KROSHCHENKO A.A., GOLOVKO V. A., BEZOBRAZOV S.V., MIKHNO E.V., KHATSKEVICH M.V., MIKHNYAEV A.L., BRICH A.L. Deep training for detecting of objects at images of documents

This paper describes deep convolutional neural networks for objects detection and classification. A comparative analysis of various deep techniques and architectures for object detection are carried out. A neural network algorithm for marking up images of text documents was developed, based on preprocessing an image that simplifies the localization of individual parts of the document and subsequent recognition of localized blocks using a deep convolutional neural network. A program of semi-automatic segmentation has been developed that makes it easier to prepare a training data set for object detection and classification.

УДК 004.89

Крощенко А.А., Головко В.А., Безобразов С.В., Михно Е.В., Рубанов В.С., Кривулец И.Ю.

ОРГАНИЗАЦИЯ СЕМАНТИЧЕСКОГО КОДИРОВАНИЯ СЛОВ И ПОИСКОВОЙ СИСТЕМЫ НА ОСНОВЕ НЕЙРОННЫХ СЕТЕЙ

Введение. Задача семантического кодирования приобрела особую важность с развитием поисковых систем. Актуальность подобных технологий связана в первую очередь с возможностью осуществления поиска в больших по объему базах. При этом особое значение имеет не столько нахождение идентичных слов, сколько осуществление поиска близких по некоторой семантической метрике слов.

Интуитивно понятно, что близкие по смыслу слова в предложении должны появляться в одних и тех же или похожих контекстах. Под контекстом в данном случае понимаются слова, располагающиеся в непосредственной близости от рассматриваемого или, иначе, целевого слова. Именно эта идея положена в основу методов семантического кодирования (например, [1, 2]). Эти методы позволяют для словаря *D* фиксированного размера, слова которого представлены в некотором коде, выполнить его преобразование в код меньшей (редуцированной) размерности (рис. 1). Параллельно с этим, благодаря специфике реализации таких методов, происходит выделение семантически значимой информации, которая может быть использована для осуществления функций поиска.

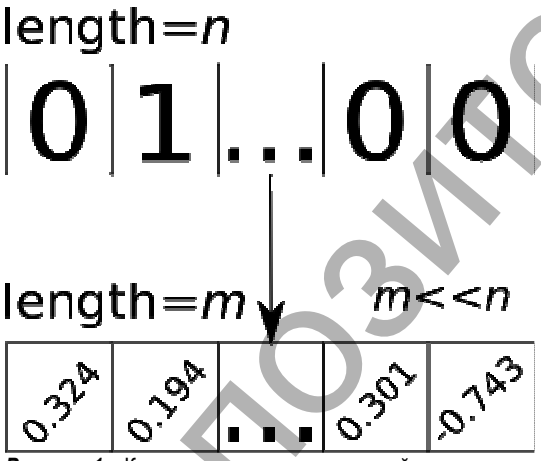

Рисунок 1 – Кодирование слов с редукцией размерности

В силу того, что слова в словаре некоторого языка почти всегда отличаются по длине, реализация какой-либо задачи сравнения слов существенно усложняется. Приведение же каждого слова словаря к вектору заданной размерности, одинакового по длине для всех слов, позволяет осуществлять сравнение искомого и проверяемого слов непосредственно путем вычисления любой (например, евклидовой метрики). Такая технология позволяет не только упростить задачи поиска, но и сделать такой поиск более интеллектуальным.

1. Метод word2vec. Одним из методов семантического кодирования, широко применяемых на практике, является word2vec. Этот подход был предложен Миколовым в 2013 году [1].

Word2vec позволяет осуществлять семантический анализ текста с выделением наиболее близких по смыслу слов. Существует два варианта метода word2vec (рис. 2), отличающихся политикой участия контекста. Под контекстом в данном случае понимается совокупность слов (слева и справа), окружающая целевое слово, взятая в пределах определенного окна.

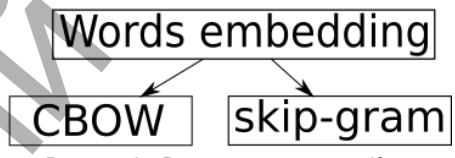

Рисунок 2 – Варианты метода word2vec

Первый вариант, называемый skip-gram, базируется на обучении нейросетевой модели, которая осуществляет формирование контекста на основе одного целевого слова, подаваемого на вход модели (рис. 3).

