УДК 004.8

Нгуен Нгок Зиеп, Ле Мань Ха

Московский физико-технический институт (государственный университет)

Нейросетевой метод снятия омонимии

Омонимы – это слова, которые звучат одинаково, но имеют разные значения. В данной работе рассмотрим метод векторного представления слов и метод классификации омонимов с помощью автоэнкодера и нейронных сетей.

Ключевые слова: омоним, нейронные сети, автоэнкодер.

Nguyen Ngoc Diep, Le Manh Ha

Moscow Institute of Physics and Technology (State University)

Classification homonyms using neural networks

Homonyms are words that are spelled in the same way but have different meanings. In this paper, we consider a metod of word presentation as a vector and classification of homonyms using an autocoder and neutral networks.

Key words: homonym, neural networks, autoencoder.

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 [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 которым 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 которых является морфологический. На этом этапе для каждого слова определяются морфологические характеристики: род, число, падеж, склонение и т.п., а также начальная форма слова (лемма). Процесс морфологической разметки осложняется омонимами [1].

2. Методы представления текста

В настоящее время теория и практика машинного обучения переживают настоящую «глубинную революцию», вызванную успешным применением методов deep learning, представляющих собой третье поколение нейронных сетей. Сети, обученные с помощью алгоритмов глубинного обучения, не просто превзошли по точности лучшие альтернативные подходы, но и в ряде задач проявили зачатки понимания смысла подаваемой информации (например, при распознавании изображений, анализе текстовой информации и т.п.). Наиболее успешные современные 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 исследователей, занимающихся глубинным обучением целыми лабораториями.

Представление слова как вектора – в настоящее время одна из самых интересных областей исследований в глубинном обучении, хотя данный подход был изначально введен Bengio [2] и др. более десяти лет назад. Самые известные модели представления слова как вектора – Continuous-Bag-of-Word и Skip-Gram, описание, представленное Mikolov [3], и реализация моделей Word2vec [4], опубликованая на сайте Google project, привлешая 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.

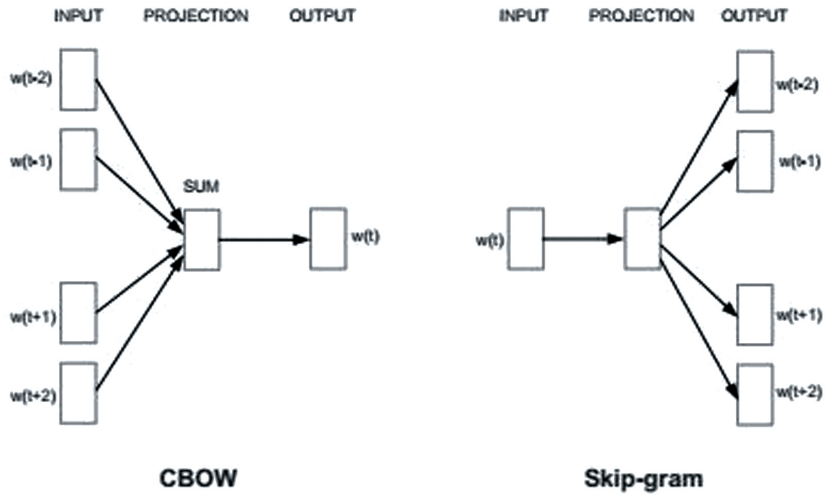

Рис. 1. Модели CBOW и Skip-gram

CBOW-модель использует окружающие слова (предложение без заданного слова), чтобы предугадать заданное слово. Входные векторы 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, чтобы предугадать предложение описания заданного слова. Эти модели могут использоваться во многих задачах анализа естественного языка, таких как классификация документов, распознавание спама или машинный перевод.

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 случае работы [6] учились вставлять английские и китайские слова в том же пространстве. Если значения слов, окружающих омонимы, разные, тогда расстояние между омонимами большое.

