референциальной совместимости удобно записывать отдельно для различных типов сущностей, из чего вытекает их (условий) организация в виде множества правил. Несмотря на то, что мы в своем исследовании ограничились оценкой подхода к отождествлению только для одного типа сущностей — лиц, правила разрабатывались для более обширного круга объектов. Например, для того, чтобы установить, что два упоминания *губернатора* в примере ниже (Таблица 1) отсылают к одному референту, необходимо прежде установить тождественность упоминаний регионов, возглавляемых губернатором.

Отождествление осуществляется попарно, поэтому каждое правило описывает условия референциальной совместимости для пары сущностей, извлеченных из двух различных текстовых фрагментов.

Таблица 1. Пример, поясняющий необходимость разработки правил отождествления для широкого круга сущностей

Текстовые упоминания	*губернатор Приморского края*	*губернатор края*
Представление в модели	губернатор → возглавляет → Приморский край	губернатор → возглавляет → край

Язык ориентирован на отождествление именно сущностей (экземпляров концептов), а не отношений. Условие референциальной совместимости описывается как конъюнкция атомарных проверок из следующего перечня (перечислены основные):

- проверка таксономического подчинения заданного объекта указанному концепту (непосредственно или опосредованно; по выбранному типу таксономических связей);
- для двух данных объектов проверка наличия общего родителя в заданной таксономии;
- проверка наличия предметного отношения заданного типа между указанными объектами;
- проверка, является ли указанный объект участником указанного отношения (с указанием роли в отношении);
- проверка того, что хоть какие-нибудь два таксономических родителя двух данных объектов связаны указанным отношением ассоциативного типа (возможно, опосредованно, с ограничением радиуса поиска);
- проверка на равенство/неравенство значений заданных атрибутов у пары объектов (а также сравнение с константой);
- проверка того, что заданный объект может характеризоваться указанным атрибутом;
- проверка того, что заданный объект характеризуется указанным атрибутом;
- проверка, является ли заданный объект априорным (создан инженером по знаниям и существовал в фактографической базе до извлечения информации);
- проверка принадлежности заданного объекта указанному представлению. Представления — это структурные элементы модели

знаний, позволяющие создавать группы элементов знаний по произвольному принципу. Например, могут быть созданы представления для группировки элементов знаний, относящихся к одной тематике, событию или процессу. В этом случае проверка принадлежности представлению эквивалентна проверке тематической принадлежности, отнесенности к событию или процессу, соответственно.

Синтаксически атомарная проверка записывается как предикат, аргументами которого являются переменные и константы. Из множества переменных, задействованных в условии референциальной совместимости, выделяются две, для которых это условие определяется. Остальные переменные играют вспомогательную роль для выражения отношений с другими объектами, фигурирующими в условии. Значениями переменных могут выступать экземпляры концептов (как извлеченные из текста, так и априорные).

Интерпретация условия референциальной совместимости состоит в переборе возможных значений переменных и выявлении таких наборов, для которых выполняются все атомарные проверки. Очевидно, что полный перебор возможных вариантов для означивания переменных при достаточно большом количестве экземпляров концептов и переменных вычислительно неэффективен (декартово произведение множеств возможных значений для каждой переменной). Для повышения эффективности были разработаны алгоритмы предобработки правил, целью которых было сужение областей значений переменных. Некоторые атомарные проверки — например, проверку таксономического подчинения — можно выполнить еще до перебора, сократив область значений переменной, фигурирующей в «таксономической проверке» (обычно сужение тем сильнее, чем ниже в таксономической иерархии находится концепт). В то же время для ряда атомарных условий предобработка сама является вычислительно неэффективной; в этом случае мы отказывались от предобработки в пользу перебора.

После успешного означивания всех переменных, входящих в условие референциальной совместимости, в выделенных переменных находится пара сущностей, для которых данным условием декларируется тождественность. Правила позволяют выразить эту тождественность либо при помощи установления между ними отношения тождественности (экземпляр отношения), либо выполнить это отождествление физически (преобразовать два экземпляра концепта в один).

В последнем случае действует эвристическая процедура, разрешающая конфликты между значениями атрибутов. Приоритет отдается наиболее специализированным значениям атрибутов, а в случае эквивалентности по этому показателю приоритет получают атрибуты экземпляра, производного от более частного концепта.

Организационно правила, описывающие условия референциальной совместимости, группируются в фазы. Порядок применения правил одной фазы недетерминирован, поэтому правила не могут учитывать эффекты отождествления других правил своей фазы. Но правила последующих фаз могут учитывать отношения тождественности и эффекты физического отождествления, выполненные на предшествующих фазах.

2.7. Методы и результаты оценки

Существуют различные подходы к количественной оценке эффективности алгоритмов разрешения кореферентности. Среди них наибольшую известность получили F-мера MUC-6 [3], бикубическая мера [4], ACE-value [5] и CEAF [6]. В своих исследованиях мы остановились на MUC-6 и CEAF-мерах.

F-мера, рассчитываемая по методике, предложенной на MUC-6, опирается на понятие компоненты связности. Эталонная разметка и результаты отождествления рассматриваются как графы, вершинами которых являются упоминания сущностей в тексте, а ребра объединяют множества упоминаний одной сущности в компоненту связности. Для расчета полноты и точности сначала подсчитываются ошибки. Для полноты вычисляется минимальное количество ребер, которые необходимо добавить в результирующую разметку, чтобы все вершины из каждой эталонной компоненты связности оказались связанными и в результирующей разметке. Для подсчета ошибок точности эталонная и результирующая разметка меняются местами. Затем, на основании известного минимально необходимого количества ребер (для образования компоненты связности) и числа недостающих ребер, вычисляются значения полноты и точности.

