Исследование зависимости качества распознавания онтологических объектов от глубины гипонимии.*

Дочкина В., Кузнецов М., Резяпкин В., Русскин А., Ярмошик Д.
yarmoshik.dv@phystech.edu

Целью работы было проведение эксперимента по распознаванию онтологических объектов на различных уровнях гипонимии моделями, использующимися для распознавания именованных сущностей. Для этого с помощью данных WordNet впервые был создан датасет, в котором для текстов на английском языке несколькими вариантами классифицированы онтологические объекты: в каждом отдельном варианте разметки используются теги «аналогичных уровней» иерархии гипонимии. Исследована связь качества распознавания с уровнем иерархии гипонимии state-of-the-art алгоритмом.

Ключевые слова: *Natural language processing (NLP), named entity recognition (NER).*

1 Введение

В последние годы прогресс в области компьютерной обработки естественного языка (Natural Language Processing) вывел эту область в число наиболее успешно развивающихся подразделов машинного обучению и искусственного интеллекта. Были достигнуты выдающиеся успехи в решении многих задач NLP, в том числе задач компьютерного перевода ([1], [2]), определения эмоциональной окраски текстов [3] и генерации изображений по текстовому описанию [4].

Одной из наиболее востребованных задач NLP является задача распознавания именованных сущностей (Named Entity Recognition). В этой задаче необходимо сопоставить слова (группы слов) обрабатываемого текста заранее определённому набору тегов (именованных сущностей). Например, в качестве тегов могут быть выбраны [Person, Time, Location, Organization]. Тогда текст «*Jim bought 300 shares of Acme Corp. in 2006*» должен быть размечен следующим образом: « $[Jim]_{Person}$ *bought 300 shares of* $[Acme\ Corp.]_{Organization}$ *in* $[2006]_{Time}$ ».

Введём определения основных понятий. *Именованные сущности* - это ? *Онтологические объекты* - это ? (И то, и то - набор заранее выбранных слов-тегов. Разница?)

Наша работа основывается на том, что онтологические объекты, такие как теги для задачи NER, образовывают иерархическую структуру по степени общности понятия. Возьмём в качестве примера слово Person. Онтологические объекты, имеющие более общее значение по отношению к нему, например, Physical object, называются гиперонимами. Онтологические объекты, выражающие по отношению к нему более частную сущность, например, Male person, называются гипонимами.

Мы ставили перед собой задачу обучить модель, близкую по архитектуре к современным NER моделям, распознавать онтологические сущности на разных уровнях гипонимии. Для этого мы с помощью базы WordNet несколькими способами произвели автоматическую разметку онтологических объектов, аналогичную разметке текстов в задаче NER, так, чтобы в каждом отдельном варианте разметки использовались теги, «близкие» друг к другу по уровню иерархи гипонимии в смысле, изложенном в разделе **??**.

*Работа выполнена при финансовой поддержке РФФИ, проект №00-00-00000. Научный руководитель: Стрижов В. В. Задачу поставил: Эксперт И. О. Консультант: Консультант И. О.

Такая, на первый взгляд, неразумная задача, позволяет нам достичь двух целей. Во-первых, мы получаем информацию о зависимости качества распознавания онтологических объектов от глубины гипонимии, представляющую теоретический (?) интерес. Во-вторых, мы ожидаем, что наша модель научится распознавать именованные сущности, несмотря на то, что в нашем датасете, размеченном с помощью словаря, им сопоставлены неправильные теги. Это весьма вероятно, потому как имена собственные и другие нетривиально определяемые именованные сущности составляют малую часть онтологических объектов в датасете и с точки зрения модели являются незначительной «грязью» в данных, которая, как показывают последние исследования [?], не является серьёзной помехой для обучения моделей, основанных на нейросетях.

2 Обзор литературы

Решения задачи распознавания именованных сущностей совершенствуются уже несколько десятилетий. На протяжении этого периода происходит постепенный переход от моделей, основанных на скрупулёзно конструируемых «руками» правилах, к системам, основанным на «обучении с учителем», конструируемым независимо от языка. Одними из первых обучаемых моделей были модели основанные на вероятностных моделях таких как HMM и CRF. С теми или иными модификациями в начале 2000-х эти методы пытались применять в различных областях, например, в медицине, где, впрочем, даже наиболее успешные работы показывали достаточно низкие результаты [5]. Тем не менее, как оказалось, даже с помощью чистого CRF при использовании оптимизационных алгоритмов, учитывающих удалённый контекст, можно получить неплохие результаты [6]. Прогресс в технологии обучения искусственных нейронных сетей позволил успешно применить нейросетевые архитектуры к задаче NER, что существенно улучшило качество распознавания именованных сущностей современными алгоритмами. При этом стоит отметить,что независимо от используемого подхода, лучший результат дают модели, адаптированные для входных данных, особенно при высокой специфичности обрабатываемых текстов, например, в твитах [7]. Одной из сильных сторон нейросетевого подхода является возможность посимвольной обработки слов [8] для получения признаков, которые могут использоваться другими алгоритмами - элементами сложной модели. Так, самые эффективные современные модели совмещают нейросетевой и вероятностный подходы [9]

3 Постановка задачи

3.1 Схема алгоритма

На первом этапе выбирается несколько корневых сущностей $\mathbf{E_0} = \{e_{01}, \ldots, e_{0n}\}$ из верхних уровней иерархии гипонимии (подробнее см **??**) С помощью WordNet для каждой сущности e из $\mathbf{E_0}$ строится дерево гипонимов с корнем в e. По методу, описанному в разделе **??**, строится соответствие между уровнями глубины вершин каждого отдельного дерева и общей шкалой глубины гипонимии. Далее, для каждого уровня гипонимии k выбирается множество сущностей $\mathbf{E_k}$, которые будут использованы для распознавания.

Следующий этап – разметка датасета. Для каждого уровня гипонимии k выполняется следующее: для каждого слова w в датасете ищется сущность $e \in \mathbf{E_0}$, такая, что w принадлежит поддереву с корнем e, и слово w помечается тегом e. Если такой сущности не находится, w помечается тегом *out-of-tag*.

Этап обучения. Для распознавания используется NER-модель ...

3.2 Выбор онтологических объектов

Литература

[1] Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. Google's multilingual neural machine translation system: enabling zero-shot translation. *arXiv preprint arXiv:1611.04558*, 2016.

[2] Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*, 2017.

[3] Sara Rosenthal, Noura Farra, and Preslav Nakov. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, 2017.

[4] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. *arXiv preprint*, 2017.

[5] Burr Settles. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications*, JNLPBA '04, pages 104–107, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[6] Flavio Massimiliano Cecchini and Elisabetta Fersini. Named entity recognition using conditional random fields with non-local relational constraints, 2013.

[7] Alan Ritter, Sam Clark, Oren Etzioni, et al. Named entity recognition in tweets: an experimental study. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1524–1534. Association for Computational Linguistics, 2011.

[8] Emily Sheng and Prem Natarajan. A byte-sized approach to named entity recognition, 2018.

[9] L. T. Anh, M. Y. Arkhipov, and M. S. Burtsev. Application of a hybrid bi-lstm-crf model to the task of russian named entity recognition, 2017.