УДК 81'27 + 81’42

ИМЕНОВАННЫЕ СУЩНОСТИ В КОРПУСЕ ТЕКСТОВ НОВОСТНЫХ СООБЩЕНИЙ: ЛИНГВИСТИЧЕСКОЕ ОПИСАНИЕ

О.И. Бабина

В работе обоснована необходимость проведения лингвистического анализа именованных сущностей с целью построения пространства признаков для репрезентации сущностей. Определен перечень классов именованных сущностей, встречающихся в корпусе новостных сообщений. На основе анализа обучающей выборки с размеченными вручную сущностями и построения конкордансов именованных сущностей выявлены лингвистические характеристики классов именованных сущностей и их локальных контекстов.

Ключевые слова: именованная сущность, контекст, корпусная лингвистика.

Понятие именованной сущности было введено 1990-х годах на конференции по пониманию сообщений (Message Understanding Conferences, MUC). Именованными сущностями называют имена собственные, которые могут обозначать персону, название организации, место, а иногда – числовую информацию, такую как время, даты, процентные отношения, валюту и т.п. [1].

Одной из подзадач анализа слабоструктурированных данных является автоматическое распознавание именованных сущностей – идентификация именованных сущностей в тексте и их классификация.

Формально задача классификации именованных сущностей может быть сформулирована следующим образом: пусть имеется множество объектов $\{O\}$ и множество классов $\{C\}$, которым эти объекты могут соответствовать.

Каждый объект во множестве $\{O\}$ представлен как вектор признаков: $\boldsymbol{O} = (o_1, o_2, \dots, o_K)$; тогда необходимо определить для O единственный класс C. Общим подходом к выявлению принадлежности объекта $\boldsymbol{O}$, представленного как вектор признаков, классу C является сопоставление вектора объекта с некоторой репрезентацией класса C в форме вектора той же размерности, что и вектор объекта $\boldsymbol{O}$. Таким образом, класс C, так же как и классифицируемый объект $\boldsymbol{O}$, представляется вектором размерности K: $C = (c_1, c_2, \dots, c_K)$.

Очевидно, что в такой постановке ключевым компонентом является этап определения пространства признаков (атрибутов, переменных, размерностей), оказывающих влияние на классификацию объектов – это основная задача, которая решается индивидуально, и, в конечном счете, выбор признаков, при сходстве подходов к сопоставлению объектов, предопределяет качество работы конкретного алгоритма.

Построение пространства признаков остается нетривиальной задачей. Учитывая, что в поставленной классификационной задаче в качестве объектов классификации выступают элементы лингвистической природы, нам представляет целесообразным выполнять отбор признаков на основе лингвистического анализа функционирования именованных сущностей на реальном языковом материале. Данная работа представляет итоги проведенного лингвистического анализа именованных сущностей в корпусе текстов.

Исследование проводилось на материале корпуса текстов новостных сообщений предметной области «Экономика» с новостных сайтов news.google.com, news.mail.ru, news.yandex.ru общим объемом 130 тыс. словоупотреблений. На основе семантического анализа содержания текстов было определено множество классов именованных сущностей, встречающихся в текстах:

C = {**PERSON, COUNTRY, CITY, REGION, COMPANY, UNION**},

где **PERSON** – имена физических лиц (например, *Путин, Дмитрий Медведев*);

COUNTRY – название государства (например, *Армения, США, РФ*);

CITY – название города (например, *Москва, БЕРЕЗНИКИ*);

REGION – название местности (например, *Нижегородская область*, *Остоженка, ЦАО, Северная Дакота*);

COMPANY – название организации, компаний, юридических лиц (например, *Уралкалий*, *Пенсионный фонд*);

UNION – искусственно созданные объединения, отделы, подразделения и т.п. (например, *Минобрнауки*, *Государственная Дума*, *Евросоюз*).

Из представленного корпуса – в обучающую выборку вошел подкорпус объемом около 7 тыс. словоупотреблений, в котором именованные

сущности были размечены вручную. В общей сложности, количество размеченных именованных сущностей составило 658 употреблений.

Основной этап исследования включал сравнительный анализ именованных сущностей между собой и с остальными лексическими единицами в тексте с целью определения отличительных характеристик, которые бы могли иметь идентифицирующую силу: способствовали бы отграничению именованных сущностей от других элементов в тексте, а также разграничению классов именованных сущностей.

В подавляющем большинстве именованные сущности представлены словами, начинающимися с заглавной буквы: *Дмитрий Медведев, Россия, Москва, Минтруд, Центробанк, «Единая Россия», Еврокомиссии* и т.д.

