9. *Anthony R.N.* Planning and Control: A Framework for Analysis. Cambridge: Harvard University Press, 1965.
10. *Ashby W.R.* An Introduction to Cybernetics. L.: Chapman & Hall, 1956.

## Literatura

1. *Anfilatov V.S., Emel'yanov A.A., Kukushkin A.A.* Sistemnyj analiz v upravlenii: uchebnoe posobie. M.: Finansy i statistika, 2005. 367 s.
2. *Klimenko I.S.* K interpretatsii printsipa neobkhodimogo raznoobraziya Eshbi primenitel'no k upravleniyu v sotsial'no-ekonomicheskikh sistemakh // Vestnik Rossijskogo novogo universiteta. Seriya "Slozhnye sistemy: modeli, analiz i upravlenie". 2012. Vyp. 4. S. 45–47.
3. *Klimenko I.S.* Teoriya sistem i sistemnyj analiz: uchebnoe posobie. M.: RosNOU, 2014. 256 s.
4. *Klimenko I.S., Belova N.A., Sharapova L.V.* K probleme opredeleniya tsennosti informatsii // Vestnik Rossijskogo novogo universiteta. Seriya "Slozhnye sistemy: modeli, analiz i upravlenie". 2018. Vyp. 2. S. 54–62.
5. *Klimenko I.S., Korovko P.G., Sharapova L.V.* K probleme otsenivaniya kachestva upravlencheskikh reshenij i effektivnosti upravleniya // Vestnik Rossijskogo novogo universiteta. Seriya "Slozhnye sistemy: modeli, analiz i upravlenie". 2017. Vyp. 1. S. 53–57.
6. *Klimenko I.S., Sharapova L.V.* Obshchaya zadacha prinyatiya resheniya i fenomen neopredelennosti // Vestnik Rossijskogo novogo universiteta. Seriya "Slozhnye sistemy: modeli, analiz i upravlenie". 2019. Vyp. 3. S. 44–56.
7. *O'Konnor Dzh., Makdermott I.* Iskusstvo sistemnogo myshleniya / per. s angl. M.: Al'pina Biznes Buks, 2006. 256 s.
8. *Frajmann A.V.* Ob osobennostyakh primeneniya printsipa neobkhodimogo raznoobraziya dlya otobrazheniya funktsij sistemnogo administratora // Vestnik Rossijskogo novogo universiteta. Seriya "Slozhnye sistemy: modeli, analiz i upravlenie". 2019. № 2. S. 64–69.
9. *Anthony R.N.* Planning and Control: A Framework for Analysis. Cambridge: Harvard University Press, 1965.
10. *Ashby W.R.* An Introduction to Cybernetics. L.: Chapman & Hall, 1956.

DOI: 10.25586/RNU.V9187.20.01.P.093 УДК 519.25+004.8

С.И. Михайлин

# ОЦЕНИВАНИЕ СТЕПЕНИ СЕМАНТИЧЕСКОЙ БЛИЗОСТИ СЛОВ ПОСРЕДСТВОМ ВИЗУАЛИЗАЦИИ ИХ ВЕКТОРНОГО ПРЕДСТАВЛЕНИЯ

Установление семантической близости между набором слов является важной задачей, решение которой позволит продвинуться в ряде направлений, связанных с коммуникацией и передачей информации. Для выявления семантической близости слов предлагается использовать визуализацию их векторного представления с последующей интерпретацией. Описан алгоритм предобработки исходного текста для достижения необходимого результата.

*Ключевые слова:* векторизация, нейронные сети, word2vec, семантика, семантическая близость, визуализация векторов.

S.I. Mikhaylin

# ESTIMATION OF THE DEGREE OF SEMANTIC PROXIMITY OF WORDS BY MEANS OF VISUALIZATION OF THEIR VECTOR REPRESENTATION

Establishing semantic proximity between a set of words is an important task, the solution of which will allow us to advance in a number of areas related to communication and information transfer. To identify the semantic proximity of words, it is proposed to use the visualization of their vector representation with subsequent interpretation. The algorithm of preprocessing the source text to achieve the desired result is described.
*Keywords:* embedding, neural network, word2vec, semantic, semantic proximity, visualization of vectors.

