А. В. Молоткова

СРЕДСТВА ВЫРАЖЕНИЯ ИМЕНОВАННЫХ СУЩНОСТЕЙ И ФАКТОВ В ПИСЬМЕННОМ ТЕКСТЕ

Преобразование неструктурированной текстовой информации в хорошо организованные и структурированные данные и знания является одной из основных задач современных технологий *Data Mining, Text Mining, Information Extraction*. Указанные технологии положены в основу создания компьютерных систем, способных правильно извлекать из текстовых массивов различные объекты и факты. Например, даты каких-либо событий, имена, адреса и телефоны персон, наименования товаров и услуг, названия компаний, описания событий в новостных сводках, мнения, оставленные пользователями социальных сетей, гостевых книг или форумов, отзывы о разнообразных продуктах и услугах и т.д. Успешность функционирования подобных систем зависит как от самой извлекаемой информации, т.е. количества и типов извлекаемых объектов, фактов и связей между ними, так и от способа представления и средств обработки полученных данных и знаний,

что непосредственно определяет класс решаемых компьютером задач: идентификации объектов, выявления и анализа фактографической информации, семантического поиска, экспертных решений, ответа на запросы, выраженные на естественном языке и т.п.

Несмотря на разнообразие существующих методов извлечения информации из неструктурированных текстовых массивов и других типов данных, до сих пор не решены ключевые проблемы информационного поиска, связанные с автоматическим построением баз данных и знаний. Наименее разработанными являются такие области, как извлечение знаний из открытых информационных ресурсов, поиск локализованных во времени фактов, а также извлечение именованных сущностей, т.е. имен и фамилий людей, названий организаций, географических названий и т.п. Проблема извлечения локализованных во времени фактов состоит в автоматическом определении временных границ. Она имеет достаточно хорошие решения в случае точного указания дат, а также при наличии взаимоисключающих фактов. Однако в реальных текстах временные границы часто задаются нечетко, например, «в детстве», «в старости» и т.д. Предположение о взаимном исключении фактов даже в классической задаче *Marriage Problem* не всегда истинно (как, например, в странах, где узаконены полигамные отношения). Проблема извлечения именованных сущностей связана, прежде всего, с тем, что даже в большом текстовом массиве они могут встречаться редко. Вторая трудность заключается в открытости информационного ресурса. Если крупные географические объекты относительно полно покрываются газетирами, то другие именованные сущности достаточно быстро могут появиться в корпусе текстов, однако их регистрация в списках требует значительного времени. Третья проблема извлечения именованных сущностей сводится к проблеме снятия омонимии. Таким образом, несмотря на все существующие подходы к автоматическому извлечению данных и знаний из больших массивов неструктурированных данных, до сих пор не решены важные проблемы информационного поиска.

Английский термин *Named Entity* (*NE*) – именованная сущность – был впервые использован в 1995 голу в рамках конференции *MUC (Message Understanding Conferences)* в связи с постановкой задачи автоматического распознавания именованных сущностей [1]. Именованной сущностью считается слово или словосочетание, предназначенное для конкретного, вполне определённого предмета или явления, выделяющее этот предмет или явление из ряда однотипных предметов или явлений. Именованная сущность обязательно имеет референт, т.е. объект внеязыковой действительности, подразумеваемый автором конкретного речевого отрезка. Референт может принадлежать реальному миру, например, фамилия, имя и отчество (ФИО) конкретного человека или название конкретной организации, либо вымышленному миру, например, быть персонажем художественного произведения.

В письменном тексте именованные сущности представлены объектами и их атрибутами. Под объектами понимаются либо персоны, имеющие такие атрибуты, как фамилия, имя, отчество, должность, звание и т.д., либо органи-

зации и топонимы, характеризующиеся такими атрибутами, как представленное в полном или сокращенном виде название предприятия, его адрес, направление деятельности и т.д. Именованные сущности характеризуются следующими типами признаков:

1) признаками уровня слова (N-граммами, суффиксами, префиксами, частями речи и т.д.);

2) признаками уровня текста (акронимами, позицией слова в предложении, наличием слова в заголовке и т.д.);

3) дополнительной информацией (газетирами, словами-указателями, например, *Inc.*, *Corp.*, стоп-словами, словами с капитализацией, которые не являются именованными сущностями и т.п.) [2].

Выделяют три категории атрибутов:

1) атрибуты, характерные непосредственно для текста документа или статьи, например, заголовок, дата создания документа, исполнитель документа;

2) атрибуты, характерные исключительно для лиц, например, год рождения, национальность, особенности внешности человека;

3) атрибуты, свойственные двум отмеченным выше категориям признаков, например, адрес, номер и марка машины и т.д.