Следует заметить, однако, что капитализация в русском языке, традиционно, кроме обозначения имен собственных, имеет смысл обозначения начала предложения.

Кроме того, из 658 размеченных лексических единиц – 180 имеют написание со строчной буквы. Анализ таких вхождений позволяет разделить эти элементы на следующие группы:

1. Обозначения стран в форме прилагательного: *российский, американский, армянский*. В нашей работе такие элементы относятся к определенному классу именованных сущностей, так как семантически соответствуют использованию существительного, обозначающего соответствующую страну.

2. Использование лексической единицы (существительного или прилагательного) в составе многокомпонентной именной группы (именованной сущности), первое слово которой пишется с заглавное буквы, а остальные со строчной: *Московская область, Европейского союза, Всемирной торговой организации, Пенсионный фонд России, Международное энергетическое агентство США, Арбитражный суд Нижегородской области, Кабинет министров Армении*. Как показывают примеры из корпуса текстов, подобные наименования образуют именную группу. В ряде случаев, слово, записанное с заглавной буквы, представляет собой прилагательное, которое синтаксически не может функционировать отдельно, только в составе более длинной именной группы.

3. Использование лексической единицы (существительного или прилагательного) в составе многокомпонентной именной группы (именованной сущности), все слова которой пишутся со строчной буквы: *министерство финансов РФ*. Как можно видеть из примеров, взятых из корпуса текстов, в таких словосочетаниях, в качестве определительной именной группы в пост-позиции, зачастую используются наименования локаций (стран, регионов, городов).

Кроме того, ряд именованных сущностей имеет начертание прописными буквами. Среди таких сущностей выделяются:

- города (обусловлено форматом статьи – обычно используется как первой слово в статье). Например, *МОСКВА, БЕРЕЗНИКИ*;
- аббревиатуры, обозначающие государства (*РФ, США*), организации (*МЭА, ЦБ, ПФР, ВТО*), включая компоненты названий организации (ОАО, ООО), локации-регионы (*МКАД, ЦАО*);
- названия организаций, записанной заглавными буквами (*ФИНАМ, РУМО, ВТБ Капитал*).

Анализ **структуры** сущностей различных типов показывает следующее:

1. Сущность **PERSON**: структурно может быть представлена с помощью следующих шаблонов:

1) имя + фамилия: например, *Александр Жилкин, Дмитрий Медведев, Владимир Путин*;

2) фамилия: например, *Присяжнюк, Медведев, Путин, Потанин*;

3) имя + отчество: например, *Владимир Владимирович*;

4) инициал(ы) имени + фамилия: например, *Д. Медведев, Я. Стурнарас, И. Сечин*.

Кроме того, в ряде случаев указание на персону осуществляется описательно: в тексте явно не указывается имя человека, однако функциональное описание однозначно определяет анафор, отсылающий к одному из имен в тексте. В качестве описаний используются выражения со следующей структурой:

1) должность + локация (страна, регион): например, *губернатор Астраханской области, президент России, президент РФ, президент Таджикистана, министр финансов Греции, президент республики*;

2) страна-прилагательное + должность: например, *таджикский лидер, российский президент, российский лидер*;

3) должность + организация: *глава Минкомсвязи РФ.*

2. Сущность **COMPANY** включают наименования компаний, фирм, банков и т.д. В основном, названия организаций записываются кириллицей. Первичные способы репрезентации именованных сущностей этой группы строятся по следующим шаблонам:

1) именная группа – (Adj) N (N <род. п.>): *Федеральная антимонопольная служба*;

2) именная группа + локация <род. п.>: *Центральный банк России, Пенсионный фонд России*;

3) аббревиатура + «именная группа»: *ГМК «Норильский никель», ОАО «Мобильные Телесистемы»*;

4) сокращенное название организации (включает часть компонентов полного названия или представляющее собой составное слово, образованное из корней словосочетания): *Центробанк, «Норникель», антимонопольная служба*;

5) сокращенное название + локация <род. п.>: *Минкомсвязи России*;

6) локация (прил.) + сокращенное название*: Европейский Центробанк*;

7) аббревиатуры: *ФАС, МТС.*

Кроме того, в отличие от других именованных сущностей, названия иностранных компаний могут не переводиться и записываться в тексте латиницей. Такие названия могут иметь следующую структуру:

1) полное название организации (словосочетание): *Shagang Group, Royal Dutch Shell, Randgold Resources, New York Mercantile Exchange*;

2) сокращенное название организации (одно из слов словосочетания): *Shagang, Shell, Randgold*;

3) аббревиатуры: *NYMEX, HSBC Bank.*

3. Сущность **UNION** включает различные отделы, департаменты, политические объединения. В текстах представляются с помощью шаблонов:

1) именная группа: *Европейский союз*;

2) локация (прил.) + тип объединения: *российское правительство*;

3) именная группа + локация <род. п.>: *кабинет министров Армении, Арбитражный суд Новгородской области*;

4) сокращенное название: *Евросоюз, Минтруд*;

5) сокращенное название + локация <род. п.>: *кабмин РФ, Минфин России*

6) аббревиатура: *ВТО, ЕС, ОПЕК.*

Заметим, что зачастую основные компоненты объединений записываются со строчной буквы.