## *Введение*

В век информатизации множество людей обменивается информацией, будь то устные или письменные сообщения, передача изображения или видео. При формировании сообщения, которое впоследствии будет отправлено адресату для последующей интерпретации, в него закладывается основная мысль. Однако само понятие смысла до сих пор не имеет однозначного строгого определения в силу сложности этого понятия, нашедшего свое формальное отображение в термине «семантика».

Нередко звучание и написание двух разных слов существенно расходится, но при этом их смысловое содержание может в значительной степени совпадать. Тем не менее полная идентичность понятий может быть достигнута только в случае их абсолютного совпадения. Например, слова «тонкий» и «стройный» сильно отличаются друг от друга с точки зрения фонетики, более того, они не являются синонимами, но заложенная в них мысль имеет определенное сходство.

Если научиться определять у некоторых наборов слов степень их взаимного семантического различия, то это позволит приблизиться к пониманию и формализации понятия семантики как таковой.

В рамках настоящей работы предпринята попытка построения метода измерения степени семантического различия языковых конструкций на основе методологии визуализации векторного представления (см., например: [4; 5; 6]). Это может позволить продемонстрировать в явном виде семантическое сходство или различие между двумя и более словами и/или предложениями.

## *Векторизация текста*

Сегодня, когда нужную информацию можно найти на множестве сайтов или запросить у своего коллеги через мессенджер, пожалуй, основным форматом передачи информации по-прежнему остается текст. Всякий письменный текст представляет собой определенную последовательность слов и символов, а изменение их порядка может повлечь за собой изменение смысла. Это означает, что определенная последовательность слов формирует контекст, который становится необходимым для адекватной интерпретации сообщений. Таким образом, разные слова, которые встречаются в одних и тех же местах текста, при похожих контекстах могут иметь существенно близкое смысловое значение.

Поскольку зрительное восприятие окружающего мира человеком является наиболее эффективным, представляется целесообразным визуализировать семантическое сходство (различие) через градации цветовой шкалы. Для реализации такого подхода необходимо перейти к векторному пространству в некотором характеристическом базисе, элементами которого являются слова.

Подход, при котором дискретные величины переводятся в непрерывные векторы, в английской литературе носит название *embedding* [4], но в русском языке он не имеет единого, устоявшегося названия, поэтому в рамках настоящей статьи будет использоваться термин *векторизация.*

Рассмотрим простейший случай векторизации предложений. Длиной вектора в таком пространстве может являться количество существующих слов в соответствующем алфавите, а областью значений каждой компоненты вектора является множество {0, 1}, где единица означает появление слова в предложении, а нуль – его отсутствие. Как меру схожести двух векторов можно использовать косинусное расстояние:

$$\cos\_\text{similarity} = 1 - \frac{AB}{\|A\|\|B\|}.$$

В таком случае два предложения, представленные двумя множествами слов, пересечение которых содержит половину используемых слов, будут иметь схожесть ~50%. Аналогично можно поступить со словами, для которых вектор будет представлен длиной, равной количеству букв в алфавите. К сожалению, такой подход не учитывает порядок слов (букв), и наличие частично одинаковых слов (букв) в предложениях (словах) не гарантирует выявления их семантической близости.

Отсюда следует, что для корректного выполнения процедуры векторизации в рассматриваемом случае необходимо использовать некоторый набор эвристических правил:

1. Необходимо учитывать порядок следования слов. Множество слов, которые предшествуют определенным словам в предложении или следуют за ними, называют его контекстом. Формально, если предложение $P$ представлено как набор слов с определенным порядком $P = \{w_1, w_2, \ldots, w_n\}$, то контекстом $C$, например, для слова $w_i$ является предложение без этого слова $C = P / w_i$.

Зачастую контекст ограничивают «окном» определенной длины, поскольку далеко отстоящие слова могут не оказывать влияния друг на друга, причем как слева, так и справа от зафиксированного слова берут одинаковое количество слов, т.е. $n/2$ (если это возможно). Например, возьмем предложение «Утром кофе помогает взбодриться» и ограничим длину окна двумя словами. Тогда получится следующий набор слов и соответствующих контекстов (табл.).