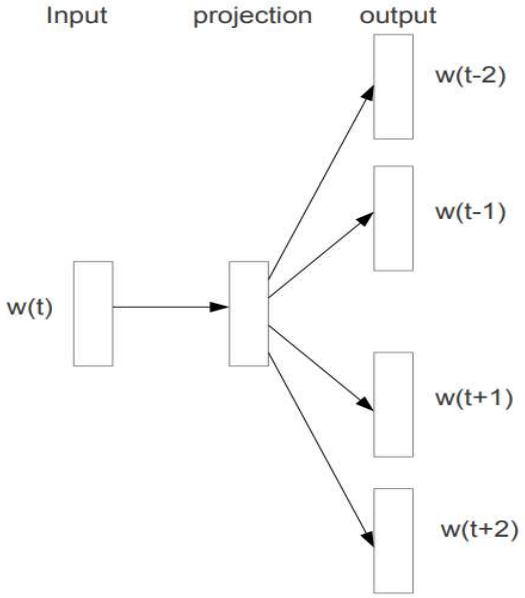

Рисунок 3 – Базовая модель НС метода word2vec (вариант skip-gram) [1]

Рубанов Владимир Степанович, *кандидат физ.-мат. наук, доцент кафедры высшей математики Брестского государственного технического университета.*
Беларусь, БрГТУ, 224017, г. Брест, ул. Московская, 267.
Кривулец Игорь Юрьевич, *аспирант кафедры информационных систем управления Белорусского государственного университета.*
Беларусь, БГУ, 220050, г. Минск, пр. Независимости, 4.

Второй вариант, называемый CBOW (Continuous Bag of Words), использует нейронную сеть для получения целевого слова на основе подаваемого контекста (рис. 4).

Следуя Миколову [1], вариант skip-gram чаще применяется в случае малой размерности обучающей выборки и позволяет хорошо представлять редкие слова и фразы. CBOW быстрее, чем skip-gram, и показывает лучшую точность для высокочастотных слов.

В основе обоих вариантов лежит использование простой по своей структуре поверхностной нейронной сети (фактически двуслойной).

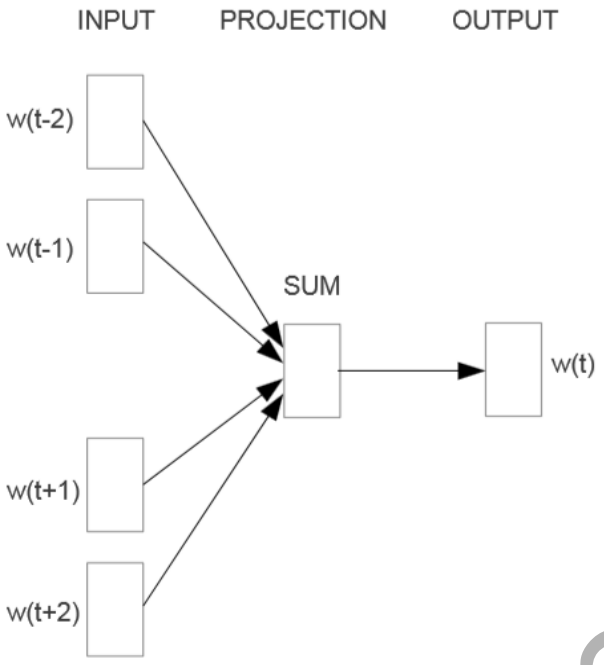

Рисунок 4 – Базовая модель НС метода word2vec (вариант CBOW) [1]

В результате применения word2vec обучается искусственная нейронная сеть, которая осуществляет отображение слова, записанного в виде унитарного кода (и, соответственно, принадлежащего словарю) в пространство меньшей размерности, которое впоследствии используется для оценки семантической близости слов. Полученное погружение может быть использовано для формирования списка семантически близких слов, а также предсказания семантических отношений. Например, **король для королевы** то же что **отец для ?**. В этом случае модель будет способна подобрать корректный ответ.

Перед непосредственным обучением нейросетевой модели выполняется предобработка текстовых данных и генерирование обучающей выборки в виде пар (**целевое_слово**, **контекстное слово**). Такое представление позволяет ускорить обучение нейронной сети.

Процесс предобработки текстовых данных можно условно разделить на следующие этапы:

1. Токенизация. Заключается в парсинге текста с его разбиением на слова и удалением знаков препинания и специальных символов (осуществляется смена регистра), **вход** – текст в формате txt, **выход** – список слов текста.

