УДК 81'322; 004.934; 004.912 (10.02.19, 05.25.05)

Е. С. Антонов

РАЗРЕШЕНИЕ ОНТОЛОГИЧЕСКОЙ И ЯЗЫКОВОЙ ОМОНИМИИ ИМЕНОВАННЫХ СУЩНОСТЕЙ С ПОМОЩЬЮ ИНТЕРПРЕТАЦИИ ДАННЫХ ОНТОЛОГИИ

Рассматривается проблема разрешения различных видов омонимии именованных сущностей на примере географических объектов в англоязычных новостных сообщениях. Разработанный метод использует интерпретируемые данные из онтологии Freebase в качестве входного вектора для нейронной сети (многослойного персептрона), что позволяет достигнуть показателей точности 95,5 % и полноты 92,9 %.

Ключевые слова: *распознавание именованных сущностей, разрешение омонимии именованных сущностей, нейронные сети, онтологии.*

Для обозначения тех или иных объектов действительности в человеческом языке часто используются имена собственные. Они несут в себе большое количество информации, так как называя всего одно слово, человек способен выразить большое количество признаков, которые помогут другому человеку однозначно идентифицировать объект, о котором идет речь. При современных объемах анализа текстов становится актуальной задача создания компьютерной программы, которая могла бы находить имена собственные в тексте. Эта задача в англоязычной литературе получила название «распознавание именованных сущностей» (Named Entity Recognition).

На первый взгляд может показаться, что большинство сущностей могут быть однозначно определены своим именем. На сегодняшний день, услышав фразу про Владимира Путина, вы едва ли усомнитесь, о ком идет речь. Для человека не составляет особого труда определить, с каким объектом действительности соотносится то или иное имя. Однако с точки зрения составителя базы данных (словаря, онтологии) такая картина выглядит утопией. По мере роста базы данных неоднозначность соответствия между множеством имен и множеством сущностей (онтологическая омонимия) быстро растет. В больших базах данных однозначных соответствий остается крайне мало (рис. 1). Человеческие способности определять сущности во многом объясняются одним простым фактом: любой человек знаком только с очень небольшим кругом сущностей, поэтому в его базе данных доля неоднозначности незначительна.

В качестве иллюстрации онтологической омонимии можно привести имя «Вашингтон»: во-первых, его носил человек (Джордж Вашингтон), во-вторых, такое название имеет 77 населенных пунктов в США, в-третьих, в западной части США есть штат Вашингтон. Но и этот список далек от полного: в некоторых контекстах слово «Вашингтон» употребляется в значении «правительство США»:

U.S. manufacturers are also concerned about ***Washington's*** *hands-off attitude towards the dollar's surge, Utsumi was quoted as saying.*

Картина может стать еще более запутанной, если включить в свое поле зрения корабли, домашних животных и вымышленные миры (фильмы, книги и др.). Можно легко догадаться, что список объектов, соответствующих имени «Вашингтон»,