**Набор слов и соответствующих контекстов предложения «Утром кофе помогает взбодриться»**

| Слово | Контекст |
|---|---|
| Утром | Кофе помогает |
| Кофе | Утром помогает |
| Помогает | Кофе взбодриться |
| Взбодриться | Кофе помогает |

2. Для осуществления векторизации необходимо учитывать и тот факт, что некоторые слова являются служебными частями речи или не имеют своей ярко выраженной семантики (например, такие части речи, как союзы, предлоги), поэтому их необходимо исключить из контекста.

3. Поскольку одинаковые слова в предложении могут иметь разное окончание, необходимо провести процесс *лемматизации* – приведения слова к лемме – «нормальной» форме [3]. Например, слово «бежал» следует преобразовать в «бежать», а «кошками» – в «кошка».

*Метод word2vec*

Для предварительной обработки текста подходящим способом векторизации выглядит алгоритм *word2vec,* который формально определяется как способ перевода слов из словесного пространства в векторное [2; 5]:

$$word2vec: \ W \to V.$$

Важнейшим свойством данной модели является то, что слова, встречающиеся в похожих контекстах, имеют схожие векторы (близость определяется косинусной мерой). Поскольку алгоритм реализуется в виде нейронной сети, а количество параметров велико, найти решение становится возможным только с помощью численных методов. Поэтому для обучения сети используется метод обратного распространения ошибки на основе градиентного спуска [1].

Алгоритм *word2vec* может работать в двух вариантах:

1. **Common Bag of Words (CBoW).** При обучении модель использует контекст слова как входное значение и выдает наиболее подходящее для этого контекста слово.

2. **Skip-Gram.** Является инвертированной моделью CBoW, т.е. на основе входного слова модель предлагает наиболее вероятный контекст.

На рисунке 1 представлен общий вид модели *word2vec* для работы в режиме CBoW. На вход модели поступает *C* векторов, представляющих собой контекст.

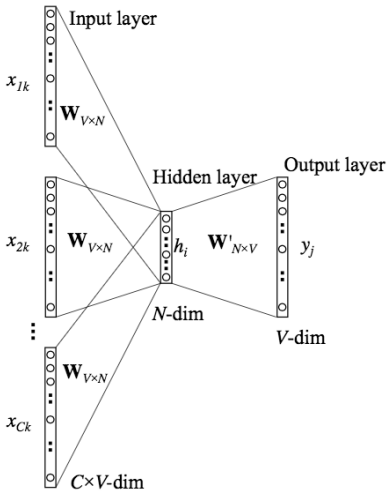


**Рис. 1.** Общий вид модели *word2vec* для работы в режиме CBoW

Здесь $x_{1k}, x_{2k}, \ldots, x_{Ck}$ – входные слова (контекст) размерности *V-dim*; $h_i$ – скрытое (латентное) представление слова в виде вектора размерности *N-dim*; $y_i$ – вероятность появления *i*-го слова в таком контексте.

Для обучения модели и реализации векторизации использовались библиотеки *gensim,* реализующая модель *word2vec,* и *nltk,* с помощью которой обрабатывался исходный текст для обучения.

*Визуальная оценка семантического сходства слов*

Для визуализации, предлагаемой в настоящей работе, нет необходимости использовать именно CBoW или Skip-Gram, поскольку наибольший интерес представляет именно промежуточное представление слов (в виде векторов), которые формируются при обучении самой модели.

Поскольку слова, имеющие схожий контекст, имеют и близкую семантику, то становится возможным визуализировать такие представления следующим образом:

1. Применяя алгоритм *word2vec,* представить слово *w* в пространстве $\mathbb{R}^n$.

2. Начертить горизонтальную полосу определенной толщины, функция которой будет состоять в отображении векторов слов с помощью цветовой палитры.

3. Разбить полосу на *n* равных отрезков, количество которых соответствует длине векторов слов (из эвристических соображений взято $n = 40$).

4. Сопоставить каждой компоненте вектора в порядке их следования отрезок на полосе.

