OSTIS-2016
(Open Semantic Technologies for Intelligent Systems)

УДК 004.822:514

НЕЙРОННЫЕ СЕТИ В СЕМАНТИЧЕСКОМ АНАЛИЗЕ

Аверкин А.Н.[*], Ярушев С.А.[**]

[*]*Вычислительный центр им. А.А. Дородницына*
Российской академии наук
Федерального исследовательского центра
«Информатика и управление»
Российской академии наук,
г. Москва, Россия
averkin2003@inbox.ru

[**]*Международный университет природы, общества и человека «Дубна»,*
г.Дубна, Московская область, Россия
Sergey.Yarushev@icloud.com

В работе проводится исследование по возможностям применения искусственных нейронных сетей в семантическом анализе. Рассматривается современное состояние дел в данной отрасли, перспективы использования искусственного интеллекта в области семантического анализа, направления и тенденции развития науки в данном направлении. Приводится обзор некоторых работ в области семантического анализа с применением технологий искусственного интеллекта и нейронных сетей.
Ключевые слова: семантический анализ, нейронные сети, искусственный интеллект, прогнозирование.

Введение

Обработка информации на естественном языке, анализ взаимосвязей между коллекцией документов и терминами, представленными в данном документе, понимание и определение направления и тематики текста – все это задачи семантического анализа. Латентно-семантическим анализом пользуются поисковые гиганты, чтобы находить тексты одной тематики. Множество работ ведется в области построения семантических моделей для обработки качества текста, понимания логических взаимосвязей, оптимизации баз знаний, а также широчайшего спектра задач.

Как дети учат язык? В частности, как они связывают структуру предложения с его значением? Данный вопрос непременно связан с более глобальным вопросом – каким образом мозг связывает последовательность символов для построения символических и суб-символических представлений? Многие ученые проводят исследования, чтобы получить ответы на данные вопросы.Одним из самых значимых аспектов в задаче обработки естественного языка является скорость, с которой происходит эта обработка. Наиболее четко эта проблема проявляется в исследованиях потенциальных способностей мозга. Наиболее подходящим подходом к моделированию данных процессов является использование нейронных сетей. Подтверждением этого может быть большое количество работ в данной тематике.

В данной работе рассматривается ряд задач в области семантического анализа с применением искусственных нейронных сетей.

1. Обзор исследований по использованию нейронных сетей в задачах классификации текста

Китайские ученые [BoYoetal., 2007] в своем исследовании представили новую модель классификации текста с использованием нейронной сети и её обучением методом обратного распространения ошибки, а также с модифицированным методом. Используется эффективный метод выбора характеристик для уменьшения размерности выборки, что за собой несет повышение производительности системы. Стандартный алгоритм обучения по методу обратного распространения ошибки довольно медленно обучается, поэтому исследователи модифицировали данный алгоритм для увеличения

скорости обучения. Традиционные слово-сопоставления на основе классификации текста используют модель векторного пространства. Тем не менее, в данном подходе не учитываются семантические отношения между терминами, которые могут приводить к ухудшению точности классификации. Латентно-семантический анализ может преодолеть проблемы, вызванные использованием статистически полученных концептуальных индексов, а не только отдельных слов. Он создает концептуальные векторные пространства, в которых каждый термин или документ представляется в виде вектора в пространстве. Это не только значительно уменьшает размерность, но также позволяет обнаруживать важные ассоциативные отношения между терминами. Исследователи протестировали свою модель на наборе из 20 новостных данных, экспериментальные результаты показали, что модели с модифицированным методом обратного распространения ошибки, предложенным в данной работе, превзошли традиционный метод обучения. А также применение латентно-семантического анализа для данной системы позволяет резко сократить размерность, что позволяет достичь хороших результатов классификации.

Как видно из исследования, использование модифицированного алгоритма обучения нейронной сети, вкупе с семантическим анализом дает хорошие результаты в задаче классификации текста.