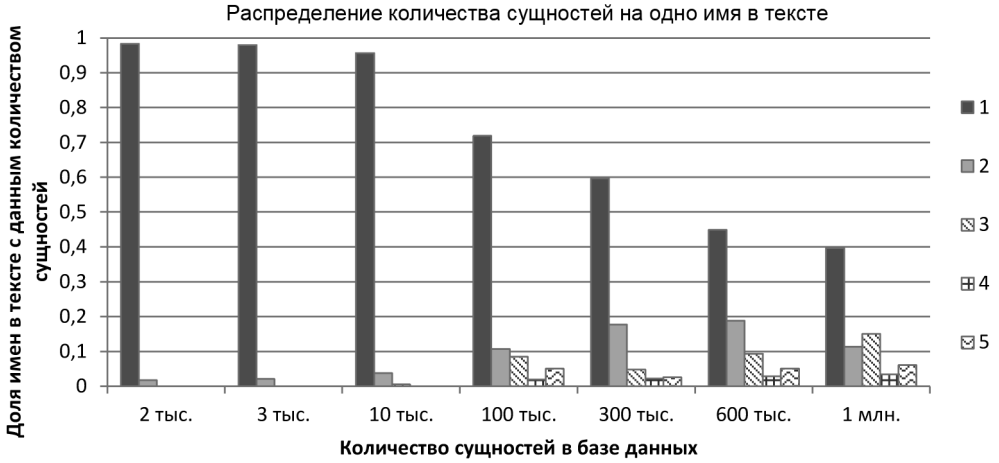

Рис. 1. Возрастание степени неоднозначности имен тексте при увеличении размера БД. Показаны случаи, когда на одно имя претендует до 5 сущностей включительно

можно продолжать буквально до бесконечности. Чем полнее будет наша база данных, тем сложнее будет совершить правильный выбор соотношения «имя ↔ сущность в БД».

Другой проблемой, связанной с увеличением количества сущностей в БД, является *языковая омонимия*, т. е. совпадение имени сущности с общеупотребительным словом языка. В онтологии Freebase некоторые географические объекты имеют имена The, N/A, Bank, Friday, Monday, New, June, March, May и т. д. В некоторых случаях совпадение является следствием ошибки заполнения онтологии, но иногда это может быть «законное» совпадение. График увеличения частоты случаев языковой омонимии представлен на рис. 2.

Проблемы языковой и онтологической омонимии традиционно решаются несколько разными методами. Вопрос о том, должна ли определенному набору слов соответствовать какая-либо онтологическая сущность, решается в рамках задачи распознавания сущностей. Признаки, на основе которых принимается подобное решение, могут иметь разную природу [1]:

• Признаки уровня слова:

– регистр символов;

– наличие в слове пунктуации, цифр, специфичных символов;

– морфологическая информация (падеж, число, лицо, префиксы, суффиксы и т. д.);

– информация о части речи;

– совпадение слова с определенным шаблоном.

• Признаки, заданные списком:

– стоп-слова;

– распространенные аббревиатуры;

– слова, которые принято писать с заглавной буквы;

– наличие в списке известных сущностей;

– типичные суффиксы организаций, географических объектов, персон и т. п.

• Признаки документа/корпуса:

– упоминание в контексте других сущностей;

– вхождения с первым символом в верхнем/нижнем регистре;

– анафора;

– определенная позиция в тексте (заголовок, подпись и т. п.);

– метаинформация документа/корпуса (URL, заголовки, списки, таблицы);

– частота имени в корпусе;

– информация о совместной встречаемости двух сущностей.

Проблема разрешения онтологической омонимии исследована хуже. На сегодняшний день опробовано две разновидности методов. Первая разновидность – модель «мешка слов» (bag of words) – использует данные о совместной встречаемости онтологической сущности (или ее категории) и определенного слова (словосочетания) [2–4]. Наилучший результат при использовании этого метода был достигнут в работе [4] (точность 91,46 %). Некоторые авторы критикуют модель «мешка слов», справедливо замечая, что ее способность к обобщению невелика [5, с. 209]. В этой и некоторых других работах [6–8] используется другая разновидность методов разрешения онтологической омонимии – интерпретация графа связей между сущностями в онтологии. Авторы работы [5] сравнивают качество кластеризации по модели «мешка слов» (F-мера 78 %), ссылкам в социальных сетях и другим мерам сходства. Наилучший результат достигается при использовании информации о совместной встречаемости всех пар имен двух сущностей (F-мера 88 %). Качество работы алгоритма, представленного в этой статье, можно сравнивать с аналогичными работами, где использовались комбинированные методы и качество достигало 90–91 % [9, 10].

