*Александр Орлов, БКЛ-162*

**Эссе по статье Т. О. Шавриной «World Vector Models as an object of linguistic research»**

Эта статья является одной из первых статей, в которых word2vec модели, раньше воспринимавшиеся компьютерными лингвистами только как один из инструментов исследования семантики слов и их сочетаемости, выступают в качестве самостоятельного объекта исследования. Автор отмечает, что векторные модели достойны изучения, так как они представляют собой информацию о дистрибуции слова в языке, собранную на большом числе контекстов за большой период времени[[1]](#footnote-1). Основная задача автора – описать особенности поведения векторных моделей для рассматриваемых языков. Этот материал может пригодиться при дальнейшей разработке механизмов оценки векторных моделей[[2]](#footnote-2). В качестве предмета исследования в статье выступают свойства рассмотренных моделей. На них стоит остановится немного подробнее.

В статье рассматриваются несколько word2vec моделей для русского и английского языков. Русские модели во всех представленных экспериментах тренировались на новостном корпусе, НКРЯ+Википедии, Taiga и Aranea. Английские моделибыли взяты из WebVectors. В статье излагается ряд экспериментов с моделями, направленных на изучение их поведения.

В первом эксперименте на материалах русского и английского языков сравнивалось число постоянных соседей (stored neigbours)[[3]](#footnote-3) для слов из списка Сводеша, случайных слов языка с высокой частотностью (top 2000) и просто случайных слов языка. Эксперимент показал, что наиболее высокий показатель наблюдался у слов из списка Сводеша. Затем шли частотные высокослучайные слова, а после этого – просто случайные слова. В этом эксперименте автор показал, что слова из списка Сводеша наименее чувствительны к смене модели как в русском, так и в английском языке.

Второй эксперимент уже был проведён на всем объёме языка. Автор получил пересечение по вокабуляру для всех моделей. Затем для каждого слова было получено число стабильных соседей и список слов был отсортирован по проценту соседей, которые остаются одинаковыми независимо от модели. В результате были получены слова с наиболее стабильными соседями. Большинство из них можно отнести к следующим классам: качества человека, эмоции, национальности, профессии, топонимы, прилагательные меры. Слова этих классов являются самыми стабильными как для русского, так и для английского. Самыми нестабильными для обоих языков оказались имена собственные (по-видимому в статье имелись в виду только имена людей, но однозначно сказать сложно).

В третьем эксперименте автор посчитал то, насколько хорошо сохраняются отношения между онтологически близкими словами (синонимами, антонимами, гипперонимами и т. п.) для моделей, построенных на новостном корпусе, Aranea, НКРЯ+wiki и Taiga. Автор смотрел на сохранение связей в моделях между онтологическими парами, выделенные в [Loukachevitch, Lashevich, 2016][[4]](#footnote-4) в окнах 10, 20, 50, 100. Было выявлено, что синонимические и антонимические связи, так часто демонстрируемые нам на различных примерах из статей, для большей части слов не воссоздаются. Наилучший результат показал корпус Aranea (сохранено 48% антонимов, 41% - причин (cause), 20% - гипперонимов, голоноимов (part holonyms) и меронимов (part meronyms), в то время как НКРЯ+wiki показал результат ниже среднего. Автор делает вывод, что те объёмы данных, которыми мы можем располагать на данный момент слишком малы для того, чтобы строить достаточно точные модели, а способности моделей по предсказанию онтологических отношений преувеличены.

В статье были выделены некоторые особенности поведения векторных моделей, однако, исследование нельзя назвать исчерпывающим.

1. Это, конечно же, верно только для моделей, которые тренировались на больших корпусах, включающих в себя несколько временных периодов. [↑](#footnote-ref-1)
2. В статье приведён краткий обзор литературы, из которого становится ясно, что немотря на актуальную на данный момент в NLP тенденцию к поиску наиболее эффективного способа вектаризации слов, в литературе до сих пор не было представлено качественного и подходящего для всех случаев механизма оценки полученных вектороввекторов. [↑](#footnote-ref-2)
3. слов, которые всегда были соседями слова (в рамках заданного окна – 10, 20, 50 или 100), независимо от калькуляции модели (всего для каждого язык было создано 15 моделей, в которых сравнивалась дистрибуция слов) [↑](#footnote-ref-3)
4. Loukachevitch N., Lashevich G. (2016) Multiword expressions in Russian Thesauri RuThes and RuWordNet. Proceedings of the AINL FRUCT 2016, pp. 66–71. [↑](#footnote-ref-4)