4. **Локации** в нашей модели представлены сущностями **COUNTRY** (названия стран), **REGION** (названия частей света – *Юг*; континентов – *Европа*; территориальных подразделений государства – *Челябинская область, Бурятия, Башкортостан*; улиц – *Остоженка*), **CITY** (названия городов). Именованные сущности такого типа репрезентируются при помощи следующих шаблонов:

1) полное название: *Санкт-Петербург, Бурятия, Российская Федерация*;

2) сокращенное название: *Россия, Соединенное Королевство*;

3) аббревиатура: *РФ, КНР*;

4) описательный оборот: *Северная столица, олимпийская столица, автомобильная столица США*.

Полное название может состоять из нескольких слов, в этом случае – основное слово именной группы представлено именем нарицательным. Семантически основное существительное таких словосочетаний дает представление о типе локации: например, *федерация, штаты* – COUNTRY; *область* – REGION и т.д.

Во многом, задача распознавания именованных сущностей сродни задаче разрешения неоднозначности: и в том и в другом случае необходимо

определить семантический аспект лексической единицы. При этом практически все работы по снятию многозначности основываются на сведениях, которые предоставляют контекст неоднозначного слова. Логично предположить, что контекст также может оказать помощь в идентификации семантического класса именованной сущности.

Обычно контекст рассматривается в одном из двух аспектов:

1) контекст представляется как *набор лексических единиц вокруг* «целевой» лексической единицы, сгруппированный без учета морфологических отношения и без учета расстояния;

2) контекст рассматривается в терминах некоторого *отношения* с целевой лексической единицей (дистанция, синтаксические связи, орфографические признаки, морфологические признаки, лексико-грамматические признаки контекста и т.д.).

Информация, которая может быть полезна и учтена для семантической идентификации слова, может включать:

1) информацию о *локальном* (или *микро*-) контексте (несколько слов в ближайшем окружении целевого слова);

2) информацию о *тематическом* контексте (несколько предложений вокруг целевого слова);

3) информацию об *экстралингвистическом* контексте, определяемым областью знаний, для которой решается проблема разрешения многозначности [2].

Однако, чем «выше» уровень информации, тем сложнее она поддается формализации. Наиболее простым способом учета контекста является учет ближайших слов целевой лексической единицы, то есть локального контекста.

При рассмотрении локального контекста необходимо решить вопрос, контекст какой длины следует рассматривать. Д. Яровски выдвинул гипотезу, что для устранения локальной многозначности достаточно 3–4 слов контекста, в то время как для смысловой многозначности нужно большее окно, которое должно состоять из 20–50 лексических единиц [3]. Существуют данные о том, что учет большего контекста (100 слов) позволяет улучшить точность разрешения неоднозначности [4], однако при этом дистанция до целевого слова обратно пропорционально зависит от важности данного контекста для определения смысла лексической единицы.

Согласно нашему наблюдению, значимые контекстные лексические единицы в корпусе текстов, чаще всего, располагаются контактно и включают не более 3 слов, например, *олимпийская столица (Сочи)*, *тайваньская металлургическая компания (China Steel).* Вместе с тем, в редких случаях в корпусе встречаются случаи, когда семантически значимый контекст располагается на более отдаленном расстоянии. Например, ***министр*** *по вопросам информационных технологий и энергетики Швеции Анна-*

Карин Хатт. Здесь слово с обозначением должности (такие слова в совей семантической структуре содержат одушевленность и, по нашему предположению, могут выступать маркерами именованных сущностей класса **PERSON**) однозначно указывает на принадлежность последующей именованной сущности (*Анна-Карин Хатт*) классу **PERSON**, при этом расположение этого контекста включает дистанцию в 8 слов.

Учитывая эти наблюдения, для выявления локального контекста для найденных именованных сущностей обучающей выборки были составлены конкордансы по материалам полного корпуса, позволяющие видеть контексты употребления данной единицы в пределах 5 слов слева и справа.

