ВОПРОСЫ ОНОМАСТИКИ
PROBLEMS OF ONOMASTICS

Daria M. Golikova
Ural Federal University, Ekaterinburg, Russia

Proper Names and Named Entities Recognition in the Automatic Text Processing.

Review of the book: Nouvel, D., Ehrmann, M., & Rosset, S. (2016). *Named Entities for Computational Linguistics*. London; Hoboken: ISTE Ltd; John Wiley & Sons, Inc., 2016.

Voprosy onomastiki, 2018, Volume 15, Issue 1, pp. 207–215
DOI: 10.15826/vopr_onom.2018.15.1.012

Language of the article: Russian

Голикова Дарья Михайловна
Уральский федеральный университет, Екатеринбург, Россия

Распознавание имен собственных и «именованных сущностей» при автоматической обработке текста.

Рец. на кн.: *Nouvel D., Ehrmann M., Rosset S.* Named Entities for Computational Linguistics / D. Nouvel, M. Ehrmann, S. Rosset. — London ; Hoboken : ISTE Ltd : John Wiley & Sons, Inc., 2016. — 170 p.

Вопросы ономастики. 2018. Т. 15. № 1. С. 207–215
DOI: 10.15826/vopr_onom.2018.15.1.012

Язык статьи: русский

Ural Federal University
named after the first President of Russia B.N.Yeltsin
ИНСТИТУТ РУССКОГО ЯЗЫКА им. В. В. Виноградова Российской АКАДЕМИИ НАУК

DOI: 10.15826/vopr_onom.2018.15.1.012
УДК 81:004 + 81’373.23 + 81’42

Д. М. Голикова
Уральский федеральный университет
Екатеринбург, Россия

РАСПОЗНАВАНИЕ ИМЕН СОБСТВЕННЫХ И «ИМЕНОВАННЫХ СУЩНОСТЕЙ» ПРИ АВТОМАТИЧЕСКОЙ ОБРАБОТКЕ ТЕКСТА

Рец. на кн.: *Nouvel D.*, *Ehrmann M.*, *Rosset S.* Named Entities for Computational Linguistics / D. Nouvel, M. Ehrmann, S. Rosset. — London ; Hoboken : ISTE Ltd : John Wiley & Sons, Inc., 2016. — 170 p.

В рецензии представлен обзор книги Дамьена Нувеля (*Damien Nouvel*), Мод Эрманн (*Maud Ehrmann*) и Софи Россе (*Sophie Rosset*) «Именованные сущности в компьютерной лингвистике» (*Named Entities for Computational Linguistics*, 2016). Работа посвящена автоматической обработке текстов, написанных на естественном языке, и распознаванию в этих текстах «именованных сущностей» (*named entities*) с целью извлечения наиболее важной информации. Под именованными сущностями в работе понимается совокупность всех единиц, так или иначе указывающих на референта. Исследователи сравнивают эту категорию с именами собственными и дефинициями и в деталях освещают все этапы создания и применения алгоритмов по автоматическому аннотированию текста, а также различные методы оценки их эффективности. Имя собственное в данном контексте — вид именованной сущности, одна из типичных отсылок к референту, которую машина должна обнаружить в тексте и связать с конкретным явлением реальности. В книге приведен подробный обзор и анализ предшествующих исследований в рассматриваемом направлении, в основном на базе английского языка. Кроме того, представлены инструменты и ресурсы, необходимые для работы с подобного рода программами: аннотированные и неаннотированные корпуса, типологии и базы знаний. Положения работы подкреплены значительным количеством показательных примеров, работа алгоритмов

проиллюстрирована с помощью наглядных схем. Рецензируемая книга дает довольно полное представление о современном состоянии практических исследований в области автоматического распознавания и анализа имен собственных и других именованных сущностей, указывает на еще не решенные проблемы в данной области и предлагает пути решения для некоторых из них.

К л ю ч е в ы е с л о в а: компьютерная лингвистика, имена собственные, автоматическая обработка текста, аннотирование, именованные сущности, корпус, база знаний.

Вопрос о природе имени собственного (ИС) уже давно рассматривается лингвистами как с теоретической, так и с практической точки зрения. Особенно актуальной эта проблема становится на стыке ономастики и компьютерной и корпусной лингвистики, когда в рамках исследований семантической разметки корпусов и автоматической обработки текста (АОТ) появляется необходимость не только «научить» машину находить имена собственные, в каком бы виде они ни были представлены, но и определять их тип и «понимать» весь текст, вычленяя из него важную информацию.

