The article deals with specific linguistic features of the formal model of automatic actualization of road accidents object database. Different types of user's queries to the database are under consideration.

В. Н. Мурашко
Минск, МГЛУ

ИМЕНОВАННАЯ СУЩНОСТЬ И ЕЕ КАТЕГОРИИ

В статье рассматривается понятие именованной сущности и основные категории, к которым ее можно отнести: человек, место, организация, дата и время. Отмечается, что выбор соответствующей категории не всегда определяется самой сутью именованной сущности. На примерах показано, что в ряде случаев категория зависит от микроконтекста ее употребления. В работе указаны основные трудности, с которыми может столкнуться система автоматического распознавания именованных сущностей в письменном тексте.

Под именованной сущностью понимают «что-либо реально существующее или вымышленное, на что можно указать или к чему можно обратиться при помощи имени собственного» [1, с. 214]. В соответствии с этим определением в задачу распознавания именованных сущностей входит не только нахождение их в тексте, но и однозначное указание на подразумеваемый объект или лицо, а также приписывание ему определенной категории. Чаще всего используется простая классификация, включающая в себя следующие три категории: ЧЕЛОВЕК (сокращенно ПЕР от *персона*), МЕСТО (сокращенно ЛОК от *локация*), ОРГАНИЗАЦИЯ (сокращенно ОРГ) [2, p. 1538]. Такая классификация является довольно простой, поскольку имеются задачи, где важно различать подтипы именованных сущностей. Например, если идет речь о компании или стране, об актере или политике и т.п. Однако разработать более подробную схему, подходящую для текстов разных жанров, намного сложнее. Хотя ни даты, ни числа не соответствуют приведенному выше определению, они часто распознаются как именованные сущности вместе с ПЕР, ЛОК и ОРГ. Для обозначения этих категорий используются сокращения ТЕМП (темпоральные выражения) и НУМ (нумерические выражения) [Там же, p. 1539]. Следует отметить, что, в отличие от трех других категорий, распознавание чисел и времени представляется значительно более легкой задачей, поскольку существует ограниченный набор способов их выражения. Ниже в таблице приведены различные именованные сущности и соответствующие им категории, содержащиеся в примере, взятом из Википедии: *Современный* [***СПбГУ***] *в* [***России***] *– преемник* [***Академического университета***], *который был учрежден одновременно с* [***Академией наук***] *указом* [***Петра I***] *от* [***28 января 1724 года***]. *В частности, в* [***1758–1765***] *годах ректором* [***Академического университета***] *был* [***М. В Ломоносов***].

Как видно из таблицы, помимо очевидной категории «человек», М. В. Ломоносову можно приписать и другие категории, соответствующие его многогранной личности. В данном случае речь идет не о более подробной

детализации, как в случае с СПбГУ (от организации к учебному заведению и дальше к вузу), а о категориях одного порядка. Тем не менее тот факт, что М. В. Ломоносов был в том числе и художником, в контексте примера не имеет такого большого значения, как то, что М. В. Ломоносов был крупным ученым.

Именованные сущности и соответствующие им категории

Название (имя)	Возможные категории
СПбГУ Академический университет Академия наук	организация, образовательное учреждение, вуз организация, образовательное учреждение, вуз научная организация, академия
Россия	место, страна, государство
Петр I М. В. Ломоносов	человек, исторический деятель, политик, правитель человек, ученый, химик, писатель, философ, художник
28 января 1724 года 1758–1765	время (дата) время (временной отрезок)

Можно сказать, что из категорий одного порядка (например, профессий) в конкретном контексте, как правило (но не всегда), имеет значение только одна из возможных категорий, и ее выбор часто не является очевидным. Более того, подобная проблема может возникнуть и с семантически несхожими категориями. Так, в рассмотренном выше примере Россия является географическим объектом, местом, чего нельзя сказать об этой именованной сущности в следующем примере, взятом из электронного издания «Утро.Ru»: ***Россия*** *отказалась от американского мяса. Россельхознадзор вводит временные ограничения на поставки продукции птицеводства США в* ***Россию***. Географические объекты не могут от чего-либо отказываться, и во втором предложении речь уже идет об организации (*Россельхознадзор*). Поэтому для *России* более подходящей категорией будет категория «страна». Таким образом, выбор соответствующей категории не обязательно следует из именованной сущности и не может быть легко осуществлен из лингвистической базы данных. Очень часто категория определяется контекстом, в котором она была упомянута.

В примере, взятом из Википедии, понятно, какой *Петр I* или какой *М. В. Ломоносов* имеется в виду. В случае Петра I нет ни одного другого известного в истории российского государства человека с таким же именем. Русская версия электронной энциклопедии приводит список из шести Ломоносовых, но имеет инициалы *М. В.* только один из них. Совершенно иначе воспринимается имя *Толстой* в следующем отрывке из «Театрального романа» М. А. Булгакова:

В час ночи мы выпили чаю, а в два Рудольфи дочитал последнюю страницу. Я заерзал на диване.

– Так, – сказал Рудольфи.

Помолчали.

– Толстому подражаете, – сказал Рудольфи.

Я рассердился.

