В последние годы прогресс в области компьютерной обработки естественного языка (Natural Language Processing) вывел эту область в число наиболее успешно развиваваищихся 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 неоходимо 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* $[2006]_{Time}$ ».

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 чатсную 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 задлаче NER, так, чтобы в каждом отедльном варианте разметки использовались теги, «близкие» друг к другу по уровню иерархи гипонимии в смысле, изложенном в разделе $$$$.

Такая, на первый взгяд, неразумная задача, позволяет нам достичь двух целей. Во-первых, мы получаем информацию о зависимости качества распознавания именованных сущностей от глубины гипонимии, представляющую теоретичский (?) 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.

Список литературы

[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.