Необходимо учитывать возможные варианты представления именованных сущностей в тексте. Такие типичные объекты как ФИО, дата, адрес, должность приводятся к одному стандартному виду. При выделении цепочек типа ФИО важным является отождествление формально различных ФИО в одном тексте. Исходя из предположения, что в тексте вряд ли могут упоминаться два лица с одним и тем же именем и фамилией, но являющиеся разными людьми, можно соединить неполные ФИО с более полными вариантами. Например, именованные сущности *А. В. Петров, Александр Петров* и *Александр Владимирович Петров* будут считаться относящимися к одному человеку [3]. Следует отметить, что выявление объектов осуществляется не только с учетом их кратких наименований (например, соотнесение отдельных фамилий или имен с полным ФИО), но также с учетом анафорических ссылок (указательных и личных местоимений, например, *этот человек, он*) и определений (например, *мэр Москвы Лужков* идентифицируется со словами *мэр, Лужков*). Такая информация выражается грамматически правильно записанными последовательностями слов и символов, которые называется значимыми компонентами текста. В состав значимых компонентов входят информационные и вспомогательные слова. Информационные слова определяют объекты и атрибуты. Здесь важную роль играют слова-классификаторы, наличие которых указывает на присутствие соответствующей информации. Например, в определенном типе текста, где важными являются приметы людей, слово *рыжий* при наличии в непосредственной близости лица указывает на то, что речь идет о внешнем виде человека. В другом типе текста слово *московский* указывает на географическое положение. К вспомогательным словам относятся те лексические единицы, без которых значимые компоненты не теряют своей сущности. К ним относятся предлоги, знаки пунктуации и так называемые шаблонные слова, указывающие на положение соответствующих информационных слов в тексте. С точки зрения их выяв-

ления значимые компоненты условно делятся на жесткие и мягкие. Жесткие значимые единицы состоят из фиксированного числа позиций или слов, например, ФИО и дата. У мягких компонентов количество позиций переменное, например, приметы человека или признаки организации. В зависимости от используемого способа выделения адрес может рассматриваться и как жесткий, и как мягкий значимый компонент. С грамматической точки зрения значимые компоненты состоят из обязательных элементов (именные слова, прописные буквы, части речи, числа) и вспомогательных элементов.

В русскоязычных новостных текстах для обозначения персон обычно используется имя (иногда также отчество) и фамилия или же просто фамилия. С учетом принципа подбора минимального языкового материала в качестве контекстуального синонима используется только фамилия. При этом необходимо проверять такие синонимы на наличие в пре- или постпозиции подходящего соответствующему объекту имени. Это может быть непосредственно имя, сочетание имени и отчества, инициалы. По умолчанию фамилия без имени идентифицируется с тем же объектом, с которым идентифицируется эта же фамилия в сочетании с именем. Более сложный анализ необходим, если в тексте упоминается несколько персон, носящих одну и ту же фамилию (например, *Мишель Обама* и *Барак Обама*). Исключение из описанного правила составляют широко известные персоны, для обозначения которых могут использоваться только фамилии. В данном случае применимо следующее правило «непротиворечивости имени»: если в тексте нет сочетания имени и фамилии-синонима, в котором имя не соответствует имени объекта, то считать данную фамилию верифицированным синонимом для данной персоны. Кроме фамилий, в качестве синонимов (которым не нужна проверка на корректность имени) могут использоваться: псевдоним, сценическое имя или прозвище человека (например, *Борис Акунин*, *Витас*); сочетание имени и фамилии (это актуально, когда существует несколько варианта правописания имени или когда у человека есть несколько имен и могут использоваться их разные, например, *Абдалла ибн Абдель Азиз Ал Сауд*, *Абдуллах бин Абдул Азиз ал Сауд*, *Абдель Азиз Аль-Сауд*); имя человека и его статус (*Королева Елизавета*, *Патриарх Московский и Всея Руси Кирилл*); имя иностранной персоны, написанное латиницей (*Britney Spears*).

В наши дни в новостных сообщениях освещаются самые разные события, относящиеся к различным сферам: политической, социальной, экономической и многим другим. В зависимости от предметной области и тематики, к которой относится текст сообщения, в нем используются определенные факты, например, «визит», «встреча», «авария», «смерть», а также их атрибуты, например, «участник встречи», «место встречи», «дата начала визита», «место происшествия». Тексты новостных сообщений о дорожно-транспортных происшествиях написаны на основе официальной информации, предоставленной полицией, и могут включать высказывания-свидетельства очевидцев и пострадавших. В отличие от русскоязычных новостных сводок об авариях в англоязычных текстах данного типа очень часто представлена личная информация о родственниках пострадавших: пол, возраст, фамилия и имя. Например, *Raul Pachecho (a 63-year-old who works nearby); Wendy*

Mateo – a 39-year-old who was hurt; Three others, including 23-year-old Ilhajjam Hamid and 42-year-old Henriquez Wellington, Yuko Hopkins, a 32-year-old. Чаще всего, указателями персон являются такие атрибуты как *police officer* (*police, cops, officers*), *driver, sources, witnesses, authorities, a Muslim woman, vehicle's passenger, a driver of a BMW*. В текст новостной сводки об аварии включают данные о месте происшествия, например, название штата, города, местности, улицы (*New Jersey,* Atlantic City, *on Broadway, West 62nd Street*) и т.д. Как отмечалось выше, именованные сущности и факты напрямую связаны с тематикой новостного сообщения. Например, в текстах данного типа преобладают такие слова, указывающие на именованные сущности, как, например, *car accident,* crash, *incident, license plates, scene, suffered, car, collision, hospital, injuries,* а также слова, обозначающие их атрибуты: *chain–reaction, Bellevue, non-life threatening* и др.