– Кому именно из Толстых? – спросил я. – Их было много... Алексею ли

Константиновичу, известному писателю, Петру ли Андреевичу, поймавшему за границей царевича Алексея, нумизмату ли Ивану Ивановичу или Льву Николаичу? [1, с. 216].

В данном случае нельзя однозначно сказать, какая из известных исторических личностей имеется в виду, хотя с большой долей вероятности читатель может предположить, что речь идет о последнем из перечисленных в отрывке персонаже. В отличие от этого примера, в абсолютном большинстве случаев неоднозначность не предполагается автором текста, и идентификация референта редко вызывает трудности у читателей. Однако для автоматических систем однозначное распознавание именованных сущностей остается далеко непростой задачей, поскольку у значительного числа имен собственных есть несколько возможных референтов.

Нередко возникает вопрос о том, нужно ли проводить границу между двумя сущностями или их можно считать единым объектом. Возвращаясь к первому примеру, рассмотрим единицы *Академический университет* и *СПбГУ*. Из текста следует, что один является преемником другого и что именно это дает возможность говорить о том, что дата основания СПбГУ – *1724 год*. С другой стороны, признавая, что СПбГУ и Академический университет – одно и то же заведение, можно лишить текст смысла, потому что утверждение *СПбГУ является преемником СПбГУ* не несет в себе никакой информации. В этом контексте имеет смысл в базе данных приписать этим двум заведениям разные коды (идентификационные номера), а в ряде других контекстов считать, что речь идет об одном и том же заведении. Подобные сложности возникают, когда нарушается исходное положение о том, что сущности являются инвариантами, то есть, что они не подвержены изменениям во времени. Особенно очевидна неоправданность этого допущения, когда речь идет, например, о странах, чьи границы неминуемо изменяются на протяжении всего времени их существования.

Все вышесказанное необходимо учитывать при разработке систем автоматического распознавания именованных сущностей в письменном тексте. Как правило, системы автоматической обработки текста построены по принципу конвейера: сначала осуществляется деление текста на слова и предложения, далее следует частеречная разметка и определение границ фраз, после чего делается синтаксический анализ и распознавание именованных сущностей. Предположим, что в тексте уже определены как имена собственные, так и все другие именные группы. Таким образом, задача состоит в том, чтобы приписать уже найденным именным группам их категории и установить их референтов в тех случаях, где это возможно и имеет смысл. Будем исходить из того, что в распоряжении разработчиков системы имеется обширный каталог именованных сущностей. Учитывая размер и постоянный рост Википедии и других ресурсов, например, *Freebase* и *Wikidata*, данное предположение имеет право на существование. Однако необходимо подчеркнуть, что ни один ресурс не является исчерпывающим.

Как и при создании других модулей обработки текста, распространенным подходом в контексте рассматриваемой проблемы является обучение статистических моделей на размеченных текстах. Подобно частеречной разметке,

многие подходы основаны на вариантах моделей марковских цепей, условных случайных полей или на классификаторах [1, с. 218]. Многие модели позволяют ответить на следующие вопросы.

1. Имеет ли данная именная группа своим референтом некую именованную сущность?

2. Какая категория из списка лучше всего соответствует именной группе в данном контексте?

3. Кто или что является референтом данной именной группы?

Во всех приведенных выше примерах у имен собственных имелся конкретный референт. Однако в следующем предложении вряд ли имеется в виду столица Франции: *Дмитрий Рогозин: «Генералы в окопах должны быть, а не в парижах».*

Эффективность параметров для систем машинного обучения зависит от языка и жанра текста. Так, во многих языках важным признаком именованных сущностей является их написание с заглавной буквы. Однако, например, в немецком языке, где все существительные пишутся с заглавной буквы, и в текстах, полученных при автоматическом распознавании речи, этот параметр оказывается бесполезным. Важным фактором является также наличие леммы слова в списке известных именованных сущностей. Для слов и именных групп, отсутствующих в списке, решение о том, идет ли речь об именованной сущности и, если да, то какого типа, принимается на основе анализа микроконтекста. Например, слово *Mrs*. в английском языке, *Frau* в немецком языке или *госпожа* в русском языке с большой долей вероятности сигнализируют о том, что за ним следует имя человека.

ЛИТЕРАТУРА

1. Прикладная и компьютерная лингвистика / под ред. И. С. Николаева, О. В. Митрениной, Т. М. Ландо. – М. : ЛЕНАНД, 2016. – 320 с.

2. *Fader, A.* Identifying relations for open information extraction / A. Fader, S. Doderland, O. Etzioni // Proc. of EMNLP-11. – 2011. – P. 1535–1545.

The article deals with the concept of the named entity and the categories it may be referred to. The main problems of automatic named entities detection and extraction from written texts are under consideration.

А. М. Савич
Минск, МГЛУ

ВИДЫ ТРАНСФОРМАЦИЙ ПРИ АВТОМАТИЧЕСКОМ ПЕРЕВОДЕ АНГЛОЯЗЫЧНОГО ОФИЦИАЛЬНО-ДЕЛОВОГО ТЕКСТА НА РУССКИЙ ЯЗЫК

В статье представлены результаты и основные выводы исследования лексических трансформаций при автоматическом переводе англоязычных текстов 17 целей ООН в области устойчивого развития на русский язык.