УДК 81'322.2

Сравнение результатов кластеризации слов по Брауну

с семантическим деревом слов WordNet

Г.Н. Чепков[1]

[1]Московский физико-технический институт (государственный университет)

chepkov@phystech.edu

«Центральной» задачей обработки естественного языка (NLP) можно считать задачу понимания текста. Однако такую задачу трудно поставить формально – и по крайней мере столь же трудно решить. На подступах к решению этой задачи находится множество легче определяемых и более обозримых задач – например, задачи правильной сегментации текста (chunking), отношений между словами (syntacticparsing), нахождения названий, собственных имен (namedentityrecognition) и отношений между ними (factextraction).Однако же, оказывается (и вряд ли это должно быть сюрпризом), что понимание (и в первую очередь, понимание значений слов) может принести и в решении этих частных задач огромную пользу – если только найти для выражения этого понимания способ – и в частности, как можно адекватнее выразить семантические отношения между словами.

Разумеется, лингвистическая наука накопила множество знаний об этом. Но практика показывает, что их перевод в алгоритмически полезную форму дорог и непрост, а результат, если и достижим, весьма далек от полного покрытия встречающегося словаря. Полнота покрытия и полнота учета имеющегося материала практически достижима только при использовании корпуса «сырых» - не размеченных человеком – текстов. Идея подобного использования состоит, в частности, в том, чтобы заменить человеческое понимание слов информацией о тех контекстах, в которых они встречаются. В последние 2 года доступен для публичного использования разработанный Google инструмент word2vec – двухуровневая нейронная сеть, позволяющаяпредставить семантические отношения слов в заданном корпусе текстов путем вложения их в много- (несколько сот) мерное векторное пространство; онаэффективно тренируется на задаче предсказания вероятностей вхождений слов в тексте на основедругого вблизи стоящегослова [1]. На языке векторного пространства оказывается возможным выразить различные типы семантических отношений между словами. Тем не менее, для более ясного понимания характера этого представления требуетсясистематическое сравнение его результатов(как

количественное, так и качественное) с традиционными лингвистическими инструментами изучения этих отношений в данном корпусе.

Одним из наиболее изученных типов семантических отношений является семантическая близость слов и понятий, в частности, выраженная в отношениях гипонимии и гиперонимии. Конечно же, близость точек в векторном пространстве позволяет проводитькластеризацию точек, в том числе иерархическую. Кроме того, существуют и более прямые методы автоматической иерархической кластеризации слов на основе их контекстного окружения на материале заданного корпуса текстов. В частности, аггломеративныйалгоритм кластеризации слов по Брауну (Brownclustering) [2] строит бинарное дерево, листья которого – слова, а узлы – кластеры, содержащие все слова в ветви, на основе максимизации правдоподобия прогноза вероятностей соседних пар слов(при условии представления слов их кластерами) при фиксированном числе кластеров. Дерево в результате оказывается построено так, что расположенные недалеко друг от друга слова – семантически сходны. Это – алгоритм жесткой кластеризации, каждому слову отвечает единственный кластер. Стоит заметить, что представление слов их кластерами помогает, в частности, учесть вклад редко встречающихся слов (sparsedata), на основе их семантической связи с более «частыми» словами.

В настоящей работе проводится сравнение результатов кластеризации слов по Брауну на основе различных корпусов текстов с ручным представлением информации, выполненным в результате многолетней работы профессиональных лингвистов. Качествокластеризации данного алгоритма сравнивается с кластеризацией, задаваемой семантическим деревом(гипонимии/гиперонимии) WordNet [3]. Это обширная лексическая база данных английского языка, разработанная учеными Принстонского университета вручную. Количественный критерий для сравнения – правдоподобие прогноза вероятностей соседних пар слов (на тестовом наборе текстов) при представлении слов их кластерами (на заданном уровне иерархии WordNet). Кроме того, изучается и качественное соответствие кластеров, структурих иерархии, а также зависимость результатов от объема корпуса.

Литература

1. *Mikolov T., Chen K., Corrado G., Dean J*. Efficient Estimation of Word Representations in Vector Space //International Conference on Learning Representations (ICLR). 2013.(http://arxiv.org/abs/1301.3781)
2. *Liang P.* Semi-supervised learning for natural language processing // Thesis (M. Eng.), MIT. 2005. (http://cs.stanford.edu/~pliang/papers/meng-thesis.pdf)

3. *Miller G.A., Beckwith R., Fellbaum C.D., Gross D., Miller, K.* WordNet: An online lexical database // Int. J. Lexicograph. v.3, 4, pp. 235–244. 1990. (http://wordnet.princeton.edu/wordnet/)