УДК 004.8

Нейросетевый метод снятия омоним

Нгуен Нгок Зиеп, Ле Мань Ха - Московский Физико-Технический Институт

Введение

В лексической системе русского языка есть слова, которые звучат одинаково,

но имеют совершенно разные значения. Такие слова называются лексическими

омонимами, а звуковое и грамматическое совпадение разных языковых единиц,

которые семантически не связаны друг с другом, называется омонимией [1].

Например:

1. Ключ(1) – «Родник» (Студеный ключ)

Ключ(2) – Стальной ключ для отпирания и запирания замка

2. Лук(1) – «Растение» зеленый лук

Лук(2) – «оружие» тугой лук

В отличие от многозначных слов лексические омонимы не обладают предметно-

семантической связью, т.е. у них нет общих семантических признаков, по ко-

торым можно было бы судить о полисемантизме одного слова. Полная лекси-

ческая омонимия – это совпадение слов, принадлежащих к одной части речи,

во всех формах. Пример: наряд(1) – «одежда», наряд(2) – «распоряжение»; За-

мок(1) – «123», Замок(2) – «здание»; Автоматическая система перевода делится

на несколько этапов, одним из которых является морфологический. На этом эта-

пе для каждого слова определяются морфологические характеристики: род, чис-

ло, падеж, склонение, и т.п., а также начальная форма слова (Лемма). Процесс

морфологической разметки осложняется омонимами [1].

Методы представления текста

В настоящее время теория и практика машинного обучения переживают настоя-

щую «глубинную революцию», вызванную успешным применением методов deep

learning, представляющих собой третье поколение нейронных сетей. Сети, обу-

ченные с помощью алгоритмов глубинного обучения, не просто превзошли по

точности лучшие альтернативные подходы, но и в ряде задач проявили зачат-

ки понимания смысла подаваемой информации (например, при распознавании

изображений, анализе текстовой информации и т.п.). Наиболее успешные совре-

менные промышленные методы компьютерного зрения, распознавания речи и

1машинного перевода построены на использовании глубинных сетей, а гиганты

IT-индустрии, такие как Apple, Google, Facebook, скупают коллективы исследо-

вателей, занимающихся глубинным обучением целыми лабораториями.

Представление слова как вектор - в настоящее время одна из самых инте-

ресных областей исследований в глубинном обучении, хотя данный подход был

изначально введен Bengio [2] и др. более десяти лет назад. Самые известные мо-

дели представления слова как вектор – “Continuous-Bag-of-Word” и “Skip-Gram”,

описание представленое Mikolov[3] и реализация моделей Word2vec [4], опублико-

ваная на сайте Google project, которая привлекла большое внимание в последние

два года. Эти модели использовали простую нейронную сеть для отображения

слов, которые ближе к слову по значению. Поэтому процесс обучения для этих

моделей быстро работает с большими данными.

Рис. 1: Модели CBOW и Skip-gram.

CBOW модель – Использует окружающие слова (предложение без заданного

слова), что бы предугадать заданное слово. Входные вектора в нейронную сеть

W(t-2),W(t-1), W(t+1), W(t+2), а выходной вектор из сети это заданное слово

W(t). В результате обучения получается 2 вектора представления слова. Входной

вектор V - это вектор представления предложения описания слова, выходной

вектор V’ – Это вектор представления слова. SKIP-GRAM модель – Использует

слово, что бы предугадать предложение описания заданного слова. Эти модели

могут использоваться во многих задачах анализа естественного языка, таких как

классификация документов, распознавание спама или машинный перевод.

2Рис. 2: Нейронная сеть, использованная для получения вектора представления

слова. Входной слой и выходной имеют V нейронов, а скрыты имеет N нейронов.

На рисунке 3 отображены вектора для чисел и животных на английском и

испанском языках [5], на рисунке можно легко заметить, что эти понятия имеют

схожие геометрические композиции. Причина в том, что что понятия, которые

основаны на реальном мире разделились одинако, как и для множества распро-

страненных языков.