В данной работе предпринята попытка использовать интерпретированные данные из онтологии

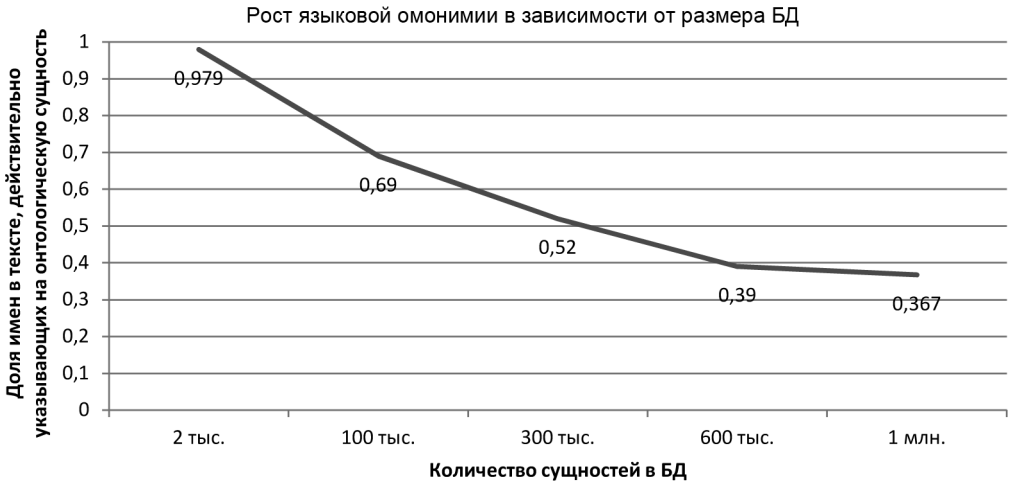

Рис. 2. Рост частоты «ложных» имен в тексте в зависимости от размера БД

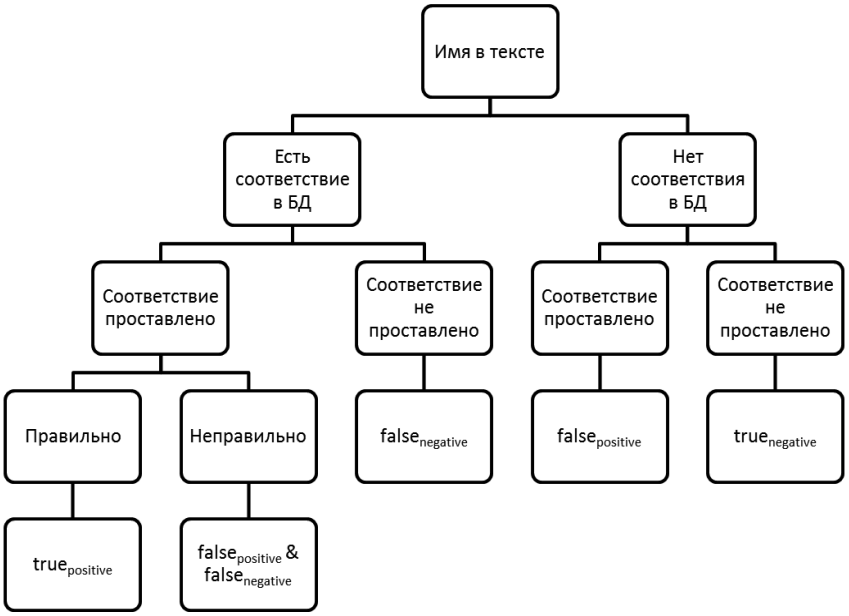

Рис. 3. Оценка работы алгоритма разрешения онтологической омонимии

Freebase для разрешения и языковой, и онтологической омонимии. Для проверки качества работы алгоритма был составлен тестовый корпус из 1700 размеченных новостных статей, где для каждого потенциального имени был указан верный вариант записи в БД, если таковой был (всего около 32 тыс. случаев). Качество работы алгоритма оценивалось по четырем параметрам (рис. 3).