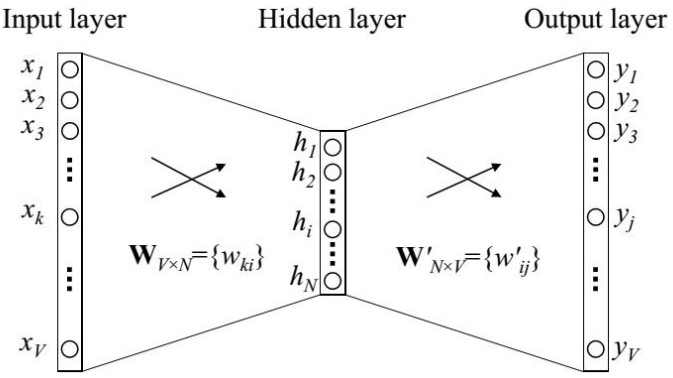

Рис. 2. Нейронная сеть, использованная для получения вектора представления слова. Входной слой и выходной имеют V нейронов, а скрытый – N нейронов

На рис. 3 отображены векторы 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 понятия, которые основаны на реальном мире, разделились одинаково, как и для множества распространенных языков.

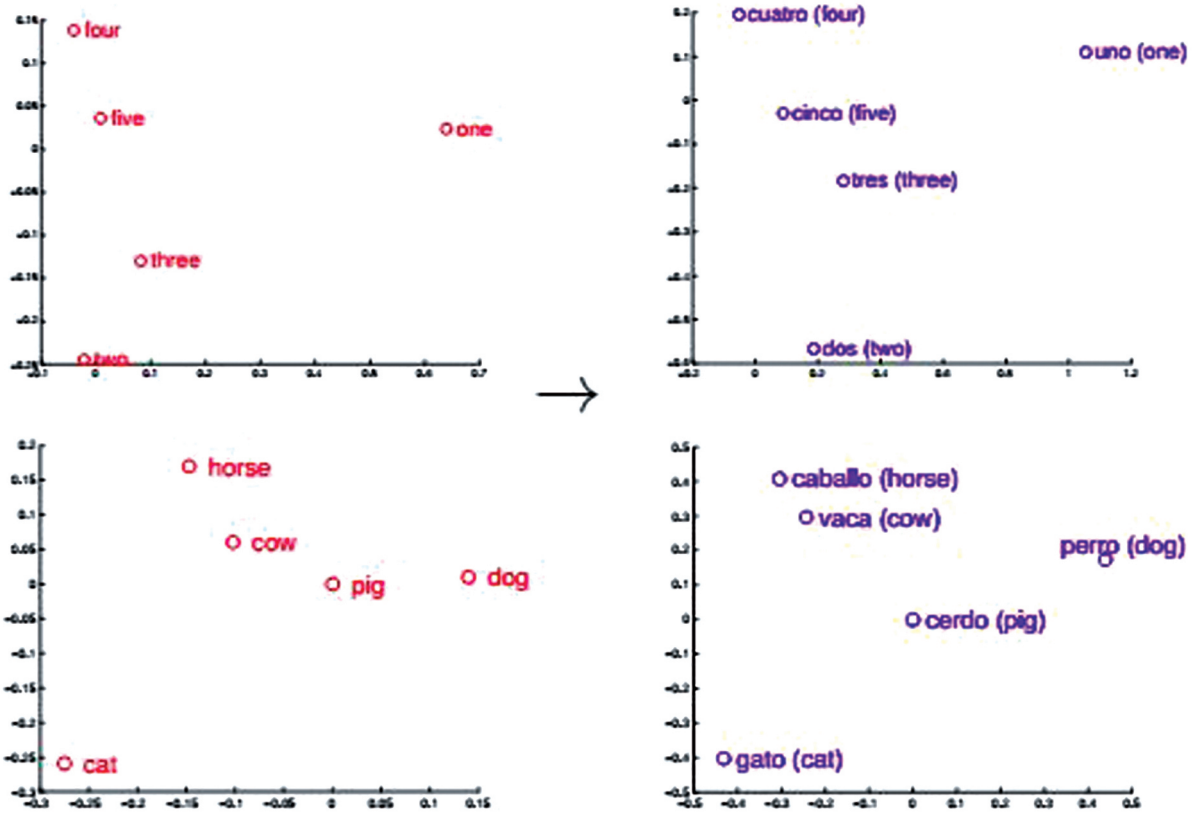

Рис. 3. Распределение векторных представлений слов чисел и животных на английском языке (слева) и испанском (справа). Пять векторов на каждом языке были спроектированы на два измерения с использованием метода PCA. Видно, что эти концепции имеют схожие геометрические представления в обоих пространствах, что позволяет предполагать возможное точное линейное отображение из одного пространства в другое