Проблемам автоматической обработки ИС и связанной с ними информации посвящена книга Дамьена Нувеля (*Damien Nouvel*), Мод Эрманн (*Maud Ehrmann*) и Софи Россе (*Sophie Rosset*) «Именованные сущности в компьютерной лингвистике» (*Named Entities for Computational Linguistics*, 2016). Под «именованными сущностями» (*named entities*, NE) подразумеваются не просто ИС, но все единицы, так или иначе указывающие на референта, включая местоимения. ИС в этом случае выступает типом NE, т. е. отсылкой к референту, которую машина должна обнаружить в тексте и связать с конкретным явлением реальности. Это понятие, таким образом, оказывается более удобным для решения задач компьютерной лингвистики. Действительно, ставя перед собой цель научить машину «понимать» текст на естественном языке, исследователи не могут ограничиться простой семантической разметкой, выделяя в тексте только имена собственные и нарицательные. Для передачи важной информации единицы текста необходимо связать с конкретным референтом, и определения только ИС, относящихся к этому референту, будет недостаточно: важно найти в тексте не только все отсылки к референту в любой лексической форме, но и дополнительную информацию, которая позволила бы снять омонимию (при ее наличии). Использование для этих целей более широкой категории NE, таким образом, кажется более чем оправданным.

Авторы подробно рассматривают, что может включать в себя класс NE в контексте АОТ, какие проблемы могут возникать при их выделении среди других лексических единиц, каким образом и по каким параметрам происходит автоматическая идентификация лексических единиц как ИС и NE.

Работа состоит из шести глав, библиографии и пяти приложений: глоссарий используемых терминов, список проведенных ранее исследований, список доступных корпусов различных типов, схемы аннотирования (схемы автоматической

разметки предложения) и перечень определений понятия «именованная сущность» (*named entity*), сформулированных учеными в рамках различных исследовательских программ.

Первая глава состоит из двух разделов. В первом разделе приведен обзор исследовательских программ по распознаванию NE начиная с 1980-х гг., когда автоматическое «понимание» содержания текста стало одной из основных задач искусственного интеллекта. Второй раздел посвящен самому понятию NE как базового элемента текста, включающего ответы на вопросы «кто?», «что?», «где?» и «когда?».

Во второй главе более подробно рассматривается определение понятия NE, сложности, связанные с этим определением, и отношение NE к именам собственным (*proper names*) и дефинициям (*definite descriptions*). Авторы отмечают, что от исследования к исследованию объем класса NE постоянно увеличивается, и разрастание это естественно, так как класс NE формируется исходя из практических потребностей АОТ. Тем не менее во второй главе исследователи дают представление о границах рассматриваемого класса и о характеристиках NE. Одна из них — семантическая и лексическая неоднородность. В качестве NE можно рассматривать любое ИС, название любой единицы реальности, и может появиться впечатление, что NE может быть чем угодно. Это связано с тем, что, во-первых, единой классификации семантических категорий, пригодной для любых практических задач, до сих пор не существует; во-вторых, в некоторые категории входят любые единицы, которые не представляется возможным классифицировать, но необходимо обработать; в-третьих, существуют «двойные» (*dual-use*) категории, в которые включаются и метонимические употребления ИС и NE, что затрудняет категоризацию и без того крайне вариативных семантически единиц.

Еще одна проблема в изучении NE — это разнообразие их текстовой реализации (*mention diversity*). Поскольку категория NE шире категории ИС, то лексически NE может выражаться практически любым способом, что затрудняет определение ее границ. В конце раздела авторы приходят к выводу, что понимание NE на данном этапе скорее интуитивное, но это следствие специфики материала, так как ИС тоже не поддаются точной категоризации.

Далее приводится обзор существующих определений NE, при этом разные подходы объединены в две группы: определения с ономасиологической и семантической точек зрения. При первом подходе NE характеризуются через семантические категории, к которым они принадлежат («Человек», «Организация», «Место» и т. д.), при втором — рассматриваются «как объединяющий “контейнер” для всех лексических единиц, близких к категории ИС, а иногда и далеких от этой категории» (с. 19). В категории NE собираются все возможные способы указания на референта, таким образом, «все упоминания референта внутри документа, какой бы ни была лексическая форма этих упоминаний, идентифицируются и включаются в состав NE. Подход не ограничивается только ИС» (с. 51). При этом значение

некоторых NE часто невозможно понять без контекста. Например, личные местоимения могут быть связаны с референтом только при наличии дополнительной информации, а чтобы понять, когда происходит действие при указании на время словом *сейчас*, необходим более широкий временной контекст.

