X

УДК 004.942, 514.18, 519.652

А. А. Харламов[1], Д. И. Гордеев[2]
[1]Институт высшей нервной деятельности и нейрофизиологии РАН, г. Москва
117485, г. Москва, ул. Бутлерова, 5а
[1]Московский государственный лингвистический университет, г. Москва
119034, г. Москва, ул. Остоженка, 38, стр. 2
[1]Высшая школа экономики, г. Москва
101000, г. Москва, ул. Мясницкая, д. 20
[2]Институт прикладных и экономических исследований РАНХиГС
(Российская академия народного хозяйства и государственной службы
при Президенте Российской Федерации), г. Москва
119571, г. Москва, проспект Вернадского, 82-84, корпус 9, офис 1805

ДИСТРИБУТИВНАЯ VS СЕТЕВАЯ СЕМАНТИКА В ДИАЛОГОВЫХ СИСТЕМАХ

A. A. Kharlamov[1], D. I. Gordeev[2]
[1]Institute of Higher Nervous Activity and Neurophysiology, Russian Academy of Sciences, Moscow
117485, Moscow, st. Butlerova, 5a
[1]Moscow State Linguistic University, Moscow
119034, Moscow, st. Ostozhenka, 38, p. 2
[1]Higher School of Economics, Moscow
101000, Moscow, st. Myasnitskaya, d. 20
[2]Institute of Applied and Economic Research, RANEPA
(Russian Academy of National Economy and Public Administration
under the President of the Russian Federation), Moscow
119571, Moscow, Vernadsky Avenue, 82-84, building 9, office 1805

DISTRIBUTIVE VS NETWORK SEMANTICS IN DIALOG SYSTEMS

В последние 8 лет возобновился повышенный интерес к сфере диалоговых агентов. Во многом это связано с внедрением машинного обучения в задачи по автоматической обработке естественного языка. Использование средств дистрибутивной и сетевой семантики позволяет использовать обобщенные данные из огромных корпусов текстов, что было более проблематичным при использовании n-грамм. Также новые языковые модели, обученные на огромных корпусах, позволяют существенно сократить затраты на дообучение моделей для новых задач (transfer learning), а в ряде случаев и вовсе обойтись без него (zero-shot learning).
Ключевые слова: диалоговые агенты; NLP; natural language processing; нейронные сети; глубокое обучение; дистрибутивная семантика; диалоговые агенты; Word2Vec; Elmo; Bert.

In the last 8 years, increased interest in the field of dialogue agents has resumed. This is largely due to the introduction of machine learning in the tasks of automatic processing of the natural language. The use of means of distributive and network semantics allows using generalized data from huge corpus of texts, which was more problematic when using n-grams. Also, new language models, trained in huge buildings, can significantly reduce the cost of additional training for new tasks (transfer learning), and in some cases even do without it (zero-shot learning). In addition, this chapter examines the established neural network architectures and promising approaches to the use of neural networks for the tasks of automatic processing of the language in general, and interactive agents in particular. Also provides an overview of the modular approach to interactive agents. The main types of modules are considered.
Keywords: interactive agents; NLP; natural language processing; neural networks; deep learning; distributive semantics; conversational agents; Word2Vec; Elmo; Bert.

Введение

Цель работы – в данной статье рассматриваются устоявшиеся нейросетевые архитектуры и перспективные подходы к использованию нейронных сетей для задач автоматической обработки языка в целом и диалоговых агентов в частности. Также приводится обзор модульного подхода к диалоговым агентам. Рассматриваются основные виды модулей.

Диалоговые системы (виртуальные собеседники, чат-боты и т.д.) ведут свою историю с самой зари появления компьютеров и основываются на идеях Тьюринга 40-х – 50-х годов о создании разумных машин. Первым виртуальным собеседником считается программа ELIZA. Идея диалога, реализованного в этой программе, основывалась на нахождении ключевых слов в вопросах собеседника. Данная область долго стагнировала. Например, программа A.L.I.C.E., выигравшая в 2004 году приз Лёбнера как лучший виртуальный собеседник, базировалась на той же технологии, что и ELIZA (рис. 1) (поиск ключевых слов и сравнение с базой данных). Однако в последние десять лет наметился некоторый прогресс в этой области. Так, лучший чат-бот 2007 года UltraHal использовал не только поиск по ключевым выражениям, но и базу данных WordNet, а победивший в 2012 ChipVivant и вовсе отказался от обширной базы сообщений (http://www.chipvivant.com/about/). Однако эти виртуальные собеседники общей тематики не привлекают внимания из-за ограниченности коммерческого использования и общих недостатков. Поэтому бизнес ориентируется на системы более узкого профиля и на системы извлечения информации (*Information retrieval*).

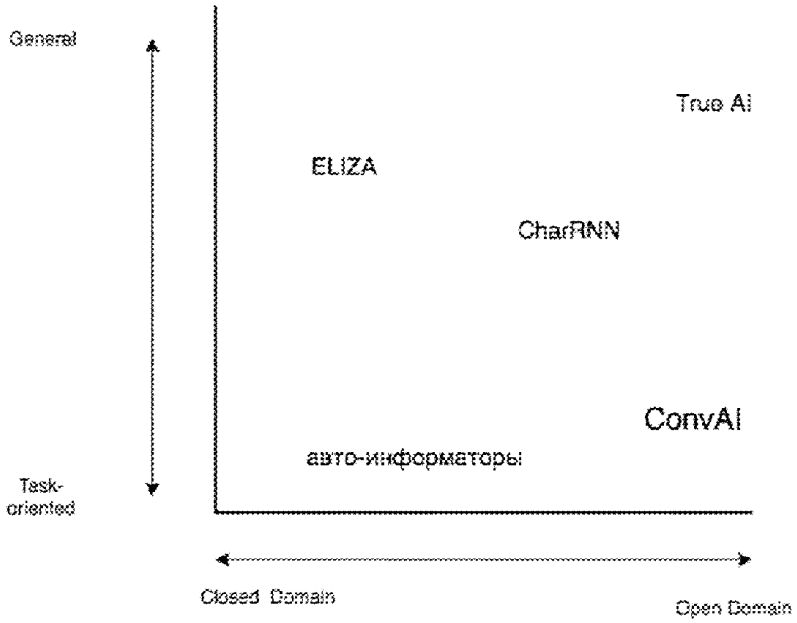


Рисунок 1 – Распределение некоторых известных диалоговых систем в пространстве «Задание – предметная область» (https://habrahabr.ru/company/mipt/blog/330228/ – пост Валентина Малых)

Ведение содержательного диалога с компьютером в широком смысле является пока нерешенной задачей. Большинство устоявшихся коммерческих подходов (например, в системах расписаний) ориентируется на нахождение некоторых именованных сущностей (NER – *named entity recognition* – распознавание именованных сущностей). Часто под этими сущностями понимаются различные имена собственные. Способы нахождения сущностей могут сильно различаться – это может быть как поиск по

ключевым словам и поиск, основанный на вручную написанных правилах, так и некоторые статистические алгоритмы. Все эти способы требуют кропотливой человеческой работы по идентификации закономерностей. Преимуществом статистических методов является то, что для аннотирования обучающего материала требуется более низкая квалификация участников проекта, в то время как для написания правил требуются лингвисты высокого уровня. Также создание аннотированного корпуса гораздо лучше масштабируется. Проведение аннотирования не требует знакомства с уже существующими аннотированными фрагментами выборки (или требует минимального знакомства), чего не скажешь о правилах, список которых после некоторого предела сложно синхронизировать на предмет взаимной корректности, что требует дополнительных организационных расходов.