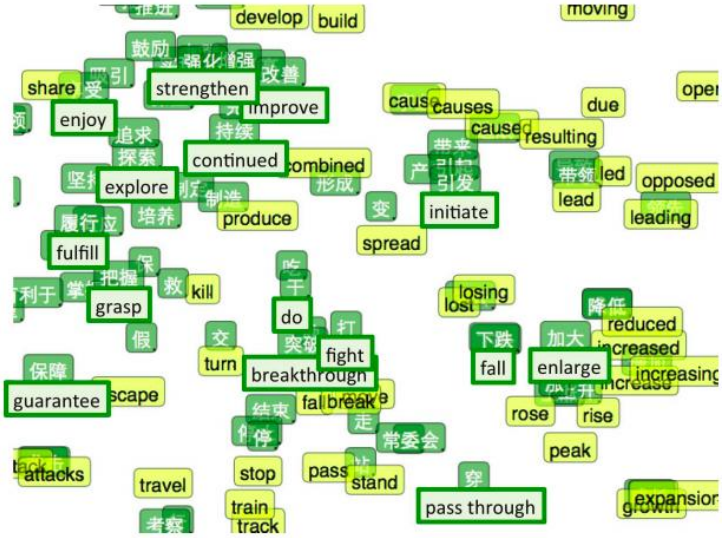

Рис. 4. Визуализация двуязычного пространства – английский и китайский

3. Нейронные сети и автоэнкодер

Нейронные сети прямого распространения часто используются для обучения с учителем и используются, например, для классификации [7].

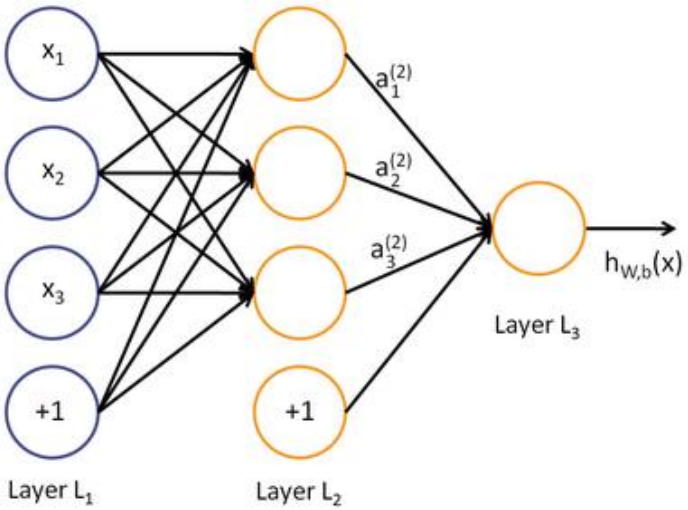

Рис. 5. Многослойная нейронная сеть

Часто используемая функция активации – сигмоида:

$$f(z_j) = \sigma(z_j) = (1 + e^{-z_j})^{-1}, \tag{1}$$

где

$$z_j = \sum w_{ij} x_i. \tag{2}$$

Полезное свойство сигмоиды – её производная функции:

$$\frac{\partial \sigma}{\partial z} = \sigma(z)(1 - \sigma(z)). \tag{3}$$

Обучение такой нейронной сети производится обычно методом обратного распространения ошибки таким образом, чтобы минимизировать среднеквадратическую ошибку сети

на обучающей выборке. Таким образом, обучающая выборка содержит пары векторов признаков (входные данные) и эталонных векторов (маркированные данные) (x, y).