Индийские ученые [Tekwanietal., 2014] провели сравнение производительности обыкновенной нейронной сети обратного распространения ошибки с комбинацией данной нейронной сети с методом латентно-семантического индексирования в задаче классификации текста. В традиционной нейронной сети с обратным распространением ошибки, процесс настройки весов блокируется в локальном минимуме, а также, скорость обучения данного типа сетей довольно низка, что влечет за собой снижение производительности. В связи с данными недостатками, ученые решили сделать комбинацию латентно-семантического индексирования и данной нейронной сети. Латентно-семантические представления в структуре данных в низко-мерном пространстве, в котором документы, термы и последовательности слов также сравнивались. Одномерная декомпозиционная техника используется в латентно-семантическом анализе, в котором многомерные матрицы термов разбиваются в набор К ортогональных факторов, в которых оригинальные текстовые данные изменены до меньшего семантического пространства. Новый вектор документов можно найти в К-мерном пространстве. Так же, находятся новые координаты запросов.

Производительность комбинации данных методов проверялась на основе методики классификации 20 новостных групп из разных категорий, таких как спорт, медицина, бизнес, политика и др. В итоге, данные метод позволяет значительно снизить размерность и получить лучшие результаты классификации текста.

2. Нейронные сети в задачах обработки естественного языка

2.1. Естественные нейронные сети в обработке языка

Одна из ключевых фигур в исследованиях головного мозга человека является В. Маунткасл [Mountcastle V., 1997]. В данной работе, он обобщил свои многолетние исследования, он утверждает, что, несмотря на разнообразие своих функций, все разделы коры головного мозга устроены, в принципе, одинаково. Это означает, что обучение и распознавание образов в коре происходит единообразно, а разнообразие ее функций есть следствие разнообразия сигналов, обрабатываемых разными участками коры.

Согласно Маунткаслу, кора имеет двумерную ячеистую структуру. Базовым функциональным элементом коры является мини-колонка диаметром около 30 мкм, состоящая из примерно 100 нейронов. Такие мини-колонки связаны между собой положительными и отрицательными латеральными связями. Причем, последние включаются резко, но с неким запаздыванием относительно первых. В результате одновременно возбуждается целый пул соседних мини-колонок, невольно заставляя вспомнить самоорганизующиеся карты Т. Кохонена [KohonenT., 2001]. В итоге, повсюду в коре мы наблюдаем самоорганизующиеся карты признаков: детекторы схожих сигналов располагаются рядом друг с другом.

Эксперименты свидетельствуют, что площадь элементарных детекторов на этих картах порядка 0.1 мм2, т.е. они содержат 102 мини-колонок или 104 нейронов. Такие функциональные единицы Маунткасл называет макро-колонками. Именно они определяют «разрешающую способность» коры и предельное число признаков, которые может запомнить человек (всего несколько миллионов). Зато надежность этой памяти гарантируется большим числом нейронов, составляющих макро-колонку. Так что мы сохраняем свою память на протяжении всей жизни даже при гибели существенной части нейронов.

Таким образом, карты Кохонена являются, по-видимому, наиболее подходящим инструментом для моделирования работы коры. Надо только научить их работе с динамическими паттернами, с которыми только и работает мозг, т.к. его основная задача – предвидение.

2.2. Исследования в задачах обработки языка и предложений

Как человек овладевают языком, а также двумя или более разными языками с одной нервной

системой, до сих пор остается открытым вопросом. Для решения данной проблемы, французские ученые, во главе с Питером Домни [X.Hinautetal., 2015] предложили построить модель, которая будет способна изучать любой язык с самого начала. В данной работе они предлагают нейросетевой подход, который обрабатывает предложения по словно, слово за словом без предварительного знания семантики слов. Предлагаемая модель не имеет «предварительно связанную» структуру, а только случайную и обученные соединения, модель основана на технологии ReservoirComputing. Ранее учеными была разработана модель для робототехнических платформ, благодаря которой, пользователи могут научить робота основам английского языка, чтобы в дальнейшем давать ему различные задания. В данной работе была добавлена способность обрабатывать редкие слова для того, чтобы можно было сохранить размер словаря довольно маленьким при обработке естественного языка. Более того, данный подход был распространен на Французский язык и показано, что нейронная сеть может изучать два языка одновременно. Даже при небольшом корпусе языка, модель способна обучаться и обобщать знания в условиях моноязычности или двуязычности. Данный подход может быть более практичной альтернативой для небольших корпусов различных языков чем другие обучающие методы, опирающиеся на наборы больших данных.

Множество исследований проводится в области обработки языка с помощью нейронных сетей [Miikkulainen R., 1996], а также в последнее время с использование так называемых EchoStateNetworks[Frank, 2006].

