УДК 004.912

ПРИМЕНЕНИЕ МОДЕЛИ ДИСТРИБУТИВНОЙ СЕМАНТИКИ WORD2VEC К АНАЛИЗУ ТЕКСТОВОЙ ИНФОРМАЦИИ

Троцюк М.А.
Брестский государственный технический университет
Научный руководитель: Кузьмицкий Н.Н., к. т. н.

Введение

Стремительное развитие информационных технологий приводит к постоянной генерации и накоплению большого количества электронной текстовой информации. Для эффективной работы с крупными массивами таких данных все более востребованными становятся автоматизированные интеллектуальные методы анализа текстовой информации. В отличие от систем анализа видео- и аудиоданных, представленных в виде отдельных уровней интенсивности пикселей и коэффициентов спектральной плотности мощности, в обработке естественных языков традиционно рассматриваются слова как некие дискретные символы, не предоставляющие системе полезной готовой информации об отношениях, которые могут существовать между ними. Например, автоматически невозможно эффективно использовать информацию, посвященную кошкам в ходе обработки собак (хотя они относятся к животным, млекопитающим, четвероногим и т. д.). Для преодоления данной проблемы предлагается использовать векторное представление слов. Теоретической базой для векторных представлений является дистрибутивная семантика – область, занимающаяся вычислением семантической близости между лингвистическими единицами (словами, словосочетаниями и т. п.) на основании их распределения в массивах лингвистических данных [1]. Дистрибутивная семантика основывается на гипотезе: лингвистические единицы, встречающиеся в схожих контекстах, имеют близкие характеристики.

Математическая модель дистрибутивной семантики

В качестве способа представления модели используются векторные пространства, отражающие информацию о дистрибуции лингвистических единиц в виде многоразрядных векторов. Координаты векторов являются числами, отражающими частоту встречаемости лингвистической единицы в данном контексте, а семантическая близость между ними вычисляется как расстояние между соответствующими векторами. В исследованиях по дистрибутивной семантике чаще всего используется косинусная мера, которая вычисляется по следующей формуле (где A и B – входные вектора):

$$Cos(A,B) = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}. \quad (1)$$

Существует множество различных моделей дистрибутивной семантики, однако в реальных приложениях зачастую возникает проблема слишком большой размерности векторов, соответствующей огромному числу контекстов, представленных в текстовом кор-

пусе. Одной из перспективных технологий получения векторов малой размерности является машинное обучение, в частности искусственные нейронные сети. При обучении таких предсказательных моделей целевым представлением каждого слова является сжатый вектор относительно небольшого размера, для которого в ходе множественных проходов по обучающему корпусу максимизируется сходство с векторами соседей и минимизируется сходство с векторами слов, соседями не являющимися.

Использование модели

Для оценки возможностей предсказательных моделей дистрибутивной семантики исследована технология анализа семантики естественных языков word2vec, разработанная группой исследователей Google в 2013 году [2]. На ее вход подается большой текстовый корпус, а выходом являются векторные представления слов, получаемые в ходе обучения. В word2vec применяются два основных алгоритма обучения: Continuous Bag of Words и Skip-gram. Первый предсказывает текущее слово, исходя из окружающего его контекста, а второй – использует текущее слово, чтобы предугадывать окружающие его слова. Порядок слов контекста не оказывает влияния на результат ни в одном из этих алгоритмов.

С помощью данной технологии и известного текстового корпуса на основе «Google News dataset» размером около 100 миллиардов слов были определены семантические аналоги для выборки из 100 слов английского языка (в таблице приведены некоторые примеры).