Стоит отметить, что своего собственного четкого определения анализируемому явлению авторы не дают, основываясь на том факте, что границы категории NE по-разному устанавливаются разными исследователями в зависимости от практических задач. Возникает вопрос: есть ли необходимость в унификации имеющихся определений и в выработке единого понимания рассматриваемой категории? С одной стороны, кажется, что это позволило бы привести исследования к единой базе, позволяя оперировать равнозначными категориями и упорядочить теоретическую сторону вопроса. С другой стороны, категория NE была выведена и понимается в настоящее время скорее эмпирически, а бóльшая часть исследований в области АОТ идет от практики к теории, понятийный аппарат «подстраивается» под практические задачи исследования, на что авторы книги не раз обращают внимание читателя. В частности, в рассматриваемой работе приводится пример из американской исследовательской программы «ACE 2005», где наряду с такими ИС, как, например, *Peter*, *Charles de Gaulle*, *Andorra*, *M 42* (туманность) и *LDC* (исследовательская лаборатория), в класс NE включаются единицы типа *military helicopter* (военный вертолет) и *land-to-air missile* (зенитная управляемая ракета). Транспорт и оружие в рамках данной программы оказались важными категориями для автоматического «понимания» того класса текстов, который интересовал исследователей и отвечал их практическим задачам, и были включены в типологию NE в рамках именно этой работы (о типологиях см. далее). Унификация и приведение к единообразию определения и характеристик NE, таким образом, кажутся крайне желательными с теоретической точки зрения, однако на практике не всегда представляются возможными.

Далее исследователи сравнивают категорию NE с ИС и дефинициями. Раздел 2.3 посвящен ИС и их традиционным характеристикам в свете практических задач компьютерной лингвистики. Приведем лишь некоторые примеры.

Один из критериев выделения ИС — написание с заглавной буквы, и он был бы очень удобен при автоматическом поиске ИС в тексте, если бы не следующее: этот критерий работает далеко не для всех языков (например, в немецком все существительные пишутся с заглавной буквы, а в грузинском, наоборот, — с маленькой); заглавную букву невозможно «увидеть» при обработке записей устных текстов; существуют имена нарицательные, которые пишутся с заглавной буквы; во многих языках представлено немало дериватов от ИС; многое зависит от графического оформления текста (он может быть написан только заглавными буквами).

С морфосинтаксической точки зрения один из критериев ИС — это отсутствие детерминантов, но и он применим далеко не для всех языков, а в тех языках, где детерминанты употребляются, можно найти немало исключений (например, ИС,

употребляющиеся с артиклем в английском и французском языках). Также авторы возражают против отсутствия у ИС лексического значения: хотя мы и не можем дать лексикографического определения ИС, но его форма имеет как социокультурные маркеры, так и, в некоторых случаях, описательные элементы (например, *Триумфальная арка*). Единичность референта как определяющий критерий тоже оказывается под вопросом — ИС может быть применимо ко многим референтам (так, имя *Анна* носят многие женщины), а имя нарицательное — только к одному (*солнце*).

В разделе 2.4 подробно рассматриваются дефиниции и их классификация, делается указание на важность контекста для понимания подобных единиц при АОТ. Авторы сравнивают два примера: *Президент Республики — это высший пост в исполнительной ветви власти Французской республики* и *Президент Республики будет участвовать в переговорах*, где *президент Республики* (*the President of the Republic*) является дефиницией (*definite description*). Очевидно, что в первом случае «существительное относится к общему понятию, значение которого ясно из значения фразы», а во втором мы видим отсылку «к индивиду, идентифицировать которого можно только с помощью референциальных точек (*referential points*)» (с. 38). Такой точкой может быть указание на место или время проведения переговоров или любая другая информация, которая позволит читателю или машине понять, о каком именно президенте идет речь. При этом и дефиниция, и ИС, и референциальные точки входят в «контейнер» NE. Дефиниция, таким образом, дополняет ИС, тогда как ИС — это «когнитивно экономичный» способ референции. Определить референта дефиниции возможно только в контексте: необходимы либо экстралингвистические знания, либо ситуативный контекст.

В разделе 2.5 рассматривается специфика NE с точки зрения значения и референции. Авторы подчеркивают, что NE не вписываются в перечень классических лингвистических категорий. Выделяются три характеристики NE: уникальность референции (референт может быть только один), ее автономность (для отсылки к референту NE достаточно своих собственных «ресурсов») и «естественная» семантическая и лексическая неоднородность. NE — это «различные лингвистические конструкции, объединенные на базе общих характеристик референта, которые “перерастают” традиционные языковые категории и не могут быть к ним сведены» (с. 46).