На основе этих четырех параметров строилась оценка точности (precision), полноты (recall) и F-мера.

$$precision = \frac{true_{positive}}{true_{positive} + false_{positive}}$$

$$recall = \frac{true_{positive}}{true_{positive} + false_{negative}}$$

$$F1 = 2 * \frac{precision * recall}{precision + recall}$$

Задачу алгоритма разрешения языковой и онтологической омонимии можно свести к приписыванию определенного веса каждому соответствию между именем в тексте и онтологической сущностью. Чтобы решить проблему онтологической омонимии, нужно добиться того, чтобы наш алгоритм приписывал правильной сущности наибольший вес. Для решения проблемы языковой омонимии каждая сущность должна пройти некоторый «порог активации», т. е. приписанный вес должен быть больше некоторого числа.

Онтология Freebase содержит разнообразные данные о географических объектах. Ключевым моментом нашего метода является интерпретация этих данных, т. е. для каждого пласта информации строилась некоторая функция, которая в зависимости от значения конкретных атрибутов приписывала каждой сущности тот или иной вес. Всего было написано 36 правил интерпретации, что в результате дало вектор из 62 чисел для каждой онтологической сущности. Эти правила можно разделить на несколько категорий:

1. Принадлежность сущности к определенному онтологическому типу (город, штат, страна, континент и т. п.).

2. Использование графа административной вложенности географических объектов (поиск более крупного объекта (штата, страны), соседей по графу (города в одной стране)).

3. Кластеризация по географическим координатам с разным радиусом.

4. Различные свойства имени (определенные суффиксы, регистр написания, частотные характеристики в корпусе, похожесть на английское слово по частоте n-грамм).

5. Правила цепочки слов с прописной буквы и слов в верхнем регистре (количество таких слов слева и справа от имени).

6. Данные о размере географического объекта (население, площадь).

7. Нахождение имени в определенном сегменте текста (заголовок, начало текста, подпись, конец текста, таблица).

8. Прочие правила (является ли сущность столицей, имеется ли в тексте неомонимичное упоминание той же сущности, длина и т. п.).

На основе множества получившихся векторов сущности был обучен многослойный персептрон. Поступающие на вход данные проходили определенную процедуру нормализации (отсечение ста-

тистических выбросов и преобразование чисел в отрезок [–1; 1]). Подбор архитектуры сети (три промежуточных слоя по 22 персептрона) и алгоритма обучения (RPROP [11]) позволили сократить время обучения одной сети до 20–40 мин, большую часть из которых занимал процесс кросс-валидации. Выходной персептрон в итоге имеет значение в промежутке [–1; 1].

После обучения персептрона подбирался «порог активации» сущности. Для этого итеративно прогонялись числа из промежутка [–1; 1] с шагом в 0,001, при этом на каждом шаге замерялось значение целевой метрики на тестовом корпусе (рис. 4).

Параметры, по которым считается целевая метрика, не предполагают однозначной обратной корреляции между точностью и полнотой, однако, как можно заметить из рис. 4, она существует (впрочем, не вполне строгая). В случае с нашим обучающим корпусом удалось достигнуть F-меры 94,2 % (точность 95,5 %, полнота 92,9 %), однако алгоритм довольно гибок в выборе приоритетной характеристики, будь то точность или полнота.

Таким образом, модель с использованием интерпретации онтологических данных оказалась

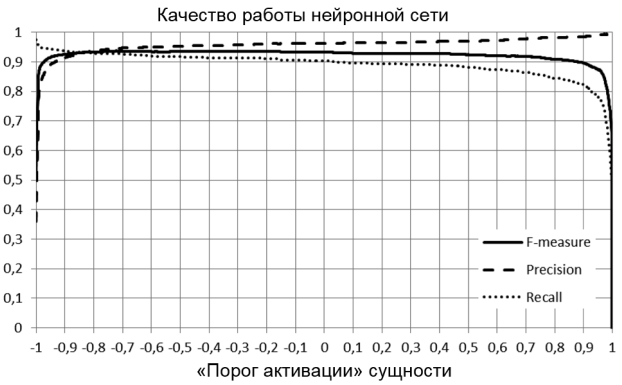

Рис. 4. Подбор «порога активации» сущности для достижения оптимального значения целевой метрики

состоятельной и позволила достичь лучших характеристик, нежели комбинация модели «мешка слов» и использования данных о связях между сущностями [9, 10]. Стоит заметить, что в нашей системе использовался на порядок больший объем БД (чуть менее миллиона сущностей). Это позволяет предполагать, что представленный алгоритм имеет лучшую обобщающую силу.

