Фролов И.В.
Научный руководитель: к.т.н. Е.Е. Канунова
Муромский институт (филиал) федерального государственного образовательного учреждения высшего образования «Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых»
602264, г. Муром, Владимирская обл., ул. Орловская, 23

Распознавание тональности текста на основе нейронных сетей

Нейронная сеть – это громадный распределенный параллельный процессор, состоящий из элементарных единиц обработки информации, накапливающих экспериментальные знания и предоставляющих их для последующей обработки. Нейронная сеть сходна с мозгом с двух точек зрения:

- знания поступают в нейронную сеть из окружающей среды и используются в процессе обучения;
- для накопления знаний применяются связи между нейронами, называемыми синоптическими весами.

Основой представленной в докладе работы является исследование методов глубинного изучения (deep learning) с использованием нейросетевых моделей для решения задач обработки текстов. Для того чтобы проводить исследования, необходимо сначала перевести естественный язык в понятный для компьютера формат, в данном случае – числовой. Для представления слов и документов в векторном виде автором использовались различные модели, такие как «мешок слов» (Bag of Words), Word2Vec, Doc2Vec [1,2].

Одним из распространенных подходов является «мешок слов» (Bag-of-Words). «Мешок слов» – это модель, которая обучается на словаре, составленном из слов всех документов.

Алгоритм построения модели:

1. Составляем словарь из всех слов, встречающихся в тексте, предварительно исключив все знаки препинания, числа и «стоп-слова».
2. Для каждого документа определяем вектор, каждая компонента которого соответствует термину из словаря, а ее значение определяется числом, сколько раз это слово встретилось в тексте. Размерность вектора соответствует мощности словаря.

В модели Word2Vec для получения хороших векторов используется машинное обучение. Одним из популярных методов является построение искусственных нейронных сетей. Изначально задается размерность векторов, которые заполняются случайными величинами. Во время обучения значения компонент векторов будут меняться, причем вектор каждого слова будет максимально схож с векторами типичных соседей и максимально отличаться от векторов слов, которые соседями данному слову не являются. Сами компоненты векторов никак не связаны с конкретными словами из словаря.

Алгоритм построения модели:

1. Составляется словарь терминов, встретившихся во всех документах. Его размер практически не ограничивается, исключаются только слова, имеющие наименьшую встречаемость.
2. Каждому термину в словаре сопоставляется частота встречаемости во всех документах
3. Для кодирования словаря строится дерево Хаффмана.
4. Производится субдискретизация частых слов (параметр задается при создании модели).
5. Для этих слов применяется один из алгоритмов CBOW (Continuous Bag-of-Words) или Skip-gram.
6. Применяется нейронная сеть прямого распространения с функцией активации иерархический softmax или негативное семплирование (negative sampling).

Алгоритм Doc2Vec - алгоритм обучения без учителя, учится получать распределенные векторы для частей текстов. Тексты могут быть переменной длины: от предложения до большого документа.

В данной модели векторные представления документов обучаются предсказывать слова в документе, точнее берется вектор документа и объединяется с несколькими векторами слов из

него, и модель пытается предсказать следующее слово с учетом контекста. Векторы слов и документов обучаются с использованием метода стохастического градиентного спуска и метода обратного распространения ошибки. Векторы документов являются уникальными, а векторы одинаковых слов в разных документах совпадают.

В качестве примера обработки текста была выбрана задача определения тональности рецензий пользователей нескольких интернет - сервисов. Определение тональности текстов является весьма актуальной задачей. Ежедневно тысячи пользователей штудируют Интернет в поисках мнений о том или ином товаре, услугах организаций и прочее. Отзывы помогают определиться с выбором не только людям, но и компаниям, что также полезно. При помощи отзывов организация может судить о качестве своей работы. Естественно, не стоит забывать о более глобальных задачах, например, исследование политических настроений в преддверии выборов или оценка существующей власти.

Задача подразумевает обучение классификаторов на имеющемся множестве размеченных данных, которые будут разделены на два подмножества: обучающее и тестовое. Каждое подмножество представлено в виде текста рецензии на русском языке и оценки, несущей позитивный или негативный мотив.

На основе исследований можно сделать вывод, что нейросетевые модели являются передовыми в области обработки текстов на естественном языке.

Литература

1 Классификация текста с помощью мешка слов http://datareview.info/article/klassifikatsiya-tekstov-s-pomoshhyu-meshka-slov-rukovodstvo/

2 Современные методы анализа тональности текста http://datareview.info/article/sovremennyie-metodyi-analiza-tonalnosti-teksta/