Рис. 3: Распределение векторных представлений слов чисел и животных на ан-

глийском языке (слева) и испанском (справа). Пять векторов на каждом языке

были спроектированы на два измерения с использованием метода РСА. Видно,

что эти концепции имеют схожие геометрические представления в обоих про-

странствах, что позволяет предполагать возможное точное линейное отображе-

ние из одного пространства в другое.

Один хороший пример двуязычного вхождения слова, приводится в работе

Socher и др. [6]. Было показано как можно научиться вставлять слова из двух

разных языков в одно общее пространство. В рассмотренном в работе случае,

учились вставлять английские и китайские слова в том же пространстве. Если

значения слов окружающих омонимы разные, тогда расстояние между омонима-

ми тоже большое.

Рис. 4: Визуализация двухязычного пространства - Английский и китайский.

Нейронные сети и Автоэнкодер

Нейронные сети прямого распространения часто используются для обучения с

учителем и используются, например, для классификации [7].

Рис. 5: Многослойная нейронная сеть

Часто используемая функция активации - сигмоид:

Полезное свойство сигмоида - её производная функции:

Обучение такой нейронной сети производится обычно методом обратного рас-

пространения ошибки таким образом, чтобы минимизировать среднеквадрати-

ческую ошибку сети на обучающей выборке. Таким образом, обучающая выбор-

ка содержит пары векторов признаков (входные данные) и эталонных векторов

(маркированные данные) (x, y).

Метод обратного распространения ошибки:

Для выходного нейрона:

Для нейронов скрытых слоев:

Коррекция весов:

Для выходного нейрона:

Для нейронов скрытых слоев:

В реальной практике, маркированных данных очень мало, для них требу-

ется много сил и времени. Автоэнкодер представляет собой алгоритм обучения

без учителя, который использует нейронную сеть и метод обратного распростра-

нения ошибки для того, чтобы добиться того, что входной вектор признаков

вызывал выход сети, равный входному вектору, т.е. y = x. Автоэнкодер явла-

ется специальной архитектурой искусственных нейронных сетей, позволяющая

применять обучение без учителя при использовании метода обратного распро-

странения ошибки. Простейшая архитектура автоэнкодера — сеть прямого рас-

пространения, без обратных связей, наиболее схожая с перцептроном и содержа-

щая входной слой, скрытый слой и выходной слой. В отличие от перцептрона,

выходной слой автоэнкодера должен содержать столько же нейронов, сколько и

входной слой [8].

Рис. 6: Автоэнкодер

Цель автоэнкодер - чтобы выход нейронной сети был наиболее близким к

входному вектору. Для того, чтобы решение этой задачи было нетривиальным,

на топологию сети накладываются особые условия:

1. Количество нейронов скрытого слоя должно быть меньше, чем размерность

входных данных.

2. Активация нейронов скрытого слоя должна быть разреженной.

Первое ограничение позволяет получить сжатие данных при передаче входно-

го сигнала на выход сети. Такая сжатие возможно, если в данных есть скрытые

взаимосвязи, корреляция признаков или структура.Второе ограничение – требо-

вание разреженной активации нейронов скрытого слоя, — позволяет получить

нетривиальные результаты даже когда количество нейронов скрытого слоя пре-

вышает размерность входных данных. Будем считать нейрон активным, когда

значение его функции передачи близко к 1 и наоборот, неактивным если зна-

чение его функции передачи близко к 0. Разреженная активация – это когда

количество неактивных нейронов в скрытом слое значительно превышает коли-

чество активных.