Как человеческий мозг обрабатывает предложения, которые человек читает или слышит? Задача понимания того, как мозг это делает, занимает одно из центральных мест в исследованиях ученых из данной области. Обработка предложений происходит в режиме реального времени. Предыдущие слова в предложении могут влиять на время обработки в сотни миллисекунд. Последние нейрофизиологические исследования позволяют предположить, что именно лобная часть головного мозга играет решающую роль в этом процессе. XavierHinaut [X.Hinautetal., 2013] провел исследование, которое дает некоторое понимание того, как определенные аспекты в данной обработке предложений в реальном времени происходят, основываясь на динамике периодических корковых сетей и пластичности в кортико-полосатой системе. Они моделируют префронтальную область BA47 при помощи рекуррентной нейронной сети, получая он-лайн вход категорий слов в процессе обработки предложений. Система обучается по парам предложений, в которых закодирован смысл как функция активации, соответствуя той роли, которую играют глаголы и существительные в предложениях. Модель изучает расширенный набор грамматических конструкций и демонстрирует возможность для генерации новых конструкций. Это демонстрирует, на сколько рано в предложении параллельный набор предикатов создает смысл. Модель демонстрирует, как он-лайн отклики на слова подвержены влиянию от предыдущих слов в предложении и предыдущие предложения в дискурсе, обеспечивающие новый взгляд на нейрофизиологию коры головного мозга для распознавания грамматической структуры. Исследование показало, что рекуррентные нейронные сети могут декодировать грамматическую структуру из предложений в реальном времени с целью получения представления о значении предложений. Это может обеспечить понимание основных механизмов кортико-полосатой функции головного мозга человека в обработке предложений.

Нейросетевая обработка естественного языка. Центральное внимание в когнитивной науке сегодня сконцентрировано на исследовании того, как нейронные сети в головном мозге используются для чтения и понимания текста. Данный вопрос исследуется огромным количеством ученых по нейрофизиологии на ряду с недавними исследованиями, которые призваны обследовать процессы головного мозга, вовлеченные в обучение неязыковых последовательностей или искусственного обучения грамматике. PeterFordDominey [Domineyetal., 2008] в своем исследовании предпринял попытку совместить данные с несколькими нейрофизиологическими моделями обработки предложений, через спецификации нейросетевой модели, архитектура которой основана на известной кортико-стриато-таламо-кортикальной (КСТК) нейроанатомии системы человеческого языка. Задача состоит в том, чтобы разработать имитационные модели, учитывающие ограничения и нейроанатомической связи, и функциональные данные изображений. В предлагаемой модели, структурные кии, закодированные в рекуррентной кортиковой нейронной сети в BA47, активируют схему (КСТК) для модулирования потока лекскической семантической информации в интегрированное представление смысла на уровне предложений. Результаты моделирования продемонстрированы в работе Caplan [Caplan D. etal.,1985].

Моделирование органа языка провел С. А. Шумский. В своей работе [Шумский, 2012], автор выдвигает три гипотезы: **первая гипотеза** состоит в том, что обработка временных рядов в коре осуществляется подобными модулями, распознающими типовые временные паттерны, каждый в своем входном потоке. Например, участок коры, ответственный за морфологический анализ слов, распознает порядка 10^5 слов и составляющих их морфем и слогов. Другой участок коры, определяющий структуру предложений, работает таким же образом, только с другим первичным алфавитом, каждый символ которого кодирует уже не букву, а целое слово. Этот участок запоминает характерные паттерны комбинирования слов в

грамматически правильные фразы. Согласно **второй гипотезе**, входом для следующего коркового модуля, ответственного за анализ временных структур более высокого порядка, служит сжатый таламусом выходной сигнал от предыдущего модуля. По **третьей гипотезе**, в «органе языка» существуют два взаимосвязанных канала «глубокого обучения»: грамматический и семантический. Аналогично дорсальному (анализ сцен) и вентральному (распознавание объектов) каналам анализа зрительной информации.