В третьей главе рассматриваются инструменты и ресурсы, необходимые для создания и проверки алгоритмов АОТ. Прежде чем находить в тексте NE, необходимо создать их типологию, т. е. выделить категории единиц, которые машина должна будет найти в тексте (это могут быть категории «Человек», «Организация», «Место», «Геополитическая единица» и т. д.; у каждой категории в зависимости от целей исследования могут выделяться подкатегории). В главе приведен краткий обзор типологий, созданных в более ранних исследованиях,

и сравнение разметки текстов в рамках разных типологий (MUC, ACE, ESTER-2, QUAERO).

Еще один инструмент, необходимый для создания работоспособной системы обработки естественного языка, — это аннотированные и неаннотированные корпуса, использующиеся для обучения и дальнейшей проверки системы. В разделе 3.2 рассмотрены основные корпуса английского, французского, немецкого, итальянского и португальского языков.

Типологии, таким образом, задают категории единиц, которые необходимо распознать в тексте, корпуса предоставляют иллюстративный материал, а третий инструмент — лексические базы и базы знаний — обеспечивает систему информацией, «касающейся NE, которую система может использовать в целях распознавания, категоризации и снятия омонимии» (с. 66). Эта информация может быть представлена либо в лексической форме (лексические базы), либо в форме энциклопедической справки (базы знаний). Лексические базы представляют собой список лексических единиц, которые могут иметь референтом ту или иную NE. В базу может быть занесена как полная форма единицы, так и ее часть. Кроме того, в эту базу входят и слова-индикаторы (*indicators*, или *trigger words*), косвенным образом указывающие на наличие и тип находящейся рядом NE, что помогает автоматически классифицировать найденные единицы. Например, сокращение *Mr* является словом-индикатором для категории «Человек» (*Person*), а слово *фестиваль* (*Festival*) указывает на наличие единицы, относящейся к категории «Событие» (*Event*), что может помочь, например, снять омонимию при полном или частичном совпадении названия события с названием места, где оно проводится. В разделе 3.3 рассматриваются такие лексические базы, как ANNIE, «WordNet», «Prolex», «Geonames», «JRC-Names» и др. Здесь же обсуждается вопрос об использовании баз знаний в АОТ. Авторы отмечают, что для обработки текстов могут использоваться базы данных разного типа (изображения, сообщения социальных сетей, сервисы вопросов и ответов), но более детально сосредотачиваются на возможности использования в АОТ таких ресурсов, как «Wikipedia».

В четвертой главе рассматривается следующий этап работы системы: непосредственные механизмы распознавания NE. Практическая цель исследователей — создать программу, способную полностью автоматически выделять NE в текстовом потоке. В разделе 4.2 авторы рассматривают ряд машиночитаемых показателей, которые могут использоваться для распознавания NE:

▪ морфологические параметры — заглавная буква, префиксы и суффиксы (например, *ville*, *saint* — частотные компоненты французских топонимов; антропонимы могут содержать такие форманты, как рус. *-вич*, швед. *-sson*, япон. *-san*; в ближайшем контексте названий организаций могут присутствовать сокращения *Inc.*, *Ltd* или *GmbH*), при этом данные признаки учитываются не только в текстах на «своем» языке, но и при работе с иностранными текстами, ведь в английском

или немецком тексте может появиться название французского города или шведское имя;

▪ совпадения с лексической базой (здесь авторы уделяют особое внимание проблеме снятия омонимии);

▪ контекстуальные признаки (учитываются признаки как в ближайшем контексте, так и в широком контексте всего документа).

Исследователи также достаточно подробно рассматривают различные модели систем по распознаванию NE: статистические и контекстуальные (HMM) модели; машинное обучение; системы, управляемые данными (*data-driven*), и т. д. Общая цель всех этих моделей — связать текст анализируемого документа с базами знаний (такими, как «Wikipedia»). NE, таким образом, не только распознается и классифицируется, но и привязывается к референту.

Эта привязка — следующий этап работы системы, которому посвящена п я т а я г л а в а книги. На данном этапе система должна снять омонимию и установить однозначное соответствие между единицей текста и каким-либо явлением в реальности. Авторы отмечают, что далеко не всегда информация о референте может быть записана в используемую системой базу данных, так как класс ИС — это открытый класс, и составить всеохватывающую лексическую или энциклопедическую базу невозможно в принципе.

