УДК 004.8

Нейросетевый метод снятия омоним

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

1

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 [10].

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 кото-

рым можно было бы судить о полисемантизме одного слова. Полная лексическая

омонимия – это совпадение слов, принадлежащих к одной части речи, во всех

формах. Пример: наряд(1) – «одежда», наряд(2) – «распоряжение»; Замок(1) –

«123», Замок(2) – «здание»;

Автоматическая система перевода делится на несколько этапов, одним из ко-

торых является морфологический. На этом этапе для каждого слова опреде-

ляются морфологические характеристики: род, число, падеж, склонение, и т.п.,

а также начальная форма слова (Лемма). Процесс морфологической разметки

осложняется омонимами [10].

2

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

1изображений, анализе текстовой информации и т.п.). Наиболее успешные совре-

менные промышленные методы компьютерного зрения, распознавания речи и

машинного 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 [1] и др. более десяти лет назад.

Самый известные модели представления слова как вектор представлены Mikolov

(2013a)[9] и реализация моделей Word2vec [11] опубликована на сайте Google

project, которая привлекла большое внимание в последние два года – это “Continuous-

Bag-of-Word” и “Skip-Gram”. 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.

2Рис. 2: Показал нейронную сеть, использованную чтобы получить вектор пред-

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

N нейронов.

Эти модели могут использоваться во многих задачах анализа естественного

языка, таких как классификация документов, распознавание спама или машин-

ный перевод.

На рисунке 3, визуализировали векторы для чисел и животных на англий-

ском и испанском языках [7], и он может быть легко видеть, что эти понятия

имеют схожие геометрические композиции. Причина в том, что, как и все рас-

пространенные языки поделиться понятия, которые основаны на реальном мире.

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

глийском языке (слева) и испанском (справа).пять Векторов на каждом языке

были спроектированы до двух измерениях с использованием РСА. Видно, что

эти концепции имеют схожие механизмы геометрические в обоих пространствах,

предполагая, что можно узнать точную линейное отображение из одного про-

странства в другое.

Один хороший пример это двуязычный слово вложение, производится в Socher

др. (2013a)[8]. Мы можем научиться вставлять слова из двух разных языков в

одной общей пространстве. В этом случае, мы учимся вставлять английские и

китайский слов в том же пространстве.

Одна вещь, мы можем сделать, чтобы почувствовать слова вложения про-

странства для визуализации их трет-СНЭ, сложной техники для визуализации

многомерных данных

4Рис. 4: Английский и китайский язык.

Таким образом, возможно снять омонимию через окружающие слова в пред-

ложении. Если значение слов окружающих омонимы разные, тогда расстояние

между омонимами тоже большое.

3

Нейронные сети и Автоэнкодер

Нейронные сети прямого распространения часто используются для обучения с

учителем и используются, например, для классификации [4].

Рис. 5: Многослойная нейронная сеть

5Часто используемая функция активации - сигмоид:

f (z j ) = σ(z j ) = (1 + e −z j ) −1

(1)

Где:

z j =

w ij x i

(2)

Полезное свойство сигмоида - её производная функции:

∂σ

= σ(z)(1 − σ(z))

∂z

(3)

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:

δ = z − y

(4)

Для нейронов скрытых слоев:

n

δ i =

δ j w ij

(5)

j=0

Коррекция весов:

Для выходного нейрона:

w i0 = w i0 + ηδy i

(6)

Для нейронов скрытых слоев:

w ij = w ij + ηδ j y j (1 − y j )y i

(7)

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 явлается специальной архитектурой искусственных нейронных

сетей, позволяющая применять обучение без учителя при использовании метода

обратного распространения ошибки. Простейшая архитектура автоэнкодера —

6сеть прямого распространения, без обратных связей, наиболее схожая с перцеп-

троном и содержащая входной слой, скрытый слой и выходной слой. В отличие

от перцептрона, выходной слой автоэнкодера должен содержать столько же ней-

ронов, сколько и входной слой [5].

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 – требование разреженной активации нейронов скрытого

слоя, — позволяет получить нетривиальные результаты даже когда количество

нейронов скрытого слоя превышает размерность входных данных. Будем считать

нейрон активным, когда значение его функции передачи близко к 1 и наоборот,

7неактивным если значение его функции передачи близко к 0. Разреженная акти-

вация – это когда количество неактивных нейронов в скрытом слое значительно

превышает количество 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, которые

кодирутся в значениях весов сети.

Мы хотим, чтобы средняя активация каждого скрытого нейронна приня-

ла значение, наиболее близкое к заданному разреженному параметру (порядка

0.05). Для этого, мы добавим в каждый нейрон скрытого слоя параметр разре-

женности ρ:

1

ρ ˆ j =

m

m

(2)

a j (x (i) )

(8)

i=1

Мы хотим, чтобы средняя активация каждого скрытого нейронна приняла

значение, наиболее близкое к ρ:

ρ ˆ j = ρ

(9)

Штравная функция:

s 2

S =

KL(ρ| ρ ˆ j ) (10)

ρ

1 − ρ

KL(ρ| ρ ˆ j ) = ρ log + (1 − ρ) log

ρ ˆ j

1 − ρ ˆ j (11)

j=1

Производная штравной функции:

ρ

1 − ρ

∂KL(ρ| ρ ˆ j )

= − +

∂ρ j

ρ ˆ j 1 − ρ ˆ j

4

(12)

Рекурсивный автоэнкодер и векторное представ-

ление 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 [6]:

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

(длина К):

8y = f (W (1) [x1, x2] + b (1) )

(13)

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.

Рис. 7: Автоэнкодер

5

5.1

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 [12]:

c i

c

(x i − v i ) 2

J(V ) =

j

(14)

j

Следующим шагом является принятие каждую точку, принадлежащую к дан-

ному набору данных и связать его в ближайший центр. В этот момент мы долж-

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

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

9должно 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.

5.2

Маркировка значений слова

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

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

Например у нас есть миллион векторов описания о слове Замок, необходимо

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

мы сможем маркировать кождое слово с значением.

Таким образом, все омонимs будут маркированы по значению, мы сможем

использавать эти резутаты длы задачи снятия омонимий.

Список литературы

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

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

pages 1137–1155.

[2] Национальный корпус русского языка, www.ruscorpora.ru

[3] Richard Socher, John Bauer, Christopher D. Manning, and Andrew Y. Ng. 2013.

Parsing with compositional vector grammars. In ACL.

[4] Stanford UFLDL tutorial, Multilayer neural networks

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

[5] Stanford UFLDL tutorial, Autoencoders

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

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

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

D. Manning, 2011

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

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

[8] 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)

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

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

10[10] Омонимы в русском языке http://gramma.ru/

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

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