Эти ограничения заставляют нейросеть искать обобщения и корреляцию в

поступающих на вход данных, выполнять их сжатие. Таким образом, нейросеть

автоматически обучается выделять из входных данных общие признаки, кото-

7рые кодирутся в значениях весов сети. Хотели, чтобы средняя активация каж-

дого скрытого нейронна приняла значение, наиболее близкое к заданному раз-

реженному параметру (порядка 0.05). Для этого был добавлен в каждый нейрон

скрытого слоя параметр разреженности ρ:

Хотели, чтобы средняя активация каждого скрытого нейронна приняла зна-

чение, наиболее близкое к ρ:

Штравная функция:

Производная штравной функции:

Рекурсивный автоэнкодер и векторное представление текста

В данном разделе посмотрим как представит текст в виде числового вектора. Для

решения классификации и других задач с помощью нейронных сетей, вход дол-

жен иметь фиксированную длину, а длинна текста явлается произвольной. Для

фиксирования размера входов нейронных сетей используется метод векторного

представления текста с помощью рекурсивного автоэнкода - входной и выходной

слои которого имеют 2K нейронов, а скрытый слой - K нейронов [9]:

Автоэнкодер объединяет два слова x1, x2 (вектор длины 2K) в один вектор y

(длина К):

Этот процесс повторяется N-1 раз для текта длинной N. В результате полу-

чается конечный вектор - семантическое векторное представление текста, этот

вектор используем как вход для системы обучения.

Кластеризация значений слова

K-means

K-means - наиболее популярный метод кластеризации. Был изобретён в 1950-х го-

дах математиком Гуго Штейнгаузом и почти одновременно Стюартом Ллойдом.

Особую популярность приобрёл после работы Маккуина. Действие алгоритма

таково, что он стремится минимизировать суммарное квадратичное отклонение

точек кластеров от центров этих кластеров [10]:

Следующим шагом является принятие каждую точку, принадлежащую к дан-

ному набору данных и связать его в ближайший центр. В этот момент нужно

пересчитать К новые центра, как барицентра кластеров в результате предыду-

щего шага. После того как вычислили эти новые К центроиды, новую привязку

должно быть сделано между теми же набора данных точек и ближайшей ново-

го центра. Цикл был сформирован. В результате этого цикла можно заметить,

что K-центры меняют шаг за шагом местоположения пока больше изменений не

сделано или, другими словами центры не двигаться больше.

Маркировка значений слова

В первых получить все возможные вектора слов, затем вектора используются,

чтобы построить вектора контекстов с применением автоэнкодер[Session 4]. На-

пример у нас есть миллион векторов описания о слове Замок, необходимо Вы-

польнять вычисления по алгоритму К-Means для этих веторов. В результате,

можно маркировать кождое слово с значением. Таким образом, все омонимы

будут маркированы по значению, тогда можно использавать эти резутаты длы

задачи снятия омонимий.

Список литературы

[1] Омонимы в русском языке http://gramma.ru/

[2] Yoshua Bengio, Rejean Ducharme, Pascal Vincent, and Christian Jauvin. 2003.

A neural probabilistic language model. In Journal of Machine Learning Research,

pages 1137–1155.

[3] Mikolov, Tomas, et al. "Distributed representations of words and phrases and

their compositionality."Advances in neural information processing systems. 2013.

[4] Word2Vec Project https://code.google.com/p/word2vec/

[5] Mikolov, Tomas, Quoc V. Le, and Ilya Sutskever. "Exploiting similarities among

languages for machine translation."arXiv preprint arXiv:1309.4168 (2013).

[6] Zou, W. Y., Socher, R., Cer, D. M., Manning, C. D. (2013). Bilingual Word

Embeddings for Phrase-Based Machine Translation. In EMNLP (pp. 1393-1398)

[7] Stanford UFLDL tutorial, Multilayer neural networks

http://ufldl.stanford.edu/tutorial/supervised/MultiLayerNeuralNetworks/

[8] Stanford UFLDL tutorial, Autoencoders

http://ufldl.stanford.edu/tutorial/unsupervised/Autoencoders/

[9] Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions,

Richard Socher, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, Christopher

D. Manning, 2011

[10] Алгоритм K-means https://ru.wikipedia.org/wiki/K-means