Одним из наиболее популярных алгоритмов, используемых для проведения NER, является CRF (*conditional random fields*), различные модификации которого до сих пор остаются конкурентоспособными. «Метод CRF имеет двух непосредственных предшественников, от каждого из которых он унаследовал часть свойств. Прежде всего, это метод скрытых марковских моделей (HMM), которые успешно используются для моделирования последовательностей, а также метод моделей максимальной энтропии (MaxEnt). CRF рассматривает условное распределение (y|x) последовательности меток y in Y, где вектор x in X состоит из наблюдаемых элементов, и наряду с MaxEnt, принадлежит к категории дискриминативных методов. Из наблюдаемых и выходных элементов конструируется набор бинарных функций-признаков (featurefunctions, potentialfunctions, factors), которые могут задаваться произвольно, и включать в себя любое количество элементов» [1]. Данный метод может быть использован для решения разных задач классификации текстов на естественном языке, однако наиболее популярен он для определения частей речи и именованных сущностей.

Начиная с 2012 года, наметился прорыв в обработке текстов на естественном языке из-за появления глубоких нейронных сетей. Однако они не улучшили кардинально результатов: так преимущество современных нейросетевых подходов по сравнению с CRF из Stanford NLP (https://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf) составляет около 3% (89% против 92%) для корпуса ConLL-2003. Кроме того, качество работы искусственных нейронных сетей (ИНС) сильно зависит от обучающих данных, которые должны быть вручную проаннотированы экспертами.

Из недостатков всех статистических методов NER стоит отметить, что они сильно зависят от объема и качества обучающей выборки, которая редко учитывает динамику объективной реальности. Кроме того, максимальное количество классов для корпусов с открытым доступом составляет только 7, что представляет некоторую сложность при необходимости более подробной классификации. Использование нерепрезентативного корпуса (большинство из них являются выборками из газетных статей) приводит к ошибкам в других сферах и контекстах.

Необходимо отметить, что информации об именованных сущностях недостаточно для современных диалоговых систем. Для ведения осмысленного диалога, не ограничивающегося заранее предопределенными вариантами ответов, необходимо понимание прагматической стороны запросов. Для этого требуется нахождение других сущностей, часто на жаргоне называемых «интентами» (*intents*) – «намерение собеседника», «прагматика сообщения». Данные сущности показывают прагматическое содержание анализируемого сообщения. Интенты обычно выделяются на основе интересующей пользователя предметной области и часто соотносятся с теми или иными операциями. Пониманию прагматики сообщений способствует наличие базы данных, где тому или иному запросу соответствует определенная операция, ранее соотнесенная человеком-оператором с запросом (например, проверка баланса для мобильных операторов, проверка номера страхового полиса для страховых организа-

ций, список транзакций за последний месяц для банковских организаций). При наличии базы данных (БД) достаточно использовать некий классификатор для определения этих категорий. На вход классификатору могут подаваться триграммы, отдельные слова, n-граммы, вектора слов (например, word2vec) или даже отдельные символы. Классификатор может основываться на современных нейросетевых подходах, учитывающих контекст (рекуррентные, сверточные нейронные сети, LSTM, Transformer-XL). Однако без наличия базы данных эта классификация представляется маловыполнимой.

Многие существующие коммерческие системы «понимания речи» часто ограничиваются этими двумя технологиями (определение интента и именованных сущностей). Так, LUIS от Microsoft (www.luis.ai/home) предлагает только эти виды информации, которая потом может быть использована чат-ботами (рис. 2).

Рисунок 2 – Пример ответа сервиса LUIS

Google Cloud Natural Language API предлагает более обширную информацию (синтаксис, анализ тональности и категорий), однако для диалоговых систем во многом данный сервис сводится к анализу тональности, а стоимость данного сервиса достаточно высока (рис. 3).

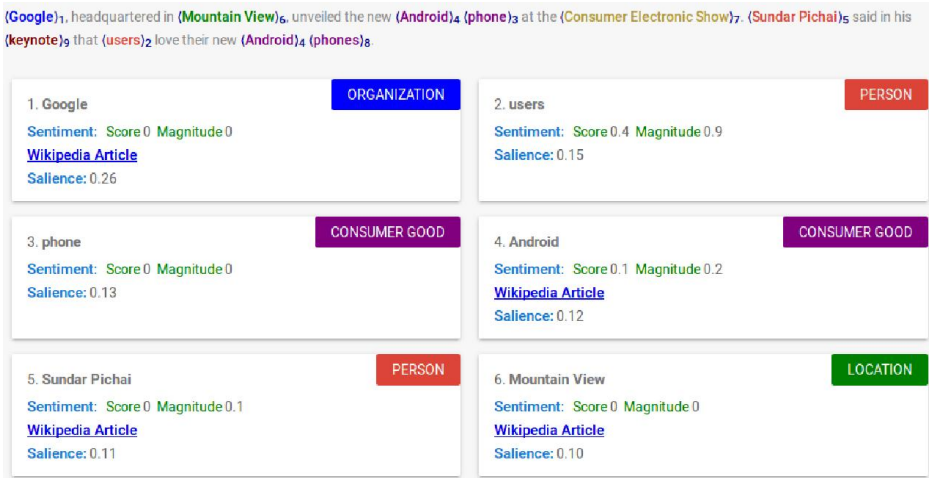


Рисунок 3 – Пример выдачи GoogleCloud API

Узкие места примерно те же, что и у NER – то есть сильная зависимость от аннотированных корпусов. Однако для уже состоявшихся компаний с опытом использования людей-операторов эта задача сильно упрощается в случае правильно сохраненной базы данных.

Потенциально, в диалоговых системах могут быть использованы различные методы представления лингвистических и экстралингвистических знаний, о которых пойдет речь ниже. Все они связаны, в основном с различными нейросетевыми парадигмами. Основное различие между ними заключается в отнесении соответствующей парадигмы к одному из двух классов подходов моделирования текстовой информации – векторному и сетевому.

Существующие обзоры

Существующие в данный момент обзоры позволяют понять основные тенденции, имеющиеся в направлении разработки искусственно-интеллектуального диалога вообще, и в анализе текстов, в частности.

Так, обзор «McKinsey Global Institute. Artificial Intelligence The next digital Frontier?» [2] посвящен общему анализу будущего искусственного интеллекта (ИИ); в том числе затрагивается и автоматическая обработка текстов на естественном языке (*Natural Language Processing* – NLP). В нем авторы уделяют внимание возможности применения ИИ по отраслям (розничная торговля, электроэнергетика, промышленное производство, здравоохранение, образование), возможности автоматизации различных профессий. Делается упор на важность NLP в сферах, где в качестве пользователей участвуют люди.

В докладе Mizuho Industry Focus [3] страницы с 27-29 посвящены системе Watson и в частности диалоговым системам, посвященным выдаче рекомендаций родителям по кормлению детей.

В докладе Natural Language Processing: Enterprise Applications for Natural Language Technologies [4] рассматривается большое количество отраслей, где может быть применена автоматическая обработка текстов на естественном языке.

В том числе, рассматриваются вопросы глубокого обучения, анализа текстов, дистрибутивная и сетевая семантика.

Глубокое обучение

Глубокое обучение характеризуется как класс алгоритмов машинного обучения, который:

1. Использует многослойную систему нелинейных фильтров для извлечения признаков с преобразованиями. Каждый последующий слой получает на входе выходные данные предыдущего слоя (или от нескольких предыдущих слоев, например, в ResNet). Система глубокого обучения может сочетать алгоритмы обучения с учителем и без учителя, при этом анализ образца представляет собой обучение без учителя, а классификация – обучение с учителем.

2. Обладает несколькими слоями выявления признаков или параметров представления данных (обучение без учителя). При этом признаки организованы иерархически, признаки более высокого уровня являются производными от признаков более низкого уровня.

