Разработка метода кластеризации слов по смысловым характеристикам с использованием алгоритмов Word2Vec

Левченко Софья Вадимовна, Кирилловых Андрей Максимович, Гимашев Линар Русланович

Студентка, студент, студент

Департамент компьютерной инженерии МИЭМ имени А. Н. Тихонова НИУ ВШЭ, Москва, Россия

E-mail: sonyalevchenko@mial.ru, andykirill@gmail.com, linarkins@gmail.com

При компьютерной обработке текста на естественном языке происходит его формальное представление, и присутствие многозначных конструкций затрудняет машинный анализ. Целью разработки метода кластеризации слов является устранение морфологической неоднозначности при компьютерном анализе. Задачей программы является детектирование-классификация текстов, построенных на сдвиге понятий или игре слов с помощью возможностей инструмента Word2Vec от Google.

Одной из основных функций многозначных слов речи является языковая игра, построенная на многозначных словах, контекстной синонимизации, использовании фразеологизмов. Чаще всего комический эффект достигается резким сдвигом, переключением рассказа из одного семантического поля в другое [1]. Пример 1:

«Штирлиц смотрит — из форточки дуло. Он закрыл форточку, и дуло исчезло.» Слово «дуло» — омоформа: сначала это глагол (ср.р. от «дуть»), затем происходит семантический сдвиг — под словом «дуло» подразумевается существительное (ср.р. им. п.).

Аналогичные краткие юмористические тексты, каламбур которых построен на сдвиге понятий или игре слов, и были объектом анализа.

Word2Vec — это набор алгоритмов обработки нейросети прямого распространения для расчета векторных представлений слов [2]. Принцип работы состоит в нахождении связей между контекстами слов, ведь слова, находящиеся в похожих контекстах, часто могут быть семантически близкими. То есть нужно максимизировать косинусную близость между векторами слов, появляющихся в близких контекстах, и минимизировать косинусную близость слов, не появляющихся в контексте друг друга [3].

Основная идея предлагаемого метода заключается в том, что для детектирования компьютером шутки нужно разрешить омонимию,

ведь она построены на сдвиге понятий. Для каламбура см. Пример 1 будет достаточен анализ слова «дуло» в контекстах «из окна дуло» и «дуло исчезло». Для каждого интересующего нас омонима можно получить вектор семантически близких к нему слов как без учета контекста, так и анализируя вокруг стоящие слова.

Для обучения Word2Vec русскому языку были использованы текстовые модели сервиса RusVectores [4]. Затем, для подключения алгоритмов Word2Vec был взят ранее разработанный модуль снятия омонимии с текстов на естественном языке, который определяет для каждого слова соответствующий ему «инфинитив», что и требуется. Используя триграммный анализ, производится морфологический разбор поступившего предложения — каждому слову присваивается свой кортеж лексических параметров. В случае нахождения неоднозначного лексического значения (омонима), значения с наименее вероятным набором тэгов устраняются.

Получив данные наборы кортежей, извлекаются начальные формы слов, у омонимов будет несколько кортежей — несколько инфинитивов. Word2Vec создает для каждого инфинитива вектор семантически схожих, часто встречаемых по контексту слов. Проанализировав такие вектора в контексте N-грамм можно разрешить омонимию и получить нужные значения для искомого слова, а значит, «понять» шутку.

Использование созданного комплекса программных модулей для снятия омонимии и модифицированных методов проекта Word2Vec от Google позволило за счет просчета частоты встречаемости слов (косинусного расстояния) получить числовую оценку интересующих нас аналогий, успешно разрешить неоднозначности, такие как омонимы, омофоны, паронимы, омоформы и др., в предложениях на русском языке — детектировать контекстный каламбур, что и предполагалось при поставновке задачи.

Литература

1. Большакова Е. И., Клышинский Э. С., Ландэ Д. В., Носков А. А., Пескова О. В., Ягунова Е. В. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. Пособие. М.: МИЭМ, 2011.
2. Word2Vec Project URL: https://code.google.com/archive/p/word2vec/
3. Distributed Representations of Words and Phrases and their Compositionality URL: http://arxiv.org/pdf/1310.4546.pdf
4. RusVectories URL: http://ling.go.mail.ru/dsm/ru/