На этапе привязки единицы текста к референту из базы знаний необходимо снять омонимию. В разделе 5.2 приводится следующий пример применения контекстуальных признаков: *Буш выступил с речью перед работниками компаний «Google» и «Apple»*. По формальным и лексическим показателям система не сможет понять, идет ли речь о 41-м или 43-м президенте США, однако наличие таких референциальных точек, как *«Google»* и *«Apple»*, однозначно указывает на Буша-младшего, так как во времена его отца этих компаний еще не существовало. Для снятия омонимии можно использовать не только конкретные единицы, но и общий контекст, тематику документа. Этот показатель не сработает в случае с Бушами, поскольку, скорее всего, в обоих случаях речь будет идти о политике, но, например, поможет снять омонимию фамилии *Маркс* (Карл Маркс или Граучо Маркс). Если развести понятия не помогает и широкий контекст, то система может просто привязывать NE к самому частотному референту: например, упоминание столицы Франции более вероятно, чем упоминание городка *Париж* (*Paris*) в Техасе.

В этой же главе приводятся числовые показатели производительности и точности систем распознавания NE (скажем лишь, что эти показатели всего на несколько процентов ниже, чем при обработке текста человеком) и опыт практического применения методики на примере BDpedia Spotlight.

Ш е с т а я г л а в а посвящена оценке качества работы систем по распознаванию NE. Авторы рассматривают классические параметры (точность и полнота), метод подсчета ошибок (*error counts*), метрики для оценки отдельных этапов

работы системы, а также методы оценки предварительной обработки данных (*evaluating preprocessing technologies*).

В заключении исследователи приводят краткий обзор возможностей для практического применения NE (информационный поиск, автоматическое реферирование, сбор информации, машинный перевод, обезличивание персональных данных и т. д.) и перечисляют проблемы, которые еще предстоит решить (анализ мультимедийных и исторических данных, обработка очень коротких сообщений и текстов на других языках и т. п.).

Книга логично структурирована, легко читается, материал иллюстрируется множеством примеров. С точки зрения ономастики работа интересна подходом к материалу и широким взглядом на категорию ИС. Стоит отметить, что авторы так или иначе упоминают множество интереснейших проблем и исследовательских задач, однако дальше простого упоминания они не идут, не указывая возможных или существующих путей их решения (впрочем, это и не является целью работы). Книга Дамьена Нувеля, Мод Эрманн и Софи Россе представляет интерес для специалистов по ономастике и компьютерной лингвистике и является ценным научным вкладом в исследования искусственного интеллекта и автоматической обработки естественного языка.

Рукопись поступила в редакцию 15.01.2018

***

Голикова Дарья Михайловна
аспирант кафедры русского языка, общего языкознания и речевой коммуникации
Уральский федеральный университет
620083, Екатеринбург, пр. Ленина, 51, ком. 306
E-mail: d.golikova@gmail.com

Golikova, Daria Mikhailovna
PhD Student, Department of Russian Language, General Linguistics and Speech Communication
Ural Federal University
51, Lenin av., office 306
620000 Ekaterinburg, Russia
E-mail: d.golikova@gmail.com

Daria M. Golikova
Ural Federal University
Ekaterinburg, Russia

PROPER NAMES AND NAMED ENTITIES RECOGNITION IN THE AUTOMATIC TEXT PROCESSING

Review of the book: Nouvel, D., Ehrmann, M., & Rosset, S. (2016). *Named Entities for Computational Linguistics*. London; Hoboken: ISTE Ltd; John Wiley & Sons, Inc., 2016. 170 p.

The reviewed book by Damien Nouvel, Maud Ehrmann, and Sophie Rosset *Named Entities for Computational Linguistics* deals with automatic processing of texts, written in a natural language, and with named entities recognition, aimed at extracting most important information in these texts. The notion of named entities here extends to the entire set of linguistic units referring to an object. The researchers minutely consider the concept of named entities, juxtaposing this category to that of proper names and comparing their definitions, and describe all the stages of creation and implementation of automatic text annotation algorithms, as well as different ways of evaluating their performance quality. Proper names, in this context, are seen as a particular instance of named entities, one of the typical sources of reference to real objects to be electronically recognized in the text. The book provides a detailed overview and analysis of previous studies in the same field, based mainly on the English language data. It presents instruments and resources required to create and implement the algorithms in question, these may include typologies, knowledge or data bases, and various types of corpora. Theoretical considerations, proposed by the authors, are supported by a significant number of exemplary cases, with algorithms operation principles presented in charts. The reviewed book gives quite a comprehensive picture of modern computational linguistic studies focused on named entities recognition, and indicates some problems which are unresolved as yet.

K e y w o r d s: computational linguistics, proper names, automatic text processing, annotation, named entities, corpus, knowledge base.

Received 15 January 2018