3. Является частью более широкой области изучения представлений данных на основе машинного обучения.

4. Формирует в процессе обучения слои на нескольких уровнях представлений, которые соответствуют различным уровням абстракции в данной предметной области; слои образуют иерархию понятий.

Глубинное обучение совершило прорыв в ряде сфер науки, особенно в сфере распознавания изображений и распознавания речи. В области обработки текстов на

естественном языке успехи не столь велики по сравнению с другими методами машинного обучения (так, отмечается, что тот же word2vec, соответствующий вышеупомянутым методам обработки текстов, не всегда работает лучше базовых методов представления текстов, напр. TF-IDF), однако в большинстве разделов этого направления удалось превзойти ранее полученные результаты. Кроме того, на общей волне популярности сильно возросло внимание исследователей к глубинному обучению.

Глубокое обучение (глубинное обучение; англ. *Deeplearning*) относится к технологиям машинного обучения. Глубокое обучение – относительно новый по меркам науки качественный уровень технологий машинного обучения, характеризующий скачок (с 2006 года) в связи с непрерывным ростом вычислительной мощности компьютеров и накоплением исследовательского опыта. Многие методы глубинного обучения были известны и апробированы раньше. С достижением рубежа необходимой производительности вычислительных систем появилась возможность решать широкий спектр задач, ранее не поддававшихся эффективному решению.

Глубокое обучение революционизировало распознавание образов и машинное обучение. Термин «глубокое обучение» был впервые представлен машинным обучением Дехтера (1986) и обучением в искусственных нейронных сетях (ИНС) Айзенбергома и др. (2000). Впоследствии он стал особенно популярным в контексте глубоких ИНС, которые ведут свое начало с 1960-х годов. Глубокое обучение может быть направлено на обучение с учителем, без учителя и с подкреплением.

Основной прорыв глубокого обучения – достижение высокой точности в задачах распознавания и кластеризации (до 95 – 99% на ImageNet). Глубокое обучение включает набор алгоритмов машинного обучения для моделирования высокоуровневых абстракций, применение многоуровневой нейросетевой архитектуры и многочисленные нелинейные преобразования.

Анализ текстов

Методы обработки больших массивов текстов с целью выявления их семантики и прагматики можно разделить на два больших класса: методы, основанные на векторном представлении элементов текста, и методы, основанные на сетевом представлении элементов текста.

1 Дистрибутивная семантика

Дистрибутивная семантика, тесно связанная с понятиями «векторное представление текстов», «парадигматическое представление текстов», и «монограммная модель текстов» (*bag-of-words*) – это область лингвистики, которая занимается вычислением степени семантической близости между лингвистическими единицами на основании их распределения (дистрибуции) в больших массивах лингвистических данных (текстовых корпусах). При этом каждой лингвистической единице, используемой в рамках анализа, присваивается свой *контекстный вектор*. Множество векторов формирует *векторное пространство*. Семантическое расстояние между понятиями, выраженными словами естественного языка, обычно вычисляется как *косинусное расстояние* между векторами выбранного векторного пространства.

Дистрибутивный анализ – это метод исследования языка, основанный на изучении окружения (дистрибуции, распределения) отдельных единиц в тексте (в их векторном представлении) и не использующий сведения ни о полном лексическом или грамматическом значении этих единиц, ни об их синтагматическом окружении на *n* шагов.

В рамках данного метода к текстам изучаемого языка применяется упорядоченный набор универсальных процедур, что позволяет выделить основные единицы языка (морфемы, слова, словосочетания, фразы), провести их классификацию (отношение их к какому-либо классу) и установить отношения взаимозаменимости между ними.

Метод дистрибутивного анализа основывается на принципе замещения: языковые единицы относятся к одному и тому же классу, если они могут выступать в одних и тех же контекстах.

Дистрибутивная семантика основывается на **дистрибутивной гипотезе**: лингвистические единицы, встречающиеся в схожих контекстах, имеют близкие значения.

Рисунок 4 – Изучение диахронического контекста для ряда слов

Векторное представление слов

С выходом работ Миколова в 2013 году векторные представления слов (wordembeddings) во многом пришли на смену n-граммам и устоялись как один из основных способов предобработки текстов (диахроническая визуализация на рис. 4).

Себастьян Рудер опубликовал недавно подробный обзор по поводу последнего состояния векторного представления слов (Word embeddings in 2017: Trends and future directions [5]) и отметил основные тренды по данному направлению.

На данный момент большое число работ посвящено переосмыслению векторов слов для включения информации о близкородственных словах и слов, не вошедших в исходный набор слов модели (out-of-vocabulary words). Часто это достигается традиционным *n*-граммным методом, что позволяет достичь определенного роста точности модели. Наиболее важным и нашумевшим исследованием в этой области за 2016 – 2017 является разработка от FacebookResearch – FastText [6], которая является имплементацией данного подхода.

Другое улучшение векторного представления слов заключается в новом подходе к словам, не входящим в обучающую выборку. Так, ряд исследований предлагает генерировать данные векторы на лету. Однако работы по этому направлению продолжаются.

Также многие работы посвящены учету омонимии в векторах-словах, однако Себастьян Рудер отмечает, что, возможно, это уже не является необходимым из-за успешности использования контекста современными нейронными сетями.

Многие работы рассматривают слова в диахроническом контексте, что позволяет рассматривать изменения в значениях слов на временной оси. Также многие ученые исследуют многоязычные и доменно-ориентированные векторные представления слов.

Так, в статье Lifelong Word Embedding via Meta-Learning [7] предлагается использовать много малоразмерных доменных корпусов и метамодель, что позволяет

улучшить предсказания для новых доменов. В других статьях [8] предлагается обогатить векторные представления слов, используя синтаксис и части речи, а также предлагается другая структура обучения, которая позволяет улучшить результаты.

В 2018 году на замену FastText пришли предобученные сложные языковые модели. Это очень похоже на исходные работы Bengio, но с применением таких более изощренных нейросетевых архитектур как Bi-LSTM и Attention.Среди нейросетевых языковых моделей можно отметить Elmo (allennlp.org/elmo), BERT(github.com/google-research/bert), ULMFit (nlp.fast.ai/) и OpenAI Transformer (blog.openai.com/language-unsupervised/) и неопубликованный GPT-2 (blog.openai.com/better-language-models/). Эти модели позволяют осуществлять обучение на меньших наборах данных с большей точностью. Также исчезает проблема омонимии для эмбеддингов.

Bert, OpenAI Transformer и GPT-2 используют на входе byte-pair encodings (BPE) («побайтовое», «парабайтовое» кодирование) (https://github.com/google/sentencepiece). При данном алгоритме наиболее часто встречающиеся n-граммы встраиваются в более крупные n-граммы (напр. фраза системами «виртуальной реальности» разобьется на n-граммы 1) система 2) ми 3)_« 4) ви 5) р 6) ту 7) аль 8) ной 9)_ре 10) аль 11) ности 12)». Это позволяет уменьшить размер входного векторного слоя для нейронной сети.

ELMo (Embeddings from Language Models)

ELMo (https://allennlp.org/elmo (ссылка не всегда открывается из-за блокировок Роскомнадзора)) (векторные представления языковых моделей) стала одной из основных работ 2018 года в этой области. Обычные языковые модели, такие как Word2Vec страдают от проблемы омонимии – слова могут иметь разные смыслы и соответственно вектора в разных контекстах. ELMo успешно решает эту проблему, так как выдает вектора слов сразу на основе всего предложения или целого параграфа. Конкретно, ELMo использует предварительно обученную, многослойную, двунаправленную, основанную на LSTM языковую модель и извлекает скрытое состояние каждого слоя для входной последовательности слов. Затем эта модель вычисляет взвешенную сумму скрытых состояний, чтобы получить вектор для каждого слова.

