Применение нейросетевых методов к задаче разрешения кореференции

Куратов Юрий Михайлович

Аспирант

Факультет ИВТ МФТИ (ГУ), Долгопрудный, Россия

E-mail: yurii.kuratov@phystech.edu

Разрешение кореференции является одной из нерешенных задач обработки естественного языка, и заключается в определении того, какие сущности (упоминания) в тексте обозначают одни и те же объекты:

> Николай Петрович познакомился с Фенечкой следующим образом. {...} «Уж не немка ли здесь хозяйка?» — пришло ему на мысль; но хозяйкой оказалась русская, женщина лет пятидесяти, опрятно одетая, с благообразным умным лицом и степенною речью. Он разговорился с ней за чаем; очень она ему понравилась.[1]

Умение извлекать подобную информацию из текстов востребовано в широком спектре задач обработки естественного языка таких, как извлечение фактов и информации из текстов, машинный перевод, диалоговые и вопросно-ответные системы. Данная задача вызывает повышенный интерес и тем, что существующие решения на данный момент обладают недостаточно высоким качеством [3]. Также если для английского языка существуют корпуса (CoNLL-2012[2]) достаточно больших размеров, то для русского языка ситуация усугубляется наличием наборов данных на порядок меньших объемов (RuCor[3] [3]) .

Решаемую задачу можно формализовать следующим образом: необходимо кластеризовать упоминания в тексте так, чтобы в одном кластере находились упоминания указывающие на один объект. Упоминания считаются уже извлеченными из текста. Идея предлагаемого решения заключается в том, чтобы научиться для пары упоминаний предсказывать вероятность того, что они лежат в одном кластере. Такой локальный подход к задаче разрешения кореференции требует построения модели с высокой точностью определения

[1] Тургенев И. С. Отцы и дети.

[2] http://conll.cemantix.org/2012/

[3] http://rucoref.maimbava.net/

наличия связей, так как дальнейшее построение кластеров чувствительно к возможным ошибкам. В качестве модели, предсказывающей наличие связи, использовалась нейронная сеть, состоящая из трех полносвязных слоев (256 - 128 - 1) с сигмоидальной функцией активации на выходе. Для построения кластеров упоминаний использовался алгоритм агломеративной кластеризации, расстоянием между упоминаниями являлась предсказанная нейронной сетью вероятность наличия связи.

Данное нейросетевое решение отличается от предложенных ранее [2], [4] тем, что использует минимальное количество дополнительных признаков и основывается на векторных представлениях слов. Нейросетевая модель показывает лучшее качество по сравнению с опубликованными результатами [2], [4] по всем метрикам, принятым для измерения качества систем разрешения кореференции (*CoNLL-F-1*, MUC, B^3 [1]). Эксперименты проводились на корпусе RuCor.

Литература

1. Luo X. On coreference resolution performance metrics //Proceedings of the conference on human language technology and empirical methods in natural language processing. – Association for Computational Linguistics, 2005. – С. 25-32.
2. Sysoev, A. A. et al. Coreference Resolution in Russian: State-of-the-art approaches application and evolvement. //Computational Linguistics and Intellectual Technologies. International Conference “Dialogue 2017” Proceedings, 2017. – С. 317-338.
3. Toldova S. et al. RU-EVAL-2014: Evaluating anaphora and coreference resolution for Russian //Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue”. – 2014. – С. 681-694.
4. Toldova S., Maxim I. Coreference Resolution for Russian: The Impact of Semantic Features //Computational Linguistics and Intellectual Technologies. International Conference “Dialogue 2017” Proceedings, 2017. – С. 339-349.