2. Опционально применяется удаление стоп-слов (т. е. наиболее распространенных и высокочастотных слов). В эту категорию могут попадать артикли, предлоги, частицы и т. д. **Вход** – список слов текста, **выход** – список с удаленными шумовыми словами. После этого этапа фактически формируется **словарь**.

3. Удаление низкочастотных слов. Часто подобные слова относятся к редко используемым либо к словам, записанным на языке, отличном от языка текста. Подобные слова заменяются на токен специального вида.

4. Предварительное кодирование слов. Производится, например, с помощью унитарного кодирования (one-hot-кодирования). Унитарное кодирование осуществляется формированием для каждого слова вектора размерности словаря, в котором устанавливается значение 1 для порядковой позиции слова в словаре и 0 – для всех остальных элементов). **Вход** – словарь. **Выход** – представление слов в виде one-hot-векторов.

Рассмотрим алгоритм применения каждого варианта метода skip-gram.

1.1 Вариант skip-gram метода word2vec. Обучение производится по следующему алгоритму.

0. Очистить словарь D.

1. Для каждого документа из обучающей выборки выполняется:

1.1 Если формат нетекстовый, распознать текст документа, иначе – переход на шаг 1.2. Вход – документ, выход – документ в формате txt.

1.2 Разбить полученный текст на слова и удалить знаки препинания и специальные символы. Вход – текст в формате txt, выход – список слов.

1.3 Удалить стоп-слова из списка слов. Вход – список слов документа, выход – список с удаленными шумовыми словами.

1.4 Добавить новые слова в словарь D.

1.5 Кодировать слова документа с помощью унитарного кода. Вход – словарь. Выход – матрица W, составленная из one-hot-векторов.

1.6 Выполнить проход по тексту документа с окном 2*window+1. Внести в обучающую выборку L центральное слово (w(t)) и контекстные слова (w(t-window),... , w(t-1), w(t+1), …, w(t+window)), представленные в виде унитарных векторов. Контекстные слова помещаются в список эталонных значений.

2. Для обучающей выборки *L* и словаря *D* обучить персептрон для модели skip-gram.

3. Сохранить полученные весовые коэффициенты.

1.2 Вариант CBOW метода word2vec. Обучение производится по следующему алгоритму.

0. Очистить словарь D.

1. Для каждого документа из обучающей выборки выполняется:

1.1 Если формат нетекстовый, распознать текст документа, иначе – переход на шаг 1.2. Вход – документ, выход – документ в формате txt.

1.2 Разбить полученный текст на слова и удалить знаки препинания и специальные символы. Вход – текст в формате txt, выход – список слов.

1.3 Удалить стоп-слова из списка слов. Вход – список слов документа, выход – список с удаленными шумовыми словами.

1.4 Добавить новые слова в словарь D.

1.5 Кодировать слова документа с помощью унитарного кода. Вход – словарь. Выход – матрица W, составленная из one-hot-векторов.

1.6 Выполнить проход по тексту документа с окном 2*window+1. Внести в обучающую выборку L центральное слово (w(t)) и контекстные слова (w(t-window),... , w(t-1), w(t+1), …, w(t+window)), представленные в виде унитарных векторов. Центральное слово помещается в список эталонных значений.

2. Для обучающей выборки L и словаря D обучить персептрон для модели CBOW.

3. Сохранить полученные весовые коэффициенты.

2 Решение задачи семантического кодирования. Разработка поисковой системы. Прототип поисковой системы, базирующейся на применении метода word2vec, представлен на рис. 5.

Как мы видим, этап функционирования системы тесно связан с этапом настройки ее параметров в ходе обучения соответствующей нейросетевой модели. В нашем случае Tranining phase включает в себя предобработку и обучение используемой сети на большом наборе документов. Эти документы могут быть представлены в различных форматах (в том числе в виде изображений). Поэтому этап предобработки может включать в себя распознавание отдельных слов и букв исходного текста, представленного в бинарном (несимвольном) формате. Для этого могут быть использованы как традиционные средства (например, OCR), так и нейросетевые модели

(например, глубокие сверточные нейронные сети). Также в этап предобработки могут входить операции, перечисленные в пункте 1. После выполнения предобработки осуществляется обучение соответствующей нейросетевой модели. В ходе этого процесса периодически может осуществляться обращение к сформированному словарю. После обучения система готова к использованию.