Задача языковой модели – предсказание того, какое слово будет следующим в тексте на основе всех предыдущих слов. Вместо использования однослойной LSTM-нейронной сети, в этой работе используется многослойная LSTM. Каждый последующий слой многоуровневой LSTM-модели принимает на вход результат LSTM-модели предыдущего уровня (рис. 5).

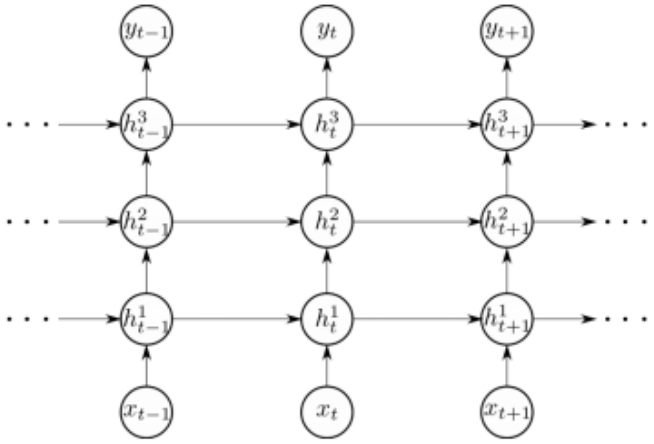


Рисунок 5 – Иллюстрация LSTM-слоя ELMO (http://mlexplained.com/2018/06/15/paper-dissected-deep-contextualized-word-representations-explained/)

Окончательное векторное представление строится по следующей формуле:

$$ELMo_k = \gamma \sum_j s_j h_{k,j},$$

где $h_{k\,j}$ – это вывод j-го LSTM-слоя для слова k, а s_j – это вес текущего слоя, который был получен уже в ходе дообучения модели на конечной задаче (например, классификации текстов с использованием векторных представлений слов из ELMO).

Авторы данной работы предлагают не отказываться от контекстонезависимых представлений слов (Word2Vec, FastText) и использовать ELMO-эмбеддинги вместе с ними.

Из проблем данной модели стоит отметить требовательность к вычислительным ресурсам и ограниченность возможности дообучения (только вектор размерностью 3).

BERT

Авторы из Google Research решили отказаться от LSTM-представлений в ELMO из-за невозможности использования параллельных вычислений на этой архитектуре, и для этого использовали модель под названием Transformer (github.com/google-research/bert) рис. 6. Хорошие англоязычные описания и визуализации этой модели можно найти в следующих статьях в Интернете:

The Illustrated Transformer (jalammar.github.io/illustrated-transformer/).

The Annotated Transformer (nlp.seas.harvard.edu/2018/04/03/attention.html).

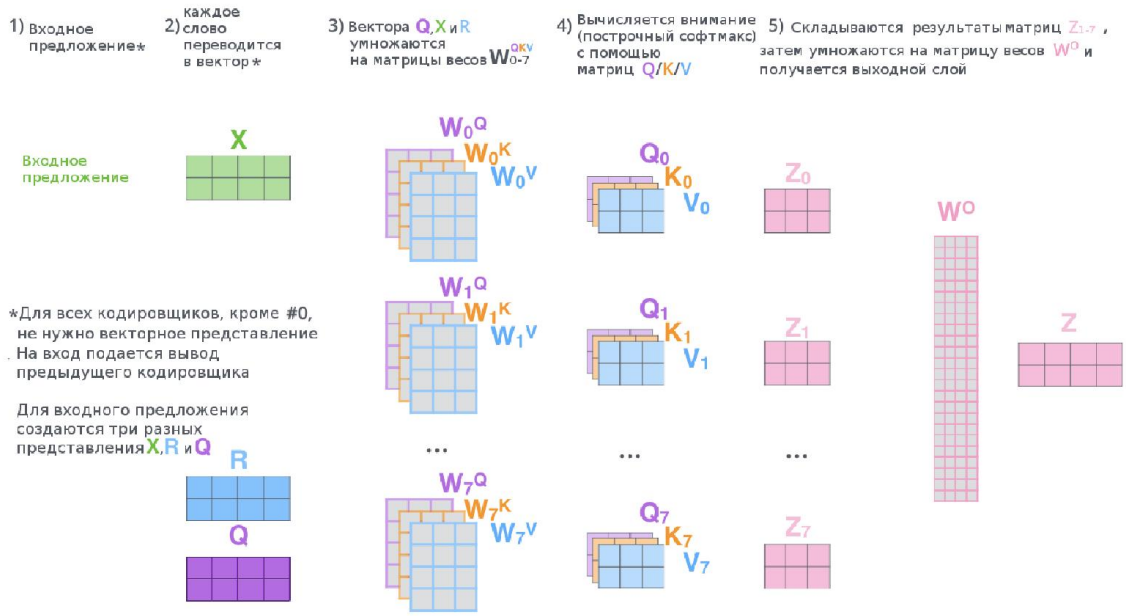


Рисунок 6 – Визуализация jalammar.github.io/illustrated-transformer/

В рамках архитектуры Bert используется, так называемое «самовнимание» (self-attention). Для каждого слова или n-граммы (в Bert используются побайтовые представления, байтовые n-граммы) инициализируются его векторные представления (используются одновременно 3 разных представления Q, K, V). Затем каждое из этих представлений умножается на 8 различных матриц весов (heads). После чего для каждой из получившихся 24 матриц вычисляется софтмакс. Затем результаты софтмакса конкатенируются и передаются на следующий слой нейронной сети.

Из интересных особенностей обучения Bert стоит отметить, что модель обучалась сразу на нескольких различных задачах. Кроме того, на вход подавались не только все предшествующие слова (или все последующие, как в ELMO), а сразу весь контекст, однако предсказываемое слово было скрыто с помощью специального символа.

Первая задача – это предсказание слова в тексте. Однако из-за двунаправленности модели (модель одновременно обучается предсказывать слово, используя контекст слева и справа), агрегирующие слои будут содержать информацию об искомом слове. Поэтому целевое слово маскируется символом <MASK>. Однако это создает проблемы при дообучении модели для решения финальной задачи, так как в текстах едва ли содержится данный символ. Поэтому только в 80% случаев (скорее всего, это эмпирический коэффициент, полученный на маленькой выборке, и следовательно может быть неоптимальным) слово заменяется на маскирующий символ. В 10% случаев предсказываемое слово заменяется на случайное. И еще в 10% случаев слово остается без изменений для увеличения предвзятости (bias) модели в сторону целевого слова.

Вторая задача – предсказание следующего предложения. Модели на вход подавались два предложения, разделенные специальным символом. На выходе модель должна предсказать, является второе предложение случайным или изначально следовало за первым. Эта задача была поставлена для повышения качества модели при работе с текстами, а не отдельными предложениями.

Bert позволяет получить наилучшие результаты после дообучения для большинства основных задач автоматической обработки текстов на естественном языке. Из недостатков стоит отметить еще большие вычислительные потребности (Google рекомендует тензорные вычислительные устройства с объемом памяти 64Gb). Цена тренировки с нуля такой модели на облачных устройствах исчисляется в несколько тысяч долларов. Для кластера из 4 самых современных потребительских GPU тренировка такой модели может занять несколько месяцев. Однако Google предоставил уже обученную модель Bert, русский язык входит в состав многоязычной модели (все языки сразу).

Морфология

Морфологический анализ – важная часть синтаксического анализа. Признаки, полученные при анализе морфологии (леммы, части речи, падежи, формы слов и т.д.), могут быть использованы для глубинного анализа текстов. Важной задачей морфологического анализа является лемматизация – нахождение нужной формы слова для данного входного слова.