5. Закрасить отрезки в соответствии с выбранной палитрой цветов, в которой каждый определенный цвет характеризует определенное число. В нашем случае числу 0 поставим в соответствие черный цвет, числу 3 – темно-серый, а –3 – светло-серый. Тогда числу 2, например, будет соответствовать темный оттенок серого, поскольку число 2 находится между числами 0 и 3.

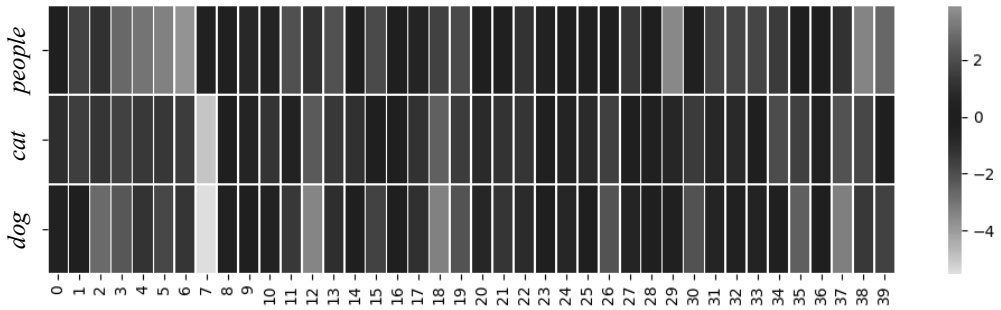


**Рис. 2.** Визуализация векторного представления слов

На рисунке 2 представлен пример визуализации векторов слов *people, cat* и *dog.* Такую визуализацию можно интерпретировать, в частности, следующим образом:

- компоненты с номерами 3, 6, 7, 13, 17, 26, 27, 30, 32, 35, 38 векторов «cat» и «dog» выглядят в цветовом отношении значительно похожими друг на друга и при этом существенно отличающимися от слова «people». Можно предположить, что эти компоненты

отображают (кодируют) нечто общее в смысловом отношении между двумя словами, которые, в частности, представляют два разных вида домашних животных;

- компоненты 8, 16, 36 во всех трех случаях имеют черный цвет: их значение близко к нулю, поскольку векторное представление для рассматриваемых слов слабо задействует данные компоненты;
- компонента 22 имеет близкие оттенки темно-серого цвета во всех трех словах, что, по-видимому, кодирует связь между домашними животными и человеком.

Как следует из анализа рисунка 2, предлагаемый способ визуализации в принципе позволяет оценить степень семантической близости слов. В частности, конкретный пример на рисунке 2 позволил продемонстрировать, что степень взаимной семантики пары слов «cat» и «dog» ощутимо больше, чем для каждой из них в сочетании со словом «people». Однако некоторые компоненты похожи во всех трех случаях, что можно интерпретировать как отражение связи человека с домашними животными.

Хотя смысловое значение компонент векторов для нас остается неясным (что обусловлено использованием сетевой модели), но на основе общей картины различия и сходства в смысловом содержании слов проявляется с очевидностью.

*Заключение*

Данные относительно семантической связи слов, полученные на основе цветовой визуализации их векторов, показывают, что такой подход, соответствующий наиболее полному отображению окружающего мира в нашем сознании, позволяет обеспечить разнообразие нюансов семантического сравнения языковых конструкций. Простота выполнения и наглядность сообщают предложенному методу признаки практической полезности.

