Секция «Вычислительная математика и кибернетика»

Атрибуция объектов текста на основе методов машинного обучения.

Драль Алексей Александрович
Аспирант
Московский государственный университет имени М.В. Ломоносова,
Механико-математический факультет, Москва, Россия
E-mail: aadral@gmail.com

"Named Entity Recognition"(известное в зарубежной литературе также, как "entity identification"и "entity extraction") занимается поиском и классификацией простых термов в тексте в заранее определенной категории, например - имена личностей, организаций, географические местоположение, время и т.п. Большинство зарубежных информационных систем выделения именованных сущностей структурируют не размеченные блоки текста на основе одной из двух иерархий: иерархии Секина [6] и иерархии от компании BBN Technologies [5] (обе изобретены в 2002 году). Современные системы для работы с английским языком обрабатывают информацию практически с человеческой точностью. Например, лучшая информационная система, представленная на 7-й конференции MUC-7 [1], имеет качество обработки информации равное 93.39% по F-мере. В то время как специальные люди, по некоторым оценкам [7], могут разметить тексты с качеством - 97.60%.

Информационные системы выделения именованных сущностей основаны на двух подходах: лингвистические правила, основанные на формальной грамматике, и статистические модели. В работе делается обзор инструментария для обработки данных, а также общедоступных сервисов, как первого (DBPedia, TrueKnowledge), так и второго типов (KnowItAll [3], YAGO-NAGO [2]). В данной работе исследуется другой подход на основе применения статистических методов машинного обучения. Особенность исследования - использование псевдо-параллельных текстов для эффективного уточнения алгоритмов машинного обучения. Существует русскоязычных сервис, использующий автоматическое извлечение данных из текста, разработанный компанией Yandex - пресс-портреты [4]. Данные о качестве обработки не разглашаются, поэтому провести сравнение алгоритмов не представляется возможным.

Выделение текстовых объектов, прежде всего именованных сущностей, а также отношений между ними, является актуальной задачей при информационном мониторинге. Именно выделение устойчивых предикативных отношений для текстовых объектов, является логическим продолжением представленной работы.

Литература

1. Elaine Marsh, Dennis Perzanowski, "MUC-7 Evaluation of IE Technology: Overview of Results 29 April 1998

2. Gerhard Weikum, Martin Theobald, "From Information to Knowledge: Harvesting Entities and Relationships from Web Sources 6 June 2010

3. Oren Etzoni, Michael Cafarella, Doug Downey, Ana-Maria Popescu Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates, "Unsupervised Named-Entity Extraction from the Web: An Expretimental Study 28 February 2005

4. доклад Татьяны Ландо в рамках NLP семинара 3 апреля 2010: http://mathlingvo.ru/nlpsemi
5. иерархия именованных сущностей от BBN Technologies: http://www.ldc.upenn.edu/Catalog/Types-Subtypes.html
6. расширенная иерархия Секина: http://nlp.cs.nyu.edu/ene/
7. сайт конференции MUC: http://www.itl.nist.gov/iaui/894.02/related_projects/muc/index.htm

Слова благодарности

Выражаю благодарность своему научному руководителю Доброву Борису Викторовичу за постановку задачи и помощь в подготовке доклада.