Так, в SyntaxNet встроен морфологический анализатор, однако многие, применяющие данный инструмент на практике, отмечают слабость морфологического анализа этой системы. То есть даже в самых совершенных системах, создаваемых крупнейшими компаниями (например, Google), анализ морфологии остается узким местом. Поэтому данная задача имеет большое значение как сама по себе, так и для использования в синтаксическом анализе.

Последние подходы, основанные на нейронных сетях, позволяют улучшить точность и морфологического анализа тоже.

Морфологический анализ – это процесс поиска морфологических характеристик слова. Цель морфологического анализа – найти, из каких морфем построено слово. Например, морфологический анализатор должен определить, что слово «кошки» – множественная форма существительного «кошка», а слово «child» – мно-

жественная форма слова «children». Таким образом, при слове «кошки» на входе морфологический анализатор должен выдать «кошка существительное, женский род, именительный падеж».

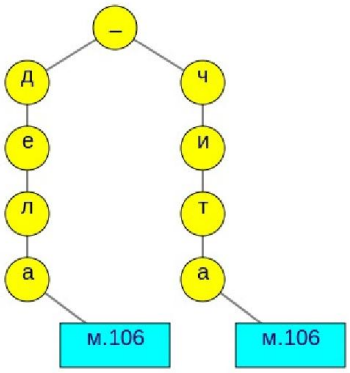


Рисунок 7 – Структура морфологического словаря

Для создания морфологического словаря (рис. 7) используется, например, структура данных «префиксного дерева», которое является видом дерева поиска, хранящего ассоциативный массив (ключ, значение), где ключи – это префиксы строк. Ключ узла графа состоит из символов пути от корня к данному узлу. Значения, ассоциированные с ключами морфологической модели, содержат префиксы данного узла. Для нахождения морфологических характеристик словоформ необходимо совершить поиск по дереву с помощью символов слова. Вычислительная сложность морфологического анализа словоформы линейна и равна O(n), где n – длина словоформы.

Также в рамках частичного решения задачи морфологического анализа существует ряд моделей для определения частей речи, данная задача также часто решается с помощью Bi-LSTMCRF (conditional random fields).

Новые статьи по данному направлению:

В связи с нейросетевым бумом в последнее время используются нейросетевые подходы к классификации текстов на основе морфологического анализа [9].

Можно отметить две модели:

– нейро-семантическая сеть на основе морфологического анализа – *Morphological Neural Semantic Networks* – MNSN;
– рекурсивный автоэнкодер (автокодировщик) морфологического анализа – *Morphological Recursive Auto Encoder* – MRAE.

Нейро-семантическая сеть на основе морфологического анализа состоит из трех последовательных частей.

1. Часть «семантические векторные представления», которая вычисляет векторные представления грамматических структур предложений, содержит автоэнкодеры по заданным грамматическим структурам (SVO, SVA, ...), которые принимают на вход слова в виде пар (вектор, морфология) и объединяют их в одну пару. Цель заключается в том, чтобы близкие по смыслу структуры предложений имели похожие векторные представления, например: «девушка читает книгу» и «женщина читает роман».

2. Часть «распределение по категориям семантического представления», которая принимает на вход объединенный вектор x и вычисляет распределение категорий по предложению, является слоем Softmax.

3. Часть «распределение по категориям текста», которая принимает на вход распределения категорий по предложениям и вычисляет распределение категорий по тексту. Распределение вероятностей для текста, состоящего из N предложений, есть среднее распределение вероятностей по предложениям.

Рекурсивный автоэнкодер морфологического анализа [10] состоит из двух частей: первая объединяет два вектора слов, а вторая объединяет два вектора морфологий. Морфологическая часть рекурсивного автоэнкодера позволяет повышать точность выбора векторов слов в процедуре формирования векторного представления текста. Векторное представление текста получается повторением процесса объединения двух слов-векторов с использованием рекурсивного автоэнкодера. На каждом этапе выбор пары слов-векторов объединение происходит с помощью данного рекурсивного автоэнкодера.

1.2 Синтаксис

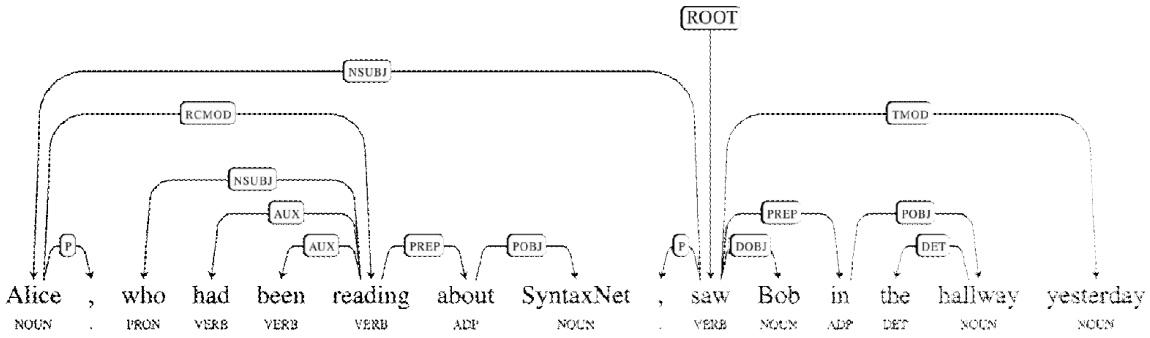


Рисунок 8 – Дерево синтаксического разбора предложения в SyntaxNet

Большое количество работ посвящено синтаксическому анализу естественного языка. Особенно большой отклик произвела работа SyntaxNet (синтаксический парсер на нейронных сетях).

SyntaxNet применяет нейронные сети к проблеме снятия двусмысленности. Вводное предложение обрабатывается слева направо, причем зависимости между предыдущим и последующим словами добавляются постепенно. В каждый момент обработки из-за двусмысленности может возникать множество гипотез. Нейронная сеть дает оценки для конкурирующих решений на основе их правдоподобия. По этой причине очень важно учитывать сразу несколько гипотез. Для этого можно использовать алгоритмы поиска кратчайшего пути на графе. Одним из таких алгоритмов является лучевой поиск. В каждый момент времени оцениваются не одна, а сразу несколько (N) гипотез разбора, на следующем шаге проверяются новые N-гипотез. Лучевой поиск позволяет избежать проклятия размерности и обычно дает более высокие результаты, чем поиск по первому лучшему совпадению, хотя и не дает гарантий оптимального решения (при ограниченном N).

Также используются результаты морфологического анализа. Кроме того, могут вычисляться векторные представления грамматических структур предложений с использованием т.н. автокодировщиков (автоэнкодеров) – нейронных сетей (one-to-vec) по заданным грамматическим структурам (SVO, SVA, ...), которые принимают на вход слова в виде пар (вектор, морфология) и объединяют их в одну пару. Цель заключается в том, чтобы близкие по смыслу структуры предложений имели похожие векторные представления, например «девушка читает книгу» и «женщина читает роман».

Для обучения отдельной грамматической структуры используется отдельный автоэнкодер: автоэнкодер объединяет элементы грамматической структуры в один вектор.

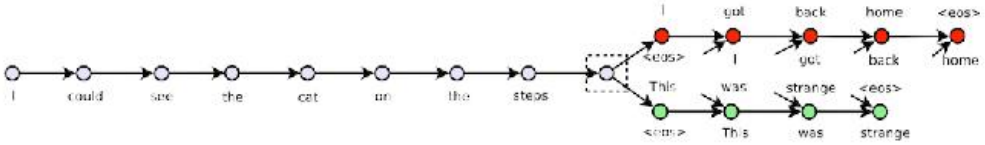
Подробный обзор синтаксических моделей для русского языка можно найти в посте habr.com/ru/company/sberbank/blog/418701/.