Анализ **контекстов** сущностей различных типов показывает следующее:

1. Сущность **PERSON**: зачастую, имя персоны сопровождается уточнением относительно должности, занимаемой указанным лицом. Указание на должность осуществляется посредством именной группы, расположенной перед упоминанием имени лица и синтаксически связанной с ним аппозитивной связью. Структура такой контекстной группы может быть представлена следующими семантическими компонентами, выраженными именными группами:

1) должность: например, *губернатор* (*Александр Жилкин*), *президент* (*Владимир Путин*), *премьер* (*Дмитрий Медведев*), *президент* (*Путин*);

2) должность + локация (страна, регион) <род. п.>: например, *президент России* (*Владимир Путин*), *министр финансов Греции* (*Яннис Стурнарас*);

3) должность + организация <род. п.>: например, *аналитик «ВТБ Капитал», глава «Роснефти», топ-менеджер РУСАЛа, замглавы Минэкономразвития, гендиректор «Норникеля».*

Наименования локаций и организаций в составе обозначения персоны выражены существительными (в общем случае, именными группами) в родительном падеже или аббревиатурами (синтаксическими группами с главным словом-аббревиатурой).

Кроме того, в предложении обозначения персон зачастую используется в позиции подлежащего, где сказуемым является глагол речевой деятельности: *сообщил, заявил, подписал, выступил.*

2. Сущность **COMPANY** может включать следующие аппозитивные контекстные лексические единицы:

1) тип организации: *авиакомпания* («*Турецкие авиалинии*»), *компания* (*ПВХ-Сибирь*);

2) локация (прил.): *австралийская* (*Fortescue Metals Group Ltd*), *индийская* (*JSW Steel*);

3) локация (прил.) + тип организации: *британская компания* (*Antofagasta Plc*);

4) тип деятельности (прил.) + тип организации: *горно-металлургическая компания* «*Норильский никель*»;

5) локация (прил.) + тип деятельности (прил.) + тип организации: *Бразильская сырьевая компания* (*Vale SA*), *тайваньская металлургическая компания* (*China Steel*).

В предложении организации используются чаще с предикатами действия: *взаимодействует, анализирует, выполняет.*

3. Сущность **UNION** в значительной степени используется независимо (как отдельный аргумент, не включающих аппозитивных уточнений). Среди некоторых регулярных левых контекстов можно упомянуть:

1) компоненты объединения: *государствах* (*Евросоюза*), *страны-члены* (*ВТО*). В этом случае объединение выступает как зависимый член именной группы в родительном падеже;

2) регуляторные механизмы: *инструменты (ВТО), нормы, правила, условие, требования, рамки*;

3) лексика экономико-политической тематики (типы организаций): *руководство, рынки, банки, биржи, ВВП, бюро, (энергетическая) группа*;

4) предлоги *к, в, с*.

4. Контексты сущностей, обозначающих локации, включают:

1) тип локации: *город, государство, республика.* Этот контекст в явной форме указывает на подтип именованной сущности;

2) прилагательное: *олимпийский Сочи.*

Локации часто употребляются в предложении в функции обстоятельства места, и используются после предлогов *в, из*.

Выявленные характеристики именованных сущностей в рассмотренном корпусе текстов предоставляют материал для определения пространства признаков, позволяющего классифицировать именованные сущности в текстах на русском языке. По результатам анализа можно сделать вывод, что значимой является лексико-грамматическая информация об именованной сущности, наличие в локальном контексте слов с определенной семантикой, которые, в свою очередь, могут отчасти быть идентифицированы посредством анализа деривационной структуры слова или внесены в специальные лексиконы. Кроме того, лексикографическое решение в случае относительно закрытых классов сущностей (названия стран, городов, русские имена и т.п.) может касаться и собственно сущностей. Составление пространства признака по итогам проведенного исследования является дальнейшей перспективой данной работы.

Библиографический список

1. Sundheim, B.M. Overveiw of Results of MUC-6 Evaluation / B. M. Sundheim // Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference (Columbia, Maryland, November 6–8 1995). – Morgan Kauffman Publishers Inc., 1995. – Pp. 13–32. – URL: http://aclweb.org/anthology/M/M95/M95-1002.pdf.

2. Hockett, C.F. A course in modern linguistics / C.F. Hockett. – New York, 1958. – 407 p.

3. Yarowsky, D. One sense per collocation / D.Yarowsky // HLT ’93: Proceedings of the workshop on Human Language Technology. – New York, 1993. – Pp. 266–271.

4. William, A. One sense per discourse / A.William, K. Gale, W. Kenneth, D. Yarowsky // HLT ’91: Proceedings of the workshop on Speech and Natural Language. – Morristown, NJ, USA: Association for Computational Linguistics, 1992. – Pp. 233–237.

К содержанию