Метод обратного распространения ошибки:

для выходного нейрона:

$$\delta = z - y; \tag{4}$$

для нейронов скрытых слоев:

$$\delta_i = \sum_{j=0}^{n} \delta_j w_{ij}. \tag{5}$$

Коррекция весов:

для выходного нейрона:

$$w'_{i0} = w_{i0} + \eta \delta y_i; \tag{6}$$

для нейронов скрытых слоев:

$$w'_{ij} = w_{ij} + \eta \delta_j y_j (1 - y_j) y_i. \tag{7}$$

В реальной практике маркированных данных очень мало, для них требуется много сил и времени. Автоэнкодер представляет собой алгоритм обучения без учителя, который использует нейронную сеть и метод обратного распространения ошибки для того, чтобы добиться того, чтобы входной вектор признаков вызывал выход сети, равный входному вектору, т.е. $y = x$. Автоэнкодер является специальной архитектурой искусственных нейронных сетей, позволяющей применять обучение без учителя при использовании метода обратного распространения ошибки. Простейшая архитектура автоэнкодера — сеть прямого распространения, без обратных связей, наиболее схожая с перцептроном и содержащая входной слой, скрытый слой и выходной слой. В отличие от перцептрона, выходной слой автоэнкодера должен содержать столько же нейронов, сколько и входной слой [8].

Цель автоэнкодера – 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 входного сигнала на выход сети. Такое 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, неактивным – если значение его функции передачи близко к 0. Разреженная активация – 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 кодируются в значениях весов сети. Необходимо, чтобы средняя активация каждого скрытого нейрона приняла значение, наиболее близкое к заданному разреженному параметру (порядка 0.05). Для этого был добавлен в каждый нейрон скрытого слоя параметр разреженности ρ:

$$\hat{\rho}_j = \frac{1}{m} \sum_{i=1}^{m} \left[a_j^{(2)}(x^{(i)}) \right]. \tag{8}$$

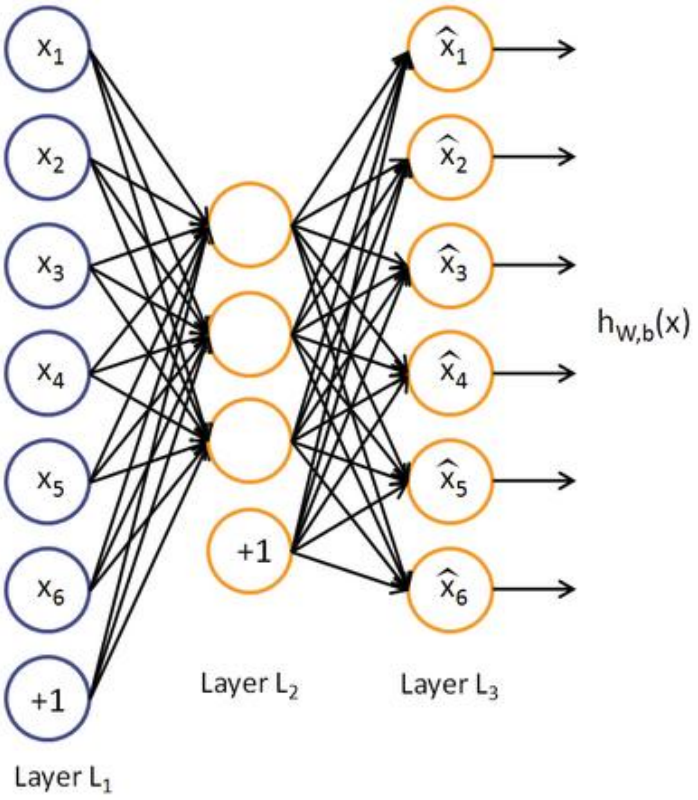

Рис. 6. Автоэнкодер

Необходимо, чтобы средняя активация каждого скрытого нейрона приняла значение, наиболее близкое к ρ:

$$\hat{\rho}_j = \rho. \tag{9}$$

Штрафная функция:

$$S = \sum_{j=1}^{s_2} KL(\rho|\hat{\rho}_j), \tag{10}$$

$$KL(\rho|\hat{\rho}_j) = \rho \log \frac{\rho}{\hat{\rho}_j} + (1-\rho) \log \frac{1-\rho}{1-\hat{\rho}_j}. \tag{11}$$

Производная штрафной функции:

$$\frac{\partial KL(\rho|\hat{\rho}_j)}{\partial \rho_j} = -\frac{\rho}{\hat{\rho}_j} + \frac{1-\rho}{1-\hat{\rho}_j}. \tag{12}$$

4. Рекурсивный автоэнкодер и векторное представление текста

В данном разделе рассмотрим, как представить текст в виде числового вектора. Для решения задачи классификации и других задач с помощью нейронных сетей вход должен иметь фиксированную длину, а длина текста является 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 K):

$$y = f(W^{(1)}[x1, x2] + b^{(1)}). \tag{13}$$

Этот процесс повторяется $N-1$ раз для текста длиной N. В результате получается конечный вектор – семантическое векторное представление текста, этот вектор используется как вход для системы обучения.