1.3 Векторное представление текстов

Работы по векторному представлению текстов очень похожи по своей идее на статьи по векторному представлению слов. Среди этих методов стоит отметить Skip-thoughts и doc2vec.

Алгоритм skip-thoughts [11] близок к векторному представлению слов (skip-gram, word2vec). Чтобы оценить сходство между двумя предложениями, используется архитектура для создания представления текста на основе алгоритма обучения без учителя. В модели skip-gram выбирается слово w_i, на основе которого требуется предсказать окружающий контекст (например, w_{i+1} и w_{i-1} для контекстного окна размером 1). Данная модель работает аналогичным образом, но на уровне предложения. То есть, учитывая (s_{i-1}, s_i, s_{i+1}), данная модель сначала кодирует предложение s_i в фиксированный вектор, затем на основе этого вектора пытается восстановить предложения s_{i-1} и s_{i+1}. Возникновение этой архитектуры вдохновлено гипотезой распределения предложений. Если сходен семантически и синтаксически окружающий контекст предложений, то сходны и сами предложения.

Doc2vecМиколова [12] очень похож на skip-thoughts (рис. 9) по своему исполнению и показывает в теории хорошие результаты, однако на практике многие инженеры столкнулись со сложностью применения этой модели.



Sentence 1	Sentence 2	GT	pred
A little girl is looking at a woman in costume	A young girl is looking at a woman in costume	4.7	4.5
A little girl is looking at a woman in costume	The little girl is looking at a man in costume	3.8	4.0
A little girl is looking at a woman in costume	A little girl in costume looks like a woman	2.9	3.5
A sea turtle is hunting for fish	A sea turtle is hunting for food	4.5	4.5
A sea turtle is not hunting for fish	A sea turtle is hunting for fish	3.4	3.8
A man is driving a car	The car is being driven by a man	5	4.9
There is no man driving the car	A man is driving a car	3.6	3.5

Рисунок 9 – Векторное представление предложения

1.4 Анализ тональности текста

Направление анализа тональности давно и активно развивается в рамках обработки текстов на естественном языке. Большую популярность это направление получило в связи с развитием нейронных сетей и методов векторного представления слов (word2vec).

В последнее время популярны методы, позволяющие оценивать тональность без привлечения аннотированных данных. Часть методов посвящена анализу тональности с помощью эмотиконов (смайликов) [13]. Многие работы в данный момент фокусируются на более специфических видах тональности. Так, часть работ посвящена анализу агрессии и кибербулинга [14]. Другие исследования посвящены анализу

юмора [15] и даже сарказма [16]. Многие исследования используют данные, которые не требуют аннотации. Так, в части исследований используются сообщения с сайта Reddit, содержащие соответствующие теги (/s – обозначают сарказм).

Необходимо упомянуть использование моделей, учитывающих контекст (использование сверточных нейронных сетей – Kim [17]) или LSTM [18].

2 Тематическое моделирование

Очень распространенной темой исследований последние двадцать лет является так называемое тематическое моделирование. Под тематическим моделированием понимается разбиение множества текстов на классы, объединенные общей темой. Тематическое моделирование может реализовываться в виде классификации, если темы заранее известны. Тогда данная операция может выполняться с помощью любого алгоритма классификации, применимого к текстам. Однако часто под тематическим моделированием понимается кластеризация текстов без заранее предопределенных тем.

Для кластеризации текстов с начала 2000-х обычно используется алгоритм LDA (Latent Dirichlet Allocation). В LDA каждый документ может рассматриваться как набор различных тематик. Подобный подход схож с вероятностным латентно-семантическим анализом [19] (pLSA) с той разницей, что в LDA предполагается, что распределение тематик априори удовлетворяет распределению Дирихле [20]. На практике в результате применения LDA получается более корректный набор тематик, чем в случае применения pLSA.

К примеру, модель (как в LDA, так и в LSA) может иметь темы, классифицируемые как «относящиеся к кошкам» и «относящиеся к собакам», тема обладает вероятностями возможности генерировать различные слова, такие как «мяу», «молоко» или «котенок», которые можно было бы классифицировать как «относящиеся к кошкам», а слова, не обладающие особой значимостью (к примеру, служебные слова [21]), будут обладать примерно равной вероятностью в различных темах.

В 2014 году К. В. Воронцовым было предложено теоретическое обобщение моделей LDA и pLSA, которое отходит от вероятностного понимания тематических моделей и решает проблему неустойчивости предыдущих методов. Данное решение было реализовано в программном пакете bigartm [22]. В ходе данной работы был также представлен обзор существующих методов тематического анализа текстов [23] (рис. 10).

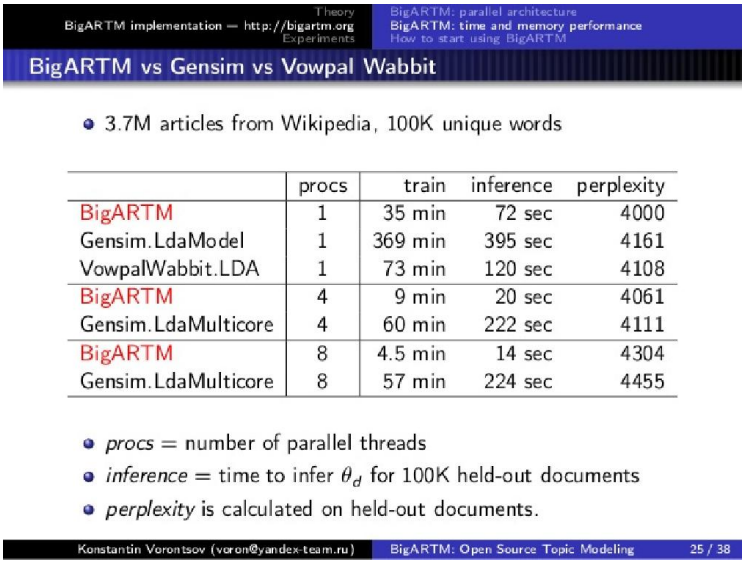


Рисунок 10 – Скорость работы разных пакетов для тематического моделирования

Монограммная модель текста, использованная в LDA и pLSA, основана на предположении, что каждое слово появляется в тексте независимо от остальных слов $p(w_1 \dots w_n) = p(w_1) \dots p(w_n)$. Это, в частности, значит, что любые перестановки слов строки $w_1 \dots w_n$имеют одну и ту же вероятность, что заведомо неверно. N-граммная модель в этих подходах не используется исключительно из-за большой мощности обучающих выборок, необходимых для генерации тематических моделей, которая физически не может быть достигнута из-за необходимости наличия соответствующего количества текстов в обучающей выборке.

3 Сетевая семантика

В отличие от методов, базирующихся на монограммной модели языка (bag-of-words, one-hot), в которой анализируемые единицы текста рассматриваются как несвязанные друг с другом, методы, используемые в сетевых подходах, работают в рамках n-граммной модели языка, модели, в которой единицы текста связаны между собой на глубину n шагов. Естественно, использование второй модели, поэтому более корректно для анализа текстов, чем использование первой.