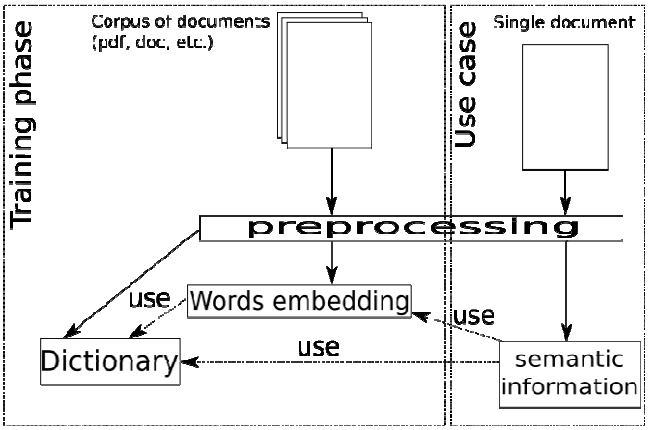

Рисунок 5 – Схема прототипа поисковой системы, построенной на основе word2vec

В фазе использования (Use case) обученная модель применяется для выявления семантической информации (а именно - определения близости различных слов текста к ключевому слову и поиска по образцу). Таким образом, возможно два варианта использования предложенной системы:

Поиск по ключевым словам

1. Подать на вход системы ключевые слова из словаря D.
2. Осуществить их преобразование с использованием сохраненных весовых коэффициентов. На выходе – преобразованные вектора.
3. Организовать сравнение по метрике для пула сохраненных в БД документов. В качестве метрики использовать евклидову норму. На выходе – результаты вычисления метрики для слов из каждого документа БД.
4. Выбрать документы, максимально соответствующие запросу.

Поиск по схожим документам

1. Подать на вход системы документ.
2. Выполнить его предобработку (как на стадии обучения). На выходе – матрица W, составленная из one-hot-векторов.
3. Выполнить умножение матрицы W на весовые коэффициенты, сохраненные на этапе обучения (прямое распространение). На выходе – векторы слов уменьшенной размерности.
4. Рассчитать сигнатуру документа, воспользовавшись данными из шага 3. Размерность итогового вектора-сигнатуры совпадает с размерностью векторов из шага 3. На выходе – сигнатура документа.
5. Выполнить поиск по базе документов с сохраненными сигнатурами.
6. Выбрать документы со схожими сигнатурами.

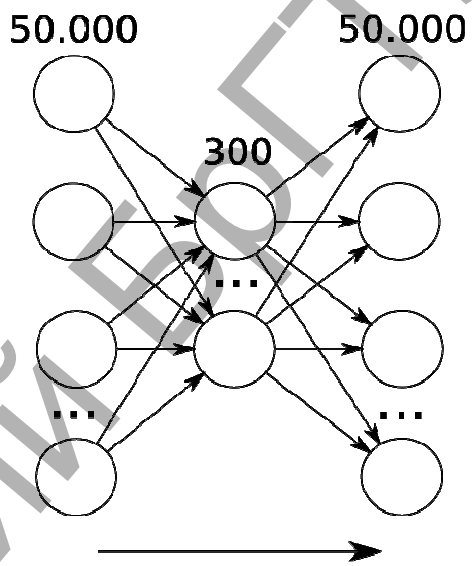

Рисунок 6 – Архитектура используемой нейронной сети (стрелка показывает направление сигнала)

Для иллюстрации работы метода word2vec приведем пример двумерной визуализации редуцированных кодов слов, полученной