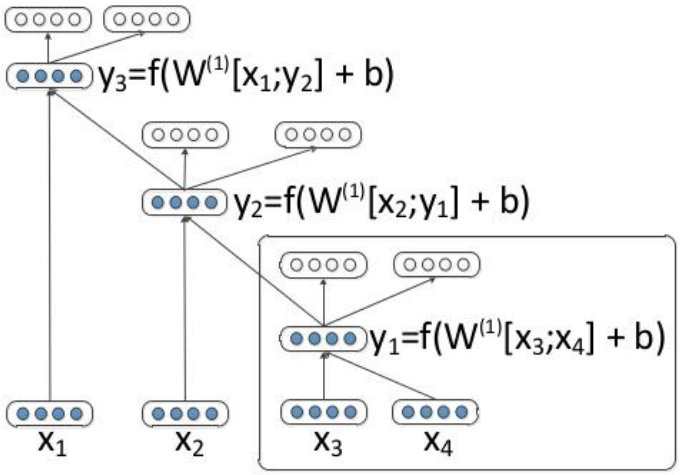

Рис. 7. Автоэнкодер

5. Кластеризация значений слова

5.1. K-means

K-means – наиболее популярный метод кластеризации, был изобретён в 1950-х годах 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]:

$$J(V) = \sum_{j}^{c} \sum_{j}^{c_i} (x_i - v_i)^2. \tag{14}$$

Функция k-means состоит из двух этапов.

1. Первая фаза предназначена для поиска приближенного значения центроидов кластеров и предварительной группировки объектов в кластеры. Все объекты группируются в ближайшие к ним кластеры с последующим расчетом расстояний между объектами и центроидами. Однократное приближение при группировке всех объектов выполняется за одну итерацию.

2. Вторая фаза предназначена для поиска точного и окончательного решения. Группировка каждого из объектов по кластерам с последующим расчетом его расстояния до центроида на этой фазе выполняется по критерию минимизации внутрикластерной суммы расстояний объектов до центроидов кластеров. На каждой итерации расчет выполняется по всем объектам исходного множества.

Недостатком итерационного алгоритма функции k-means является возможность попадания решения в локальный минимум внутрикластерной суммы расстояний точек кластера до его центроида по k кластерам. Эта проблема может быть решена соответствующим выбором начальной точки в начале процедуры кластеризации объектов.

5.2. Маркировка значений слова

Во-первых, следует получить все возможные векторы слов, затем векторы используются, чтобы построить векторы контекстов с применением автоэнкодера [4]. Например, имеется миллион векторов описания слова «замок», далее необходимо выполнить вычисления по алгоритму K-Means для этих векторов. В результате можно маркировать каждое слово со значением. Таким образом, все омонимы будут маркированы по значению, тогда можно использовать эти результаты для задачи снятия омонимий.

Литература

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

2. *Bengio Y., Ducharme R., Vincent P., Jauvin Ch.* A neural probabilistic language model. In Journal of Machine Learning Research, 2003. P. 1137–1155.

3. *Mikolov T.* [*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 T., Le Quoc V., Sutskever I.* 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.* Bilingual Word Embeddings for Phrase-Based Machine Translation // EMNLP. 2013. P. 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. *Socher R., Pennington J., Huang E.H., Ng A.Y., Manning Ch.D.* Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions. 2011.

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

References

1. Homonyms in Russian language. http://gramma.ru/

2. *Bengio Y., Ducharme R., Vincent P., Jauvin Ch.* A neural probabilistic language model. In Journal of Machine Learning Research, 2003. P. 1137–1155.

3. *Mikolov T.* [*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 T., Le Quoc V., Sutskever I.* 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.* Bilingual Word Embeddings for Phrase-Based Machine Translation. EMNLP. 2013. P. 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. *Socher R., Pennington J., Huang E.H., Ng A.Y., Manning Ch.D.* Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions. 2011.

10. K-means algorithm. https://ru.wikipedia.org/wiki/K-means

Поступила в редакцию 09.11.2015.