Для проверки данных гипотез, был создан программный комплекс «семантический процессор *Голем*», способный выявлять иерархии языковых паттернов при обучении на больших текстовых массивах. Обучение проводилось на текстовом массиве объемом 6 ГБ, состоящем из материалов русскоязычных интернет-СМИ. Чтобы приблизить условия эксперимента к обучению устной речи ребенком, все слова приводились к строчным буквам. Объем обучающей выборки примерно соответствует языковому опыту 20-летнего человека (при восприятии $\sim 10^5$ слов в день). Обучение заняло около двух месяцев работы современного ПК.

В результате, разработанный Шумским комплекс довольно уверенно может распознавать имена, фамилии, города, страны и некоторые другие понятия. Также, удалось добиться понимания и того, какие понятия в данном предложении соотносятся с какими, можно сказать, что Голем способен достаточно адекватно распознавать и индексировать смысловое содержание предложений.

Заключение

В данной работе проведен обзор современных работ в области исследования современного языка, исследования по изучению работы головного мозга и понимания им языка. Каждое исследование, содержит применение нейросетевых технологий в задачах семантического анализа и моделирования работы головного мозга.

Исходя из результатов, полученных в каждом исследовании, можно сделать вывод, что нейронные сети в задачах семантического анализа показывают высокую производительность и расширяют возможности в анализе текстовых данных, а также являются незаменимой технологией в задачах моделирование мозговой деятельности, в частности моделирование обучения новым языкам и применение данных технологий в построении роботов, способных самостоятельно изучать языки и понимать их смысл.

Работа выполнена при поддержке гранта: РФФИ №14-07-00603

Библиографический список

[Шумский, 2012]Шумский, С. А. Мозг и язык: Гипотеза о строении «Органа языка».

[Bo Yo et al., 2007] Yu B., Xu Z., Li C. Latent semantic analysis for text categorization using neural network //Knowledge-Based Systems. – 2008. – Т. 21. – №. 8. – С. 900-904.

[Domineyetal., 2008]Peter Ford Dominey, Toshio Inui, Michel Hoen. Neural network processing of natural language: II. Towards a unified model of corticostriatal function in learning sentence comprehensionand non-linguistic sequencing// Brain and Language. – 2008 doi:10.1016/j.bandl.2008.08.002.

[Frank, 2006] Frank, S. L. (2006). Strong systematicity in sentence processing by an Echo State Network. In Proc. of ICANN 2006, pp. 505–514.

[X. Hinaut et al., 2015]Hinaut X. et al. A Recurrent Neural Network for Multiple Language Acquisition: Starting with English and French.

[X. Hinaut et al., 2013] Hinaut X., Dominey P. F. Real-time parallel processing of grammatical structure in the fronto-striatal system: a recurrent network simulation study using reservoir computing //PloS one. – 2013. – Т. 8. – №. 2. – С. e52946.

[Kohonen T., 2001]Kohonen T. Self-Organizing Maps.SpringerVerlag. 2001.

[MiikkulainenR., 1996]Miikkulainen, R. (1996) Subsymbolic case-role analysis of sentences with embedded clauses. CognitiveSci 20: 47–73.

[Mountcastle V., 1997]Mountcastle V. The columnar organization of neocortex // Brain. 1997. V. 120. P. 701–722

[Caplan D. et al.,1985]Caplan, D., Baker, C., &Dehaut, F. (1985).Syntactic determinants of sentence comprehension in aphasia. Cognition, 21, 117–175.

NEURAL NETWORKS IN SEMANTIC ANALYSIS

Averkin A.N.[*], Yarushev S.Y.[**]

[*]*Institution of Russian Academy of Sciences Dorodnicyn Computing Centre of RAS, Moscow, Russia*

averkin2003@inbox.ru

[**]*Moscow region State Educational Institution for higher professional education Dubna International University for Nature, Society and Man, Dubna, Russia*

Sergey.Yarushev@icloud.com

In this paper we study the possibilities of application of artificial neural networks in the semantic analysis. The current state of affairs in the industry, the prospects for the use of artificial intelligence in the field of semantic analysis, trends and tendencies of development of science in this direction. A review of some of the works in the field of semantic analysis with the use of artificial intelligence and neural networks.

Conclusion

Based on the results obtained in each study, we can conclude that the neural networks in the problems of semantic analysis shows high performance and extend the capabilities of the analysis of the text data.Alsoitindispensable technology for modeling brain activity, in particular modeling of learning new languages and the use of Information technologies in building robots that can independently learn languages and understand their meaning.