Таблица 1 – Выявленные семантические аналоги и косинусные расстояния между ними

Исходное слово		**Исходное слово**		**Исходное слово**	
Black_Sabbath		Belarus		Linux	
Аналог	***Cos(A, B)***	**Аналог**	***Cos(A, B)***	**Аналог**	***Cos(A, B)***
BLACK_SABBATH	0.709492	Ukraine	0.821953	GNU_Linux	0.806687
Led_Zeppelin	0.662612	Belarusian	0.788542	Linux_OS	0.751182
Judas_Priest	0.662190	Belarussian	0.774779	Unix	0.743429
Iron_Maiden	0.655039	Moldova	0.741623	Ubuntu_Linux	0.740498
Fleetwood_Mac	0.644911	Russia	0.730356	Red_Hat_Linux	0.740256
guitarist_Tony_Iommi	0.641936	Kazakhstan	0.720747	Linux_kernel	0.732418
Bassist_Geezer_Butler	0.641535	Latvia	0.710845	Linux_distributions	0.730769
Led_Zep	0.637333	Minsk	0.699619	Ubuntu	0.726819
Glenn_Danzig	0.633546	Azerbaijan	0.698007	Debian_Linux	0.723264
Van_Halen	0.626795	Lithuania	0.682613	OS	0.717712

Анализ результатов показывает, что технология word2vec с высокой эффективностью способна находить семантически близкие слова. Для получения количественных оценок точности модели проведено ее тестирование с помощью наборов, представляющих собой семантически близкие пары слов вида (Athens Greece), (Rome, Italy). Суть теста заключается в том, что, отняв от векторного представления слова «Greece» вектор слова «Athens» и прибавив «Rome», должен получиться вектор, близкий к «Italy». Исследуемая модель продемонстрировала верный результат в 76.8% случаев для 13000 тестовых наборов, что показывает работоспособность технологии. Ограничением моделей word2vec является необходимость использования больших текстовых корпусов

для обучения и зависимость от языка, в связи с чем представляет интерес исследование более сложных моделей типа seq2seq, используемых в частности в машинном переводе.

Заключение

В представленной работе рассмотрено применение модели дистрибутивной семантики к анализу текстовой информации на примере технологии word2vec. Данные модели могут применяться для решения целого ряда практических задач: выявления семантической близости слов и словосочетаний, определения тематики документов, кластеризации документов для информационного поиска, определения тональности высказываний. Далее планируется рассмотреть другие методы машинного обучения в задачах анализа текстовой информации, выполнить сравнение точности работы различных моделей и эффективности их программных реализаций.

Список цитированных источников

1. Дистрибутивная семантика // Википедия [Электронный ресурс]. – Режим доступа: https://ru.wikipedia.org/wiki/Дистрибутивная_семантика – Дата доступа: 14.10.2017.

2. Word2vec // Google Code Archive [Электронный ресурс]. – Режим доступа: https://code.google.com/archive/p/word2vec/ – Дата доступа: 13.10.2016

3. Ian Goodfellow. Deep Learning [Text] / Ian Goodfellow, Yoshua Bengio, Aaron Courville – MIT Press, 2016 – 652.

УДК 004.925.8

ТЕХНОЛОГИЯ МОДЕЛИРОВАНИЯ ТЕЛ ВРАЩЕНИЯ В ПРОСТРАНСТВЕ

Фесько В.В., Орлова А.С.
Белорусский государственный университет
информатики и радиоэлектроники, г. Минск
Научный руководитель: Баркова Е.А., к. физ.-мат. наук, доцент

Всем известно, что телами вращения называют объёмные тела, возникающие при вращении плоской геометрической фигуры, ограниченной кривой, вокруг оси, лежащей в той же плоскости.

Такие тела имеют широкое применение в науке и технике. Тела вращения наиболее распространены в машиностроении; используются при конструировании космических зондов и спутников. Автомобили и корабли, станки, научные установки имеют в своей конструкции рассматриваемые тела.

С машинами и различными конструкциями все понятно, но что насчет человека? Тела вращения широко используются в медицине. Посмотрите хотя бы на бионические протезы. И робототехника при создании очередного человекоподобного механизма не обходится без таких тел. Конечно же, такое широкое применение обусловлено полезными свойствами тел вращения: тут и специфическое отражение лучей света, и обтекаемость, и гладкость поверхности, и просто эстетическое удовлетворение.