3.1 N-граммная модель

Создание методов, способных использовать большие окна анализа даже ценой потери контекста в данном окне, во многом смогли вытеснить Марковские цепи из задач анализа текста. Существует ряд исследований по теоретической составляющей word2vec, например, статья Голдберга [24] о том, что векторные представления слов крайне похожи на факторизацию PMI (*Pointwise mutual information* – точечная взаимная информация). Тем самым, многие исследовательские группы возвратились к n-граммам для их использования в анализе текстов [25]. N-граммы используются в качестве одного из входных признаков в новых нейросетевых контекстах для решения задач самого разного рода. Так, в системе FastText используются n-граммы для распознавания слов, близких по написанию и контексту. Также n-граммы используются и в сейчас популярныхByte-PairEncodings (https://github.com/bheinzerling/bpemb; https://github.com/google/sentencepiece).

При анализе текстов нет достоверного априорного знания о равенстве распределений слов в разных позициях строки. Поэтому используются n-граммные модели с целью ввести контекстную привязку через условные вероятности. Обычно используется «односторонняя» n-граммная модель, а именно принятая при использовании n-грамм «правосторонняя» модель, в которой вероятность очередного слова строки задается в зависимости от предшествующих ему $(\mathbf{n-1})$ слов, что записывается как $\mathbf{p(w_n|w_1 \dots w_{n-1})}$.

Тогда: $\mathbf{p(w_1 \dots w_{n-1}w_n) = p(w_n|w_1 \dots w_{n-1})\ p(w_1 \dots w_{n-1})}$.

В терминах вероятности “быть справа” имеем:

$$\mathbf{p(w_1 \dots w_{n-1}w_n) =}$$
$$\mathbf{p(w_n|w_1 \dots w_{n-1})p(w_{n-1}|w_1 \dots w_{n-2}) \dots p(w_3|w_1w_2)p(w_2|w_1)p(w_1)}$$

Оценкой вероятности n-граммы служит частота ее встречаемости:

$$\mathbf{\hat{p}(w_i|w_{i-n} \dots w_{i-1}) = f(w_i|w_{i-n} \dots w_{i-1}) = C(w_{i-n} \dots w_{i-1}w_i)/C(w_{i-n} \dots w_{i-1})}.$$

Так, для биграммной модели оценкой вероятности биграммы является частота ее появления в тексте. В триграммной модели требуется статистика совместного

появления пар слов. Такая модель для английского языка была построена и работала успешно в качестве модели языка в системах распознавания речи. Для русского языка построение триграммной модели оказалось сопряжено с определенными трудностями – не оказалось нужного количества русскоязычных текстов для обучения модели. По этой причине с более граммными моделями трудности усугубляются [26].

3.2 Частный случай n-граммного представления: формирование семантической сети

Если нет достоверного априорного знания о равенстве распределений слов в разных позициях строки, необходимо ввести контекстную привязку, то есть перейти к n-граммной модели текста.

Тогда $\mathbf{p}(\mathbf{w_1} \dots \mathbf{w_{n-1}}\mathbf{w_n}) = \mathbf{p}(\mathbf{w_n}|\mathbf{w_1} \dots \mathbf{w_{n-1}})\ \mathbf{p}(\mathbf{w_1} \dots \mathbf{w_{n-1}})$.

Неудобство ситуации заключается в том, что при формировании такой модели русского языка доступный по объему корпус текстов позволяет сформировать только биграммную модель. Формирование уже триграммной модели затруднено из-за отсутствия необходимого количества текстов. Поэтому формируют так называемую 2,5-граммную модель, в которой вместо третьего в цепочке слова используют грамматические классы слов.

Однако имея некоторое подспорье в виде однородной семантической сети, где вершины – это слова текста, а дуги обозначают ассоциативные связи слов в предложениях текста, можно пользоваться n-граммными моделями, вычисляя соответствующие вероятности не в лоб, а пересчитывая их итеративно.

$$\mathrm{p}(t_i^2) = \frac{\sum_{j=1}^{J_i} \mathrm{p}(w_j|w_i)\mathrm{p}(t_i^1)}{\sum_{i=1}^{I}\sum_{j=1}^{J_n} \mathrm{p}(w_j|w_i)\mathrm{p}(t_i^1)},$$

$$\mathrm{p}(t_i^3) = \frac{\sum_{j=1}^{J_i} \mathrm{p}(w_j|w_i)\mathrm{p}(t_i^2)}{\sum_{i=1}^{I}\sum_{j=1}^{J_i} \mathrm{p}(w_j|w_i)\mathrm{p}(t_i^2)},$$

…

$$\mathrm{p}(t_i^n) = \frac{\sum_{j=1}^{J_i} \mathrm{p}(w_j|w_i)\mathrm{p}(t_i^{n-1})}{\sum_{i=1}^{I}\sum_{j=1}^{J_i} \mathrm{p}(w_j|w_i)\mathrm{p}(t_i^{n-1})},$$

где $\mathrm{p}(t_i^1) = p(w_2)$и $\mathrm{p}(w_j|w_i)$ – одинаковая для всех шагов итерации вероятность появления последующего слова текста при условии появления предыдущего слова.

Так, n-граммная модель может быть реализована, в том числе на основе искусственных нейронных сетей с использованием нейронной сети на основе нейроподобных элементов с временной суммации сигналов для вычисления частот встречаемости и совместной встречаемости слов в тексте, с последующим применением хопфилдоподобного алгоритма итеративного перевзвешивания весовых коэффициентов слов [27].

3.3 Reinforcement learning / Q-learning / Policy-gradient

Andrew Ng в декабре 2016 г. сказал, что 2017 год будет годом Reinforcement Learning (обучение с подкреплением). Результаты ICLR-2018 это подтверждают. Обучение с подкреплением в виде Q-learning (сейчас часто используются как синонимы) не требует аннотированных данных для обучения: в ходе подбора весов нейронная сеть ориентируется на данные некой целевой функции (Q-function). Такой функцией может выступать оценка пользователя в приложении («Довольны ли Вы нашим сервисом?») или некие формальные показатели в виде длительности или лексического разнообразия диалога. И 2018 год ознаменовал целый ряд работ в этом направлении.

Многие статьи посвящены новым архитектурам в сфере обучения с подкреплением. Часть статей ориентируется на создание новых алгоритмов обучения без учителя [28]. Часть работ фокусируется на полностью автоматическом создании целей для обучения нейронной сети [29], в том числе делаются попытки автоматически понимать парадигматику сообщений для генерации текстов [30]. Часть исследований рассматривает аспекты задач, которые разворачиваются во времени [31]. К числу таких задач можно отнести речь и текст.

4 Нейроинформатика. Ассоциативная память – среда для формирования пространства знаний

Несмотря на развитие вычислительной техники и достижения научного направления Искусственный интеллект, мозг остается единственным вычислителем, который эффективно решает интеллектуальные задачи. Эта его способность реализуется благодаря возможности одновременного параллельного анализа большого объема информации, а также благодаря единообразному способу представления и обработки этой информации, независимо от модальности информации: речевой, текстовой, зрительной [32].

Центральная нервная система человека (естественная нейронная сеть) является тем субстратом, на котором реализуются все его творческие способности (как следствие формируемых в естественной нейронной сети когнитивных информационных сетей (http://crm.ics.org.ru/journal/article/1804/). Большая часть интеллектуальных функций человека реализуется на основе кортикоморфной ассоциативной памяти – коры больших полушарий головного мозга – с участием гиппокампа. Поэтому рассмотрение архитектуры, свойств, функциональности ассоциативной памяти человека является естественным способом найти наилучшие решения для реализации интеллектуальных приложений.

Необходимо заметить, что подход к анализу информации со стороны ассоциативной памяти раскрывает одно важное **отличие мозга человека от автоматических систем обработки информации**. В отличие от мозга, все существующие системы в процессе обработки информации пытаются уменьшить объем обрабатываемой информации, снять вариативность. Мозг же, наоборот, наращивает мощность представления вариантов, делая, таким образом, обработку информации все более тонкой и точной.