Список литературы

1. Nadeau D., Sekine S. A survey of named entity recognition and classification // Linguisticae Investigationes. Amsterdam, Netherlands: John Benjamins Publishing Company, 2007. Vol. 30. Iss. 1. P. 3–26.
2. Bunescu R., Paska M. Using Encyclopedic Knowledge for Named Entity Disambiguation // Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006). Stroudsburg, PA, USA: Association for Computational Linguistics, 2006. P. 9–16.
3. Fader A., Soderland S., Etzioni O. Scaling Wikipedia-based Named Entity Disambiguation to Arbitrary Web Text // Proceedings of the WikiAI 09 – IJCAI Workshop: User Contributed Knowledge and Artificial Intelligence: An Evolving Synergy. Pasadena, CA, USA: IJCAI Organization, 2009.
4. Gentile A. L., Zhang Z., Xia L., Iria J. Cultural Knowledge for Named Entity Disambiguation: A Graph-Based Semantic Relatedness Approach // Serdica Journal of Computing. Sofia, Bulgaria: Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, 2010. Vol. 2. Iss. 4. P. 217–242.
5. Han X., Zhao J. Named Entity Disambiguation by Leveraging Wikipedia Semantic Knowledge // Proceedings of the 18th ACM conference on Information and knowledge management. New York, NY, USA: Association for Computing Machinery, 2009. P. 215–224.
6. Malin B., Arnoldi E., Kathleen M. C. A Network Analysis Model for Disambiguation of Names in Lists // Computational & Mathematical Organization Theory. Hingham, MA, USA: Kluwer Academic Publishers, 2005. Vol. 11. Iss. 4. P. 119–139.
7. Minkov E., Cohen W. W., Ng A. Y. Contextual Search and Name Disambiguation in Email Using Graphs // Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. New York, NY, USA: Association for Computing Machinery, 2006. P. 27–34.
8. Bekkerman R., McCallum A. Disambiguating Web Appearances of People in a Social Network // Proceedings of International World Wide Web Conference. New York, NY, USA: Association for Computing Machinery, 2005. P. 463–470.
9. Kanada Y. A method of geographical name extraction from Japanese text // Proceedings of the eighth international conference on Information and knowledge management (CIKM'99). New York, NY, USA: Association for Computing Machinery, 1999. P. 46–54.
10. Smith D. A., Crane G. Disambiguating geographic names in a historic digital library // Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries. London, UK: Springer, 2001. P. 127–136.
11. Riedmiller M., Braun H. RPROP - A Fast Adaptive Learning Algorithm // Proceedings of the International Symposium on Computer and Information Science VII. Heidelberg, Germany: Springer, 1992.

Антонов Е. С., аспирант.
Московский государственный университет им. М. В. Ломоносова.
Ленинские горы, МГУ, д. 1, стр. 51, Москва, ГСП-1, 119991
E-mail: me@eantonov.name

Материал поступил в редакцию 13.09.2012.

E. S. Antonov

NAMED ENTITY DISAMBIGUATION BASED ON ONTOLOGY DATA INTERPRETATION

The article deals with the issue of named entity disambiguation (geographical objects in English news). The research is based on the method of ontology data interpretation. The method uses features, extracted from Freebase ontology, as the input for a neural network (multi-layer perceptron). This strategy makes possible to achieve precision value of 95.5 % and recall 92.9 %.

Key words: *named entity recognition, named entity disambiguation, neural networks, ontology.*

Lomonosov Moscow State University.
MSU, Faculty of Philology, Russia, 119991, Moscow, GSP-1, 1-51 Leninskie Gory, 1 Humanities Building.
E-mail: me@eantonov.name