Расчет CEAF-меры опирается на поиск оптимального отображения один-к-одному между эталонными сущностями и сущностями, полученными в результате отождествления (причем каждая сущность представляет собой множество ее упоминаний). В графовой нотации это можно представить следующим образом. Определяется полный двудольный граф, вершинами которого выступают сущности

эталонной и результирующей разметки (соответственно попадающие в разные доли), а взвешенные дуги отражают степень сходства для каждой пары сущностей. Степень сходства может вычисляться как общее число упоминаний у двух сущностей (mention-based CEAF) или как «локальная» F-мера, отражающая степень сходства двух сущностей (entity-based CEAF). Задача состоит в поиске паросочетания с максимальным суммарным весом и решается венгерским алгоритмом (известным также как алгоритм Куна — Манкреса). После того как оптимальное соответствие и соответствующая суммарная степень сходства найдены, несложно вычислить полноту и точность. Для этого найденная степень сходства делится на степень сходства множества эталонных сущностей с собой (степень сходства, когда все эталоны найдены) или на степень сходства множества результирующих сущностей с собой (степень сходства, когда все результаты корректны), соответственно.

CEAF-мера дает более интуитивную оценку качества (легко интерпретируема) и сбалансированно дифференцирует результаты (одинаково штрафует за ошибки в точности и полноте). В отличие от нее MUC-мера имеет перекос в пользу результатов с избыточной связностью — высокой полнотой, но низкой точностью — в случаях, когда в тексте небольшое количество сущностей с большим числом упоминаний для каждой. В то же время много оценок приводится именно по MUC-мере, появившейся раньше CEAF, поэтому для возможности сравнения с результатами других подходов мы выполняем расчет качественных показателей и по ней тоже.

Для экспериментальных исследований метода была размечена небольшая тестовая коллекция, состоящая из 100 текстов новостей. Получены следующие результаты:

Методика оценки	F-мера	Точность	Полнота
MUC-6	60,32	82,44	47,56
mention-based CEAF	68,78	84,04	58,20
entity-based CEAF	72,08	72,77	71,40

Выводы

Исследован метод разрешения кореферентности текстовых упоминаний лиц, основанный на проверке условий референциальной совместимости построенных по ним текстовых экземпляров. Референциальная совместимость текстовых экземлпяров определяется как концептуальная совместимость их родительских концептов при непротиворечивости индивидных признаков экземпляров. Как следует из результатов экспериментов, предложенный метод позволяет решать задачу установления кореферентности упоминаний лиц на приемлемом качественном уровне. Нам представляется, что дополнение алгоритма дискурсивным компонентом позволит существенно повысить как точность, так и полноту результатов.

Список литературы

[1] Sarawagi S. *Information Extraction* // Foundations and Trends in Databases, 2008. **Vol. 1**, no. 3, p. 261–377. ↑[]

[2] Кормалев Д., Куршев Е., Сулейманова Е., Трофимов И. *Технология извлечения информации из текстов, основанная на знаниях* // Программные продукты и системы, 2009, № 2, с. 62–66. ↑2.1, 2.2

[3] Vilain M., Burger J., Aberdeen J., Connolly D., Hirschman L. *A Model-Theoretic Coreference Scoring Scheme* // Proceedings of the Sixth Message Understanding Conference (MUC-6), 1995, p. 45–52. ↑2.7

[4] Bagga A., Baldwin B. *Algorithms for Scoring Coreference Chains* // Proceedings of the Linguistic Coreference Workshop at The First International Conference on Language Resources and Evaluation (LREC'98), 1998, p. 563–566. ↑2.7

[5] The ACE 2003 Evaluation Plan : NIST, 20 August 2003, ftp://jaguar.ncsl.nist.gov/ace/doc/ace_evalplan-2003.v1.pdf. ↑2.7

[6] Luo X. *On coreference resolution performance metrics* // Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT'05). — Stroudsburg, PA, USA : Association for Computational Linguistics, 2005, p. 25–32. ↑2.7

Об авторах:

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

Игорь Владимирович Трофимов

Старший научный сотрудник Исследовательского центра искусственного интеллекта ИПС им. А. К. Айламазяна РАН, специалист по технологиям автоматической обработки естественного языка, извлечения информации, автоматического планирования.

e-mail: igor@warlock-98.botik.ru

Образец ссылки на эту публикацию:

Е. А. Сулейманова, И. В. Трофимов. *О подходе к отождествлению сущностей в рамках задачи извлечения информации из текстов* // Программные системы: теория и приложения : электрон. научн. журн. 2013. Т. 4, № 1(15), с. 15–30.

URL: http://psta.psiras.ru/read/psta2013_1_15-30.pdf

E. Suleymanova, I. Trofimov. *A method for coreference resolution within information extraction.*

Abstract. The paper presents an ontology-based method for establishing coreference relationships between person-type mentions. The method is centered around the concept of referential consistency of markables, which is defined as both semantic (conceptual) consistency and consistency of individual features. Some experimental results are given. (in Russian).

Key Words and Phrases: information extraction, coreference resolution, structured domain knowledge.