Обработка специфической информации в мозге приводит к формированию представления знаний о мире (модели мира) в виде так называемых когнитивных семантических сетей (http://crm.ics.org.ru/journal/article/1804/), которые опосредуют структурное представление информации (http://urss.ru/cgi-bin/db.pl?lang=Ru&blang=

ru&page=Book&id=32713). Когнитивные семантические сети возникают в колонках коры полушарий головного мозга виртуально, в процессе обработки информации на субстрате естественных нейронных сетей. При формировании когнитивной информационной сети ее вершины – образы событий разной сложности, хранящихся в колонках коры, – ранжируются по степени важности (связности) в рамках конкретных квазитекстов (описывающих фрагменты предметных областей, в свою очередь представленных корпусами квазитекстов, в том числе – естественно-языковых текстов), что позволяет выявлять «тематическую структуру» этих моделей предметных областей и ассоциативно навигировать по хранилищу – ассоциативной памяти. Ранжирование осуществляется с участием ламелей гиппокампа, в которых формируются представления о целых ситуациях (описываемых отдельными предложениями квазитекста), в которые входят упомянутые образы событий.

Единообразие представления информации различных модальностей в мозге позволяет объединять разномодальную информацию в едином пространстве обработки, что дает еще одно интересное свойство кортикоморфной ассоциативной памяти – позволяет поддерживать процесс принятия решения за счет привлечения к этому процессу более полного многомодального описания.

Технология формирования подобных семантических представлений оказывается удобной для выявления смысла (ассоциативного соотнесения входной информации с тематической структурой модели мира) в больших массивах информации различной природы. В первую очередь, это касается обработки текстовой информации, которая в 90% случаев исчерпывает (в содержательном смысле) современные информационные потоки. И поскольку языковой компонент модели мира (предметной области) изоморфен ее (модели мира) многомодальному компоненту, эта технология оказывается целиком адекватной смысловой обработке текстов.

Кроме того, в процессе такой обработки, в силу ассоциативности принципов обработки, исключаются повторы записи информации в длительное хранение, т.е. такое представление оказывается компактным и легко масштабируемым по сравнению с другими методами хранения, используемыми в традиционных видах микроэлектронной памяти.

И наконец, мозг представляет собой параллельный потоковый вычислитель, что существенно отличает его от традиционных фон-Неймановских архитектур. В силу того, что достигается физический предел применимости закона Мура из-за приближения проектных норм в микроэлектронике к размерам атома, переход к параллельным архитектурам вычислителей, в том числе на основе кортикоморфной ассоциативной памяти, оказывается способом решения проблемы дальнейшего роста мощности вычислительной техники. Этот же параллелизм обработки позволяет реализовать энергоэффективные вычисления.

Основополагающие исследования по архитектуре сенсорных систем:
(https://books.academic.ru/book.nsf/62139471/Зрение+и+мышление),
(https://royallib.com/book/hyubel_devid/glaz_mozg_zrenie.html),
моторных систем (https://search.rsl.ru/ru/record/01001318643), отдельных органов мозга (http://www.mathnet.ru/links/1e9bdd73f148bc1e386cd8a081a32c69/trspy252.pdf), (https://eknigi.org/estestvennye_nauki/124836-gippokamp-i-pamyat.html), и всего мозга человека в целом (http://2dip.su/список_литературы/123674/) были предприняты в последней трети прошлого века. Однако и сейчас проводится колоссальное количество нейроэлектрофизиологических исследований, посвященных изучению архитектуры мозга в целом [33] и отдельных его органов – например, гиппокампа, что показано в статье «The Hippocampusasa Predictive Map» [34].

Понимание того, как мозг обрабатывает информацию, позволяет реализовать как локальные механизмы обработки информации, так и их объединение в глобальный инструмент.

Архитектура интеллектуальной системы, решающей задачу ведения интеллектуального диалога, должна включать в себя, по крайней мере, три основных модуля: (1) модуль, формирующий и хранящий модель мира, включающую, в свою очередь, языковой компонент; (2) модуль, формирующий и хранящий модели отдельных ситуаций; а также (3) модуль, формирующий план целенаправленного поведения, контролирующий также выполнение этого плана.

Эти три модуля использует человек в процессе целенаправленного поведения. Модель мира у него формируется в колонках коры полушарий большого мозга. А модели ситуаций, формирующиеся в ламелях гиппокампа, используются фронтальной корой для построения плана поведения. Это не значит, что надо просто моделировать кору и гиппокамп, но требуется воспроизведение архитектуры обработки информации, как она реализуется мозгом человека [32].

5 Объединение тенденций

В последнее время исследователи стремятся не только увеличить глубину искусственных нейронных сетей, придумать позаковыристей алгоритм их функционирования, но они озаботились их макроархитектурой, начиная понимать, что мозг человека глубоко неоднородная естественная нейронная сеть, и сложность его архитектуры неслучайна. Здесь они отталкиваются от работ, связанных с пониманием особенностей работы головного мозга.

5.1 Память

Большое внимание сейчас уделяется различным структурам памяти, которые позволяют хранить нейронным сетям больший объем информации, что приближает их к человеческой памяти. Сейчас существует много экспериментов в этой области. Исследователи или видоизменяют уже испытанные методы вроде Long short-term memory (LSTM – долгая краткосрочная память), recurrent neural networks (RNN – рекуррентные нейронные сети), или пробуют новые подходы, будь то эпизодическая память или другие структуры (рис. 11).

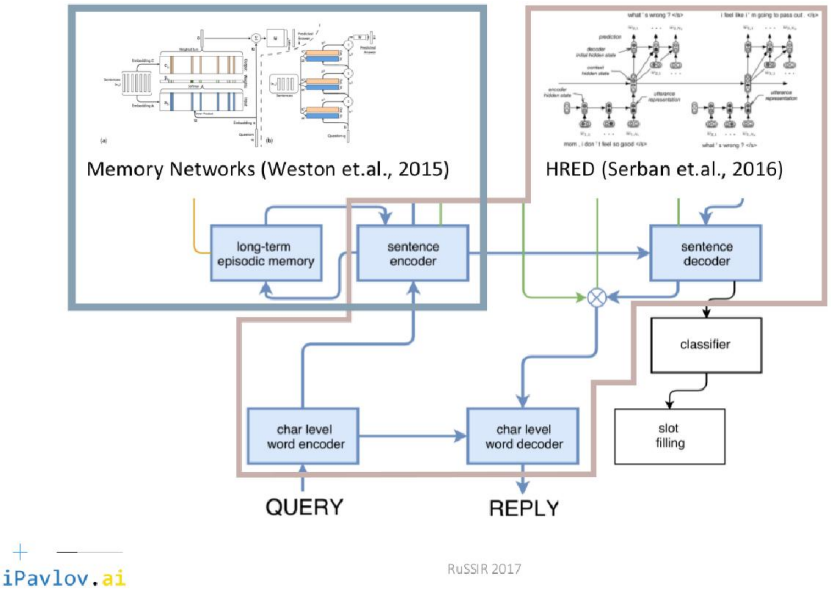


Рисунок 11 – Структура памяти в составных нейронных сетях для диалоговых агентов

Часть исследований в данной тематике посвящена эпизодической памяти [35]. Многие исследования сфокусированы на улучшениях структур памяти рекуррентных нейронных сетей [36], другие же напротив нацелены не на эпизодическую, а долговременную память [37]. Некоторые исследователи подходят к этой же проблеме с другой стороны и предлагают способы структурирования памяти [38]. Также стоит отметить использование моделей с памятью для генерации текстов [39]. Модели памяти в большинстве случаев лишь вдохновлены нейрофизиологией, но обычно остаются от нее далеки.

5.2 Внимание

