Named entity and fact extraction by RCO in Dialogue Evaluation 2016

В докладе рассказывается об опыте участия компании “ЭР СИ О” в подготовке и проведении соревнования Dialogue Evaluation 2016 между системами компьютерного анализа текста в части выделения именованных сущностей и фактов. Анализируется методика проведения этого соревнования и результаты разметки тестового корпуса экспертами, а также различные способы сравнения результатов работы систем с экспертной разметкой и соответствующие оценки качества.

Named entity and fact extraction by RCO in Dialogue Evaluation 2016

A.E. Ermakov, P.Y. Polyakov

The paper describes the experience of the participation of RCO LLC in the preparation and running of the Dialogue Evaluation 2016 competition between automatic text analysis systems in terms of named entity and fact extraction. The methodology for running this competition and the results of experts’ markings of the test corpus are being analyzed, as well as various methods of comparing the results performed by the automatic systems with the experts’ markings and corresponding quality assessments.

Опыт участия RCO в Dialogue Evaluation 2016: выделение именованных сущностей и фактов

Введение

В рамках мероприятия по тестированию и сравнению анализаторов текста Dialogue Evaluation 2016 авторы от лица компании “ЭР СИ О” готовились принять участие в разделе “Выделение именованных сущностей и фактов”, а именно:

- в Дорожке 1 по выделению в тексте упоминаний именованных сущностей: персон, организации и локаций;
- в Дорожке 2 по выделению атрибутов, нормализации и идентификации сущностей, выделенных в Дорожке 1;
- в Дорожке 3 по извлечению из текста упоминаний фактов заданных типов с их участниками в заданных ролях.

Участие авторов в Dialogue Evaluation 2016 как представителей компании “ЭР СИ О” (*http://www.rco.ru*) было мотивировано желанием:

а) получить объективную независимую оценку качества работы собственных лингвистических алгоритмов, воплощенных в программе RCO Fact Extractor (*http://www.rco.ru/?page_id=3554*), в том числе провести сравнение с качеством работы других алгоритмов, прежде всего, алгоритмов Организаторов – компании ABBYY.

б) обеспечить разработчиков лингвистических анализаторов корпусом размеченных русских текстов, пригодным для исследования алгоритмов в части обработки определенных классов именованных сущностей и фактов.

Будучи включенными в состав Оргкомитета Dialogue Evaluation 2016 по разделу “Выделение именованных сущностей и фактов”, авторы доклада принимали активное участие в формировании требований к эталонной разметке текстов асессорами, правил сравнения результатов работы тестируемых программ-анализаторов с результатами эталонной разметки, а также в определении общей методики оценки результатов. К сожалению, членам оргкомитета не удалось выработать консолидированную позицию по всем пунктам в постановочной части Дорожек 2 и 3. Кроме того, не был проработан и описан в деталях алгоритм сравнения результатов работы систем прогонов участников с эталонной разметкой, на основании которого

организаторами были написаны программы-компараторы для сравнения результатов машинной и человеческой разметки. По этой причине, а также вследствие того, что программы-компараторы не предоставляли достаточной информации для интерпретации результатов своей работы, алгоритм работы программ-компараторов, если не изучать исходный код программ, представляется «чёрным ящиком». Это, на наш взгляд, лишает возможности оценивать преимущества и недостатки различных подходов, используемых участниками дорожек для решения поставленных задач, путём сравнения результатов прогонов участников, и, тем самым, снижает научную ценность результатов. Тем не менее, тестирование и сравнение анализаторов текста в задаче «Выделение именованных сущностей и фактов» проводится на конференции «Диалог» впервые, и поэтому эта работа очень ценна в плане уточнения постановки задач и доработки алгоритмов программ-компараторов.

В соревновании разметка обучающей и тестовой выборок производилась одновременно, до получения результатов прогонов участников. Однако, более широкораспространённой является методика, когда случайно выбранное подмножество тестовой выборки, или подмножество, сформированное по принципу «общего котла», размечается после получения результатов прогонов участников. Такая методика позволяет получить более объективные оценки.

Очень близкое по направленности соревнование было проведено в 2005 году в рамках РОМИП: дорожка поиска фактов [1]. В этой дорожке было выделено две подзадачи: выделение именованных сущностей (аналог Дорожки 1 настоящего соревнования) и выделение фактов заданных типов (аналог Дорожки 3). Интересно, что в 2005 году организаторы также столкнулись с серьёзными методологическими сложностями, в частности, «работа асессоров затруднялась расплывчатыми ответами систем» [1], по этой причине не была завершена оценка полноты результатов, присланных участниками. Тем не менее, были оценены результаты по 3000 упоминаниям фактов (аналоги ownership и occupation в настоящем соревновании). Для сравнения, в рамках настоящего соревнования оценено всего 254 упоминания фактов в обучающей и 702 в тестовой коллекциях. В 2005 году у лучших систем точность выделения фактов с типами ownership и occupation была близка или превышала 0,9. В настоящем соревновании у лучшей системы точность выделения фактов с типом ownership составляет 0,52, с типом occupation – 0,78. По опыту нашей компании, если система выделяет факты с точностью меньше 0,9, то в плане практического использования она бесполезна – слишком много шума. Поэтому, если исключить вариант, что современные системы перестали решать практические задачи, то остаётся предположить недостатки в методологии вычисления точности в текущем соревновании.

В данной работе мы изложили наши замечания и предложения по каждой дорожке, направленные на повышение качества проведения тестирования. Мы также решили принять участие в соревновании по Дорожке 1 в части выделения упоминаний именованных персон, так как в отношении данной задачи у нас не имелось серьезных замечаний.

Извлечение сущностей

В Дорожке 1 от систем участников требовалось выделить в тексте все упоминания именованных сущностей, относимых к классам персон, организаций и локаций, причем, только такие упоминания, в которых сущность была обозначена именем собственным. Для выделения упоминания сущности необходимо было указать позицию этого упоминания в тексте: смещение от начала и длину упоминания. Ошибка в указании позиции сущности, когда выделена только часть ее упоминания, или в упоминание ошибочно добавлено слово, понижает и точность, и полноту итоговой оценки.

Алгоритм RCO выделения упоминаний именованных персон в тексте основан на словарях известных фамилий, имен и отчеств, а также на бессловарном морфологическом анализе слов, написанных с заглавной буквы, в с целью сформировать гипотезы о возможном словоизменении и

лексическом разряде слова, похожего на фамилию или отчество [2]. Для каждого предложения цепочка множественных морфологических описаний слов (омоформов) обрабатывается правилами, распознающими различные способы написания ФИО и подобных им наименований российских, азиатских, арабских, китайских и прочих персон. Правила проверяют порядок и тип элементов наименований персон, их грамматическое согласование, учитывают определенный контекст - приложения, обозначающие должность, титул, обращение, а также наличие омонимии с именами нарицательному. В случае срабатывания каждое правило формирует множество возможных вариантов написания соответствующей сущности в тексте, которые затем ищутся в тексте и могут добавить в число выделенных по правилам упоминаний сущностей новые варианты этих упоминаний – более краткие упоминания, которые сами по себе не могут быть достоверно классифицированы как обозначения именованных сущностей – например, при омонимии с именами нарицательными с заглавной буквы, или в заголовке, где все слова написаны прописными буквами. Таким образом, если по отдельному слову *Заяц* или *Владимир* нет возможности определить, обозначает ли оно человека, то при наличии рядом в тексте упоминания *Михаила Зайца* или *Владимира Петровича* можно с уверенностью принять решение. Другой типичный пример – выделение единого имени организации *РОГА И КОПЫТА* в заголовке типа *РОГА И КОПЫТА УХОДЯТ С РЫНКА* становится возможным при наличии в тексте полного упоминания *ООО “Рога и копыта”*. Так же разрешается омонимия между именами персон и объектов других типов – географическими наименованиями, названиями организаций, артефактов. В ходе этого процесса различные варианты написания наименования объекта в тексте отождествляются между собой – разрешается кореферентность, их атрибуты (для персон – элементы ФИО, для организаций – собственно название, организационно-правовая форма, тип организации, географические атрибуты и т.п.) идентифицируются и стандартизуются. Дополнительно в ходе отождествления разных наименований, употреблявшихся в тексте в разных грамматических формах, выбираются наиболее достоверные гипотезы о словоизменении их атрибутов из числа порожденных морфологическим анализатором. Например, по отдельному упоминанию в тексте *Дмитрия Танка* невозможно установить, склоняется ли его фамилия и какова будет ее нормальная форма, однако наличие в тексте упоминания *Дмитрий Танка* либо *Дмитрию Танку* позволяет уверенно решить проблему. На основе множества собранных и стандартизованных значений атрибутов для каждого объекта строится стандартизованное наименование по соответствующим правилам синтеза, для европейских персон это цепочка *Фамилия Имя Отчество* в именительном падеже. В случае сохранившейся неоднозначности разбора решение о выделении именованной сущности делегируется следующему этапу обработки текста - синтаксическому анализу, в котором принимается то решение, которое позволяет получить наилучший граф синтаксического разбора. По результатам синтаксического анализа “собираются” имена дополнительных объектов, например организаций вида *Московский государственный институт стали и сплавов* “. В целом, различные составляющие процесса выделения упоминаний именованных сущностей реализуются разными эвристическими алгоритмами на разных этапах лингвистического анализа текста, что делает затруднительным систематическое описание всех особенностей нашей системы в ограниченном объеме статьи .

На тестовой выборке наша система показала следующие результаты: точность 0.956, полнота 0.873, F1-мера 0.912. Сдвиг в сторону точности заложен в алгоритм специально, чтобы сохранять приемлемую точность в процессе многостадийного анализа, включающего также анафору, синтаксический анализ и выделение фактов. Было бы интересно сопоставить эти цифры с уровнем согласия между собой асессоров, но, к сожалению, разметка одних и тех же документов разными асессорами в данном тестировании не проводилась.

Деление именованных сущностей на организации и локации не является однозначным, что было признано де факто Организаторами – так, в правилах разметки текстов для людей и программ-анализаторов был выделен особый класс упоминаний сущностей в тексте –

“организации-локации”, к которому относились упоминания географических объектов в роли организации вида *Россия подписала договор с Ираном*, *Россия прекратила поставки газа Украине*. Объективных оснований для выделения подобных словоупотреблений в отдельный класс мы не обнаруживаем – с теоретической точки зрения, мы не можем дать четкого определения границ данного класса, с практической точки зрения мы не видим прикладных задач, в которых подобное подразделение сущностей оказалось бы полезно. Единственный критерий, на основании которого можно уверенно относить сущности к классу “организации-локации” в эталонной разметке – это употребление наименования географического объекта в роли субъекта определенных пропозиций – сочетаний определенных глаголов с определенными распространителями: *Москва включает в себя десять административных округов*, vs. *Москва включает механизмы политического урегулирования*, *Южная Африка является домом черной цапли*, vs. *Южная Африка является поставщиком алмазов*. Отметим, что в лингвистической литературе отмечаются как типичные подобные случаи употребления организаций в роли персон при глаголах, подразумевающих одушевленность своего субъекта: *институт проголосовал за/вышел на митинг* – здесь имя организации обозначает совокупность входящих в нее людей. Очевидно, что в этих же контекстах “организаций-персон” могут фигурировать и географические объекты, т.е. мы вроде бы как имеем класс “локация-организация-персона”, тогда как на самом деле мы, таким образом, просто пытаемся выделить роль “одушевленный субъект”. Нам кажется, что требование различать “объективное” значения слова - его денотата и значение конкретного словоупотребления тексте - “ролевого” значения слова при формулировании требований к выделяемым сущностям является избыточным для данной дорожки.

Если обосновывать требования к разметке практическими соображениями построения эффективных прикладных систем, в том числе ставить задачей соревнования выявление областей применения испытуемых систем, то нам очевидна необходимость как минимум следующих дополнительных требований:

А) Необходимо различать и выделять как отдельные классы локации особых видов, не являющие объектами административно-территориального деления (страны, районы, населенные пункты, улицы) и географического (моря, реки, горы). Просмотр сущностей, выделенных как локации в обучающих текстах, показал, что к таковым разметчиками отнесены леса (*Химкинский лес*), памятники (*Юрию Долгорукому*), магазины (*Библио-глобус*), планеты солнечной системы и даже искусственные космические объекты (*МКС*). Перечень типов таких объектов и правила их классификации с перечислением списков конкретных слов должны быть заранее определены и анонсированы участникам, а оценки должны проставляться для каждого класса сущностей. Распознавание же всех подобных именованных объектов в одной куче для практических задач бессмысленно.

Б) Не должны выделяться упоминания персон в составе географических объектов и организаций *стадион имени Ленина*, *МГУ им. М.В. Ломоносова*. При построении информационно-поисковых систем и автоматических классификаторов текста трактовка *Ленина* и *Ломоносова* как самостоятельных сущностей в подобном контексте приводит к информационному шуму.

В) Отдельно от локаций должны выделяться упоминания географических объектов в составе организаций и должностей вида *президент РФ*, *Голос Америки*. Выделение самостоятельных географических объектов в подобных синтаксических позициях обычно приводит к шуму в прикладных системах – содержание текста часто не имеет никакого отношения к такому косвенно упоминаемому объекту. Соответствующие неоднозначности мы встречали в ручной разметке, например, в составе выделенной разметчиками организации *Голос Америки* локация *Америка* также выделена как самостоятельная сущность, тогда как в составе организации *Движение общежитий Москвы* локация *Москва* не выделена. В соответствии же с наблюдаемыми по факту правилами разметки, даже прилагательные, образованные от географических объектов,

трактовались как локации общего вида наравне с прочими: *российская компания* – должна быть выделена локация *Россия*, *подмосковный губернатор* – должна быть выделена локация *Подмосковье*. Практически в Дорожках 1 и 2 по выделению сущностей требование выделения таких локаций было просто равнозначно требованию включить в систему словарь соответствий географических прилагательных объектам. Однако для Дорожки 3 по выделению фактов данная неоднозначность, на наш взгляд, вылилась в принципиальную проблему – требование того, что по упоминанию *подмосковный губернатор Борис Громов* должен быть выделен факт *Occupation { Who=Борис Громов, Position=губернатор, Where=Подмосковье }* логично приводило к тому, что для "*новокузнецкого хоккеиста Александра Орехова*" местом работы становился Новокузнецк, а для *российского президента/программиста* в качестве места работы выступала *Россия*. Даже в отношении государственных чиновников такая интерпретация представляется не совсем верной – их местами работы являются соответствующие госучреждения, а не географические объекты. Наименование же географического объекта должно просто входить в состав наименования должности, что не было учтено в методике разметки или сравнения результатов работы систем, вследствие чего семантически наиболее точное выделение факта нашей системой в виде *Occupation { Who=Борис Громов, Position=губернатор Подмосковья }* приводило к ошибкам и полноты (формально не выделено *Where=Подмосковье*), и точности (формально неверно выделена должность). Приведем еще ряд примеров спорной разметки, встретившихся нам в обучающей выборке размеченных текстов: *Интернет* как организация отмечен 9 раз, *Белый Дом* как организация – 1 раз и как локация – 3 раза, *Первая пилотируемая экспедиция* – организация...

Помимо собственно классификации именованных сущностей существует проблема, связанная с определением границ упоминаний сущностей. Так, в разметке словосочетания *немецкой компании Harles und Jentzsch* слово *немецкой* не включено разметчиками в состав объекта и потому его выделение анализатором текста будет оштрафовано, в разметке же словосочетаний *северовосточный китайский город Дацин*, *тихоокеанский порт Козьмино* в соответствии с разметкой требуется включение всех слов в состав объекта. Еще больше возникает произвола при разметке словосочетаний с однородными членами типа *компаний Oracle и Sun Microsystems*. Такая же проблема наблюдается с определением границ должностей для фактов типа Occupation. Мы предлагаем дополнить инструкцию для разметчиков просьбой указывать все возможные варианты названия организации, и при сравнении полученной эталонной разметки с прогонами участников считать допустимым любой из указанных возможных вариантов. Кроме того, для каждого документа необходимо получить разметку от нескольких разметчиков, которые работают с документом независимо друг от друга, чтобы, как минимум, иметь сильную и слабую оценку сравнения прогонов с эталонной разметкой.

Дорожка 2 по выделению именованных сущностей в дополнение к Дорожке 1 предполагала выделение в тексте всех упоминаний сущности, определение атрибутов имени и составление наиболее полного варианта имени. Кроме того, значения атрибутов должны были нормализованы и дублирующиеся значения отброшены. Неотождествлённые между собой упоминания одной и той же сущности (например, *Большой театр* и *главный театр страны*) штрафовались. Также штрафовалось добавление к сущности атрибутов, не указанных в тексте – например, добавление отчества к фамилии президента, если отчество в тексте не употреблялось, что фактически означало запрет на использование нелингвистических источников знаний. При этом для нас остается вопросом, как тогда предполагалось сравнивать синонимичные варианты атрибутов – *Россия*, *РФ* и *Российская федерация*, *Петербург* и *Санкт-Петербург*, *Володя* и *Владимир* и им подобные.

Для получения хорошего результата на данной дорожке необходимо было решить две задачи: установить кореферентность между сущностями и собрать и классифицировать атрибуты сущности, нормализовав их и удалив дубликаты. По нашему мнению, каждая из этих двух задач

заслуживает отдельной дорожки. Их объединение в одну дорожку затрудняет постановку задачи, и это, возможно, стало причиной недостаточно детального описания дорожки и недоработок в программе компаратора (например, программа штрафует за замену *ё* на *е*).

Описание типа сущности/факта	**Код в эталонной разметке**	**Примеров в обучающем корпусе**	**Примеров в тестовом корпусе**
Сущности		*2470*	
Локации + Локации-организации	Loc, LocOrg	1056	
Организации	Organ	673	
Персоны	Person	741	
Факты (всего)		*264*	
Сделка (всего)	Deal	30	
сделка: займы	Type займ/возвращение займа	6	
сделка: инвестиции	Type инвестиции	1	
…	Type инвестиция	2	
сделка: купля/продажа	Type купля/продажа	11	
сделка: штрафы	Type наложение/выплата штрафа	7	
сделка: услуги	Type услуги	1	
сделка: участие в мероприятии	Type участие в мероприятии	2	
Встреча	Meeting	4	
Работа (всего)	Occupation	207	
работа: начало	Фаза конец	7	
работа: конец	Фаза начало	6	
Владение предприятием	Ownership	14	
Часть целого	isPartOf	9	

Таблица 1. Статистика типов эталонных сущностей и фактов в обучающем и тестовом множествах. Последний столбец должен быть заполнен после публикации Организаторами всего размеченного корпуса.

Извлечение фактов

Технология, используемая нами для извлечения из текста описаний заданных событий и фактов в структурированной форме фреймов с определением всех требуемых участников и их ролей подробно описано в статье [3]. Соответствующее решение воплощено в программном компоненте RCO Fact Extractor SDK (http://www.rco.ru/?page_id=3554), который позволяет разработчикам информационно-поисковых и аналитических систем включать возможности лингвистического и семантического анализа текста в собственные приложения. Стандартные лингвистические настройки RCO Fact Extractor SDK позволяют выделять более двадцати классов различных сущностей, в том числе пять классов именованных, и около двухсот классов событий и фактов (http://www.rco.ru/wp-content/uploads/2015/04/RCO_Object_Description.pdf). Данная система

(в версии 2005 года) участвовала в дорожке по извлечению фактов в рамках семинара РОМИП [4] (результаты можно найти в приложении http://romip.ru/romip2005/appendix_h.pdf). Однако, в настоящем мы не стали принимать участие в Дорожке по извлечению фактов по причинам, описанным ниже.

Из Таблицы 1 видно, что число примеров на все типы фактов в размеченной обучающей выборке оказалось ничтожно мало, за исключением фактов типа Occupation (работа), причём в абсолютном большинстве случаев данный тип факта выражается в тексте простой последовательностью сущностей должность → персона → организация. В результате, третья дорожка свелась к извлечению малоинтересного для прикладных задач факта, который, к тому же, не требует интеллектуальных алгоритмов анализа текста, и для получения высоких оценок достаточно научиться ловить подряд идущие сущности должность → персона → организация.