*Автор благодарит профессора И.С. Клименко за полезное обсуждение.*

## Литература

1. *Горбань А.Н., Россиев Д.А.* Нейронные сети на персональном компьютере. Новосибирск: Наука, 1996. 276 с.
2. *Золотарев О.В., Шарнин М.М., Еромасова А., Тезадова Ф.М.* Современные подходы к обработке многоязычных текстов, основанные на методах дистрибутивной семантики // Сборник трудов международной научной конференции по физико-технической информатике – СРТ2018 (Пущино, 28–31 мая 2018 г.). Протвино, 2018. С. 43–47.
3. *Camacho-Collados J., Mohammad T.P.* On the Role of Text Preprocessing in Neural Network Architectures: An Evaluation Study on Text Categorization and Sentiment Analysis // Cornell University. URL: https://arxiv.org/abs/1707.01780 (date of the application: 23.08.2018).
4. *Mikolov T., Chen K., Corrado G., Dean J.* Efficient Estimation of Word Representations in Vector Space // Cornell University. URL: https://arxiv.org/abs/1301.3781 (date of the application: 16.01.2013).
5. *Mikolov T., Sutskever I., Chen K., Corrado G., Dean J.* Distributed Representations of Words and Phrases and Their Compositionality // Cornell University. URL: https://arxiv.org/abs/1310.4546 (date of the application: 16.10.2013).
6. *Molino P., Wang Y., Zhang J.* Parallax: Visualizing and Understanding the Semantics of Embedding Spaces via Algebraic Formulae // Cornell University. URL: https://arxiv.org/pdf/1905.12099.pdf (date of the application: 27.01.2020).

## Literatura

1. *Gorban' A.N., Rossiev D.A.* Nejronnye seti na personal'nom komp'yutere. Novosibirsk: Nauka, 1996. 276 s.
2. *Zolotarev O.V., Sharnin M.M., Eromasova A., Tezadova F.M.* Sovremennye podkhody k obrabotke mnogoyazychnykh tekstov, osnovannye na metodakh distributivnoj semantiki // Sbornik trudov mezhdunarodnoj nauchnoj konferentsii po fiziko-tekhnicheskoj informatike – SRT2018 (Pushchino, 28–31 maya 2018 g.). Protvino, 2018. S. 43–47.
3. *Camacho-Collados J., Mohammad T.P.* On the Role of Text Preprocessing in Neural Network Architectures: An Evaluation Study on Text Categorization and Sentiment Analysis // Cornell University. URL: https://arxiv.org/abs/1707.01780 (date of the application: 23.08.2018).
4. *Mikolov T., Chen K., Corrado G., Dean J.* Efficient Estimation of Word Representations in Vector Space // Cornell University. URL: https://arxiv.org/abs/1301.3781 (date of the application: 16.01.2013).
5. *Mikolov T., Sutskever I., Chen K., Corrado G., Dean J.* Distributed Representations of Words and Phrases and Their Compositionality // Cornell University. URL: https://arxiv.org/abs/1310.4546 (date of the application: 16.10.2013).
6. *Molino P., Wang Y., Zhang J.* Parallax: Visualizing and Understanding the Semantics of Embedding Spaces via Algebraic Formulae // Cornell University. URL: https://arxiv.org/pdf/1905.12099.pdf (date of the application: 27.01.2020).

DOI: 10.25586/RNU.V9187.20.01.P.099 УДК 519.64+004.75+004.94

Р.С. Хабаров, Ю.С. Фоменко

# МЕТОДИКА ОРГАНИЗАЦИИ ОБРАБОТКИ ДАННЫХ В ЕДИНОЙ ТЕРРИТОРИАЛЬНО-РАСПРЕДЕЛЕННОЙ ИНФОРМАЦИОННОЙ СИСТЕМЕ ДИСТАНЦИОННОГО ЗОНДИРОВАНИЯ ЗЕМЛИ

Представлена методика организации обработки данных с космических аппаратов дистанционного зондирования Земли в Единой территориально-распределенной информационной системе, основанная на оптимизации назначения смешанных приоритетов в сети массового обслуживания с учетом специфики обработки данных дистанционного зондирования Земли. Приведен пример реализации предложенной методики для обработки данных с космического аппарата «Ресурс-П».
*Ключевые слова:* обработка данных дистанционного зондирования Земли, сети массового обслуживания, оптимизация сетей массового обслуживания, многоканальные системы, приоритетное обслуживание, смешанные приоритеты.

R.S. Khabarov, Yu.S. Fomenko

# DATA PROCESSING ORGANIZATION METHODOLOGY IN THE UNIFIED TERRITORIALLY DISTRIBUTED EARTH REMOTE SENSING INFORMATION SYSTEM

A methodology for data processing organizing from Earth remote sensing spacecraft in the Unified Territorial Distributed Information System is presented. It is based on optimization of the mixed priorities