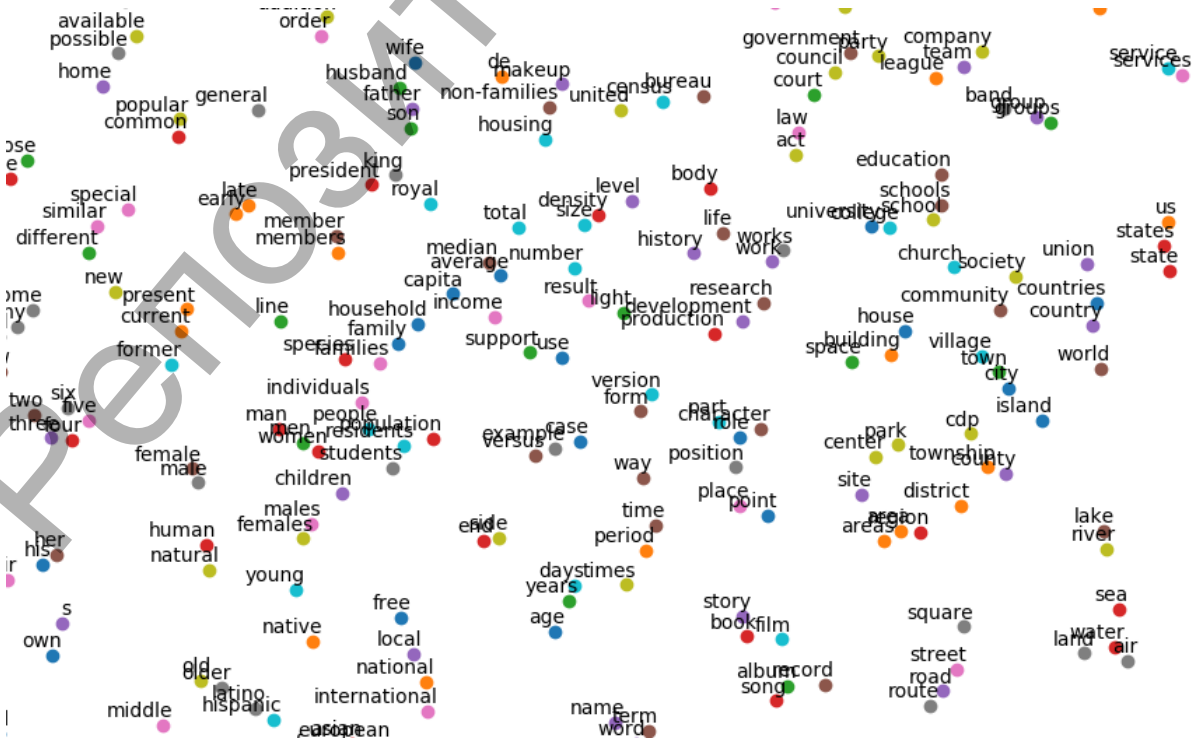

Рисунок 7 – Карта семантического сходства

нами для выборки из 100.000 англоязычных документов википедии и общего размера словаря 50.000 слов. В данном эксперименте использовалась упрощенная архитектура skip-gram, представленная на рис. 6 и включающая 50.000 входных нейронов, соответствующих целевому слову, 300 скрытых и 50.000 выходных нейронов, соответствующих контекстному слову. Структура представленной сети схожа с автоэнкодером, за исключением того, что в качестве эталонных значений используется контекстное слово, а не само целевое слово, как принято для автоэнкодеров. Автоассоциативные нейронные сети так же могут применяться для решения задачи семантического кодирования [3]. Таким образом, весовые коэффициенты первого слоя обученной нейронной сети представляют собой матрицу погружения для формирования редуцированного кода исходных слов.

При обучении были сформированы пары слов (целевое_слово, контекстное слово), которые подавались на нейронную сеть минибатчами по 128 пар в каждом. После обучения к редуцированным кодам слов был применен алгоритм t-SNE [4] для уменьшения размерности данных. Фрагмент полученная двумерная карта семантического сходства изображена на рисунке 7.

Приведенный рисунок иллюстрирует тот факт, что с помощью параметров обученной нейронной сети можно осуществлять поиск близких в семантическом плане слов, например, слова **lake**, **river**, **sea** и **water** попадают в одну группу, а слова **album**, **record**, **song** – в другую.

Базируясь на полученных результатах, можно реализовать систему, осуществляющую поиск близких слов и их подсветку в исследуемом тексте, результат работы которой для заданного url-адреса и списка искомых слов представлен на рис. 8.