Второй по количеству примеров тип факта – Deal (сделка) – имеет слишком общее определение: индикация некоторого «экономического взаимодействия» между персонами и организациями. Где находится граница «экономического взаимодействия» не определено. Из-за этого возникает произвол в понимании и разметке фактов данного типа, вот типичный пример, иллюстрирующий подобный произвол: файл book_129.facts обучающей выборки содержит факт “сделка: наложение/выплата штрафа” с двумя участниками Participant=*США* и Participant=*Куба*, которому в файле book_129.txt соответствует текст “*Куба была отрезана от всемирной паутины кабельных телекоммуникаций в результате наложенного Вашингтоном эмбарго на торговлю с коммунистическим государством*“. На наш взгляд, наложений санкций никак не может считаться сделкой, роли Кубы и США совершенно различны, а требование выделения участника *США* вместо *Вашингтон* вообще не имеет никаких оснований. Ещё одна проблема в определении данного типа факта – участники факта не дифференцируются по их ролям в факте, например, для вроде понятного подкласса фактов “сделка: купля-продажа” все именованные участники – компания-покупатель, компания-продавец и компания-товар – не должны были различаться при выделении. Практический смысл решения задачи извлечения фактов в такой постановке теряется.

Ещё одна проблема связана со сложными правилами сравнения выделенных участников факта при оценке результата. Согласно заявленному регламенту оценки результатов, в зачет шло как совпадение строк исходных упоминаний именованных сущностей в тексте, выделенных по Дорожке 1, так и совпадение цепочек нормализованных атрибутов сущностей, выделенных по Дорожке 2 на основе собираемых с разных мест текста упоминаний этих атрибутов. При таком подходе в случае выдачи исходных упоминаний из текста оцениваемая система заведомо проигрывает там, где участники факта выражаются анафорически - местоимениями или именами нарицательными, а в случае выдачи нормализованных имен возникают проблемы, описанные выше в отношении Дорожки 2. Мы предлагаем разрешить участвующим в дорожках системам выдавать участников фактов как в виде позиций в тексте (для систем, которые могут определить позицию участника факта в тексте), так и в виде стандартизированных наименований в соответствии с Дорожкой 2 (для систем, которые хорошо справляются с Дорожкой 2). Чтобы увеличить процент интересных с практической точки зрения фактов в тестовой коллекции, мы предлагаем перейти от модели сплошного вычитывания новостных сообщений к модели поиска документов по ключевым словам, характерным для интересующего нас факта. При этом список ключевых слов, по которым был произведен отбор документов, должен быть опубликован как часть описания дорожки.

Заключение

В работе описан опыт участия компании RCO в дорожках Dialogue Evaluation 2016. В виду того, что соревнование по выделению именованных сущностей и фактов в текстах на русском языке проводится впервые, и постановка задач и методика оценки результатов прогонов

недостаточно проработаны, мы приняли решение участвовать только в Дорожке 1 в части выделения упоминаний именованных персон, так как в отношении данной задачи у нас не имелось серьезных замечаний. Наша система показала высокие результаты, однако, оценить, насколько далеки они от теоретического максимума – уровня согласия асессоров в разметке тестовой коллекции, нет возможности в виду отсутствия соответствующих данных на момент написания статьи. Существенная часть статьи посвящена описанию проблем, с которыми сталкиваются участники соревнования и нашим предложениям, как эти проблемы можно устранить в будущих дорожках.

Литература

Кураленок И., Некрестьянов И., Некрестьянова М. Обзор РОМИП'2005 // Труды третьего российского семинара по оценке методов информационного поиска. - Санкт-Петербург: НИИ Химии СПбГУ, 2005, С. 7-22.

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.

Ермаков А.Е., Плешко В.В. Семантическая интерпретация в системах компьютерного анализа текста // Информационные технологии. - 2009. – N 6. – С. 2-7.

Плешко В.В., Ермаков А.Е., Голенков В.П., Поляков П.Ю. RCO на РОМИП 2005 // Труды третьего российского семинара по оценке методов информационного поиска. - Санкт-Петербург: НИИ Химии СПбГУ, 2005, С. 23-39.