Автоматическое извлечение именованных сущностей и фактов из письменных текстов может проводиться в рамках следующих подходов [4, с. 252–254].

1. Извлечение сущностей и фактов на основе признаков подразумевает наличие фиксированного набора признаков и весов использования этих признаков в контексте извлекаемых элементов текста. Для каждого извлекаемого элемента определяется характеризующий его вектор признаков. Процесс извлечения сводится к распознаванию некоторого сегмента текста на основе вероятностного анализа признаков, обнаруженных в контексте этого сегмента. Недостатками данного подхода является использование ограниченного размера контекста (как правило, не более 2–3 слов), необходимого для обеспечения требуемой точности извлечения. Учет контекста большего размера приводит к снижению полноты распознавания и к увеличению размера необходимой репрезентативной выборки, на которой собирается статистика для расчета оценок вероятностей.

2. Извлечение сущностей и фактов на основе ядер устраняет часть недостатков предыдущего подхода за счет алгоритмического определения меры подобия между представлениями сопоставляемых текстовых сегментов. Суть данного подхода сводится к замене скалярного произведения векторов, отражающих признаковое представление распознаваемых элементов, некоторой функцией – ядром. Такая функция задается алгоритмически и учитывает более сложное представление распознаваемых элементов и их контекстов, как правило, – древовидное, описывающее синтаксическую структуру текстового сегмента. Для древовидных представлений расчет ядра, чаще всего, сводится к сопоставлению всех вложенных деревьев в исходные. Недостатком данного подхода является высокая сложность расчета ядер и определения синтаксической структуры текстового сегмента.

3. В основе подхода, основанного на сопоставлении образцов, лежит понятие «образец» и правила сопоставления образцов с фрагментами текста. Образцы представляют собой цепочки ограничителей (символы, слова, части речи и семантические классы), являющиеся своего рода лексическими шаблонами текстовых сегментов. В этом отношении данный подход аналогичен ядерному подходу при условии, что текстовые сегменты имеют «плоское» представление, а не древовидное.

4. Подход, основанный на фразовых образцах, является своего рода компромиссом между подходом, основанным на ядрах, и подходом, основанным на сопоставлении текстовых образцов. Текстовые сегменты, также как и в основанном на ядрах подходе, представляются в виде деревьев, отражающих результаты синтаксического анализа предложений. Однако вместо сложного расчета ядер в данном случае выполняется более простая процедура сопоставления сегмента текста с синтаксическим шаблоном. Чаще всего, для отражения синтаксических связей между единицами текста используются контекстно-свободные грамматики, которые позволяют оценить вероятность применения того или иного правила для конкретного фрагмента текста и выбрать правило с максимальной степенью вероятности. При этом на некоторой обучающей выборке текстов, предварительно размеченной человеком, осуществляется ручное составление формальных правил и вычисление оценок вероятностей их употребления. Разметка содержит также указания на правила грамматики, которые необходимо применять при извлечении данных элементов.

ЛИТЕРАТУРА

1. *Кузнецов, И.* Методики выявления объектов и связей, заданных в неявном виде [Электронный ресурс] / И. Кузнецов. – Режим доступа : http://www.dialog-21.ru/digests/dialog2012/materials/pdf/%D0%9A%D1%83%-D0%B7%D0%BD%D0%B5%D1%86%D0%BE%D0%B2_%D0%98_%D0%-9F.pdf. – Дата доступа : 18.03.16.

2. *Котельников, Д. С.* Итерационное извлечение шаблонов описания событий по новостным кластерам / Д. С. Котельников, Н. В. Лукашевич // Тр. XIV Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» RCDL–2012, Переславль-Залесский, 15–18 окт. 2012 г. – Переславль-Залесский, 2012. – С. 292–298.

3. *Голованова, Е. И.* Соотношение естественного и искусственного начал в языке профессиональной коммуникации / Е. И. Голованова // Языки профессиональной коммуникации : матер. Междунар. науч. конф. – Челябинск, 2003. – С. 26–33.

4. *Кормалев, Д. А.* Обобщение и специализация при построении правил извлечения информации / Д. А. Кормалев // Тр. X нац. конф. по искусственному интеллекту «КИИ–2006», Обнинск, 25–28 сент. 2006 г. : в 3 т. – М. : Физматлит, 2006. – Т. 2. – С. 572–579.