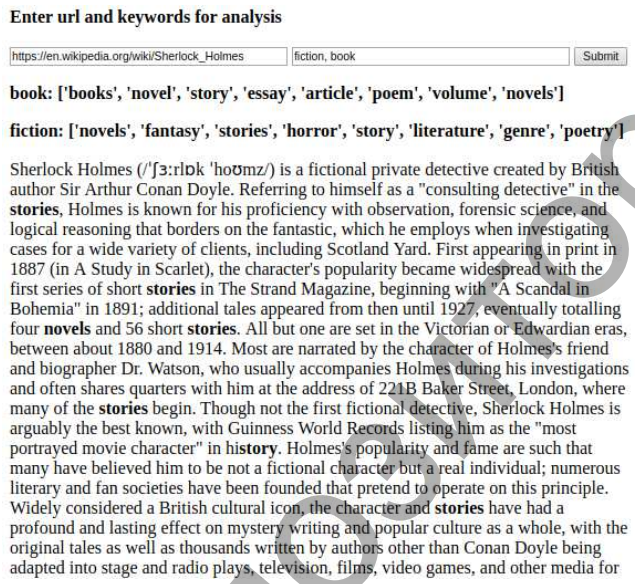

Рисунок 8 – Поисковая система (отображение результатов)

Помимо этого, метод word2vec может использоваться для формирования базы знаний в определенной предметной области (например, [5]) и извлечения семантических реляций в целом [6].

Заключение. В данной статье рассматривается и анализируется применение метода word2vec для решения задач семантического кодирования.

На основании полученных практических результатов разработан прототип поисковой системы, базирующейся на использовании выделенной семантической информации для осуществления релевантного поиска в базе документов. Предложено два основных сценария осуществления такого поиска. Осуществлена подготовка обучающей выборки на базе корпуса документов англоязычной версии Википедии, включающей более 100 тысяч оригинальных статей. Полученная выборка использовалась в экспериментальной части работы для проверки эффективности разработанного прототипа поисковой системы.

Так как применяемая структура нейронной сети для осуществления кодирования по методу word2vec является по сути поверхностной, остается открытым для исследований вопрос сравнительной характеристики word2vec и методов семантического кодирования, реализуемых с использованием глубоких нейронных сетей [7] и сверточных нейронных сетей.

СПИСОК ЦИТИРОВАННЫХ ИСТОЧНИКОВ

1. Mikolov, T. Efficient estimation of word representations in vector space / T. Mikolov, K. Chen, G. Corrado, J. Dean // arXiv [Web-resource]. – 2013. – Mode of access: https://arxiv.org/pdf/1301.3781.pdf. – Date of access: 12.12.2017.
2. Pennington, J. GloVe: Global Vectors for Word Representation / J. Pennington, R. Socher, and C. Manning // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). – 2013. – P. 1532–1543.
3. Головко, В.А. Семантическое кодирование на основе глубоких автоассоциативных нейронных сетей / В.А. Головко, А.А. Крощенко // Open Semantic Technologies for Intelligent Systems: материалы VI Международной научно-технической конференции «Открытые семантические технологии проектирования интеллектуальных систем» – Минск : БГУИР, 2016. – С. 313–318.
4. Van der Maaten, L. Visualizing High-Dimensional Data Using t-SNE / L.J.P van der Maaten, G.E. Hinton // Journal of Machine Learning Research. – Volume 9. – 2008. – P. 2579–2605.
5. Xiong, S. Deep Knowledge Representation based on Compositional Semantics for Chinese Geography / S. Xiong, X. Wang, P. Duan, Z. Yu and A. Dahou // In Proceedings of the 9th International Conference on Agents and Artificial Intelligence. – Volume 2: ICAART. – 2017. – P. 17–23.
6. Pelevina, M. Making Sense of Word Embeddings / M. Pelevina, N. Arefyev, C. Biemann, A. Panchenko // arXiv [Web-resource]. – 2017. – Mode of access: https://arxiv.org/pdf/1708.03390.pdf. – Date of access: 12.12.2017.
7. Головко, В.А. Применение нейронных сетей глубокого доверия для выделения семантически значимых признаков / В.А. Головко, А.А. Крощенко // Open Semantic Technologies for Intelligent Systems: материалы V Международной научно-технической конференции «Открытые семантические технологии проектирования интеллектуальных систем» – Минск : БГУИР, 2015. – С. 481–486.

Материал поступил в редакцию 13.02.2018

KROSHCHENKO A.A., GOLOVKO V. A., BEZOBRAZOV S.V., MIKHNO E.V., RUBANOV V.S., KRIVULETS I.Yu. The organization of semantic coding of words and search engine on the basis of neural networks

In this paper we investigate applying of word2vec for solution of semantic hashing task. Based on received results we developed prototype of searching system, which uses extracted semantic information for relevant search in base of documents. We proposed two main scenarios for execution such search. We prepared training database based on corpus of 100.000 original documents from Wikipedia. Received database has used in experimental part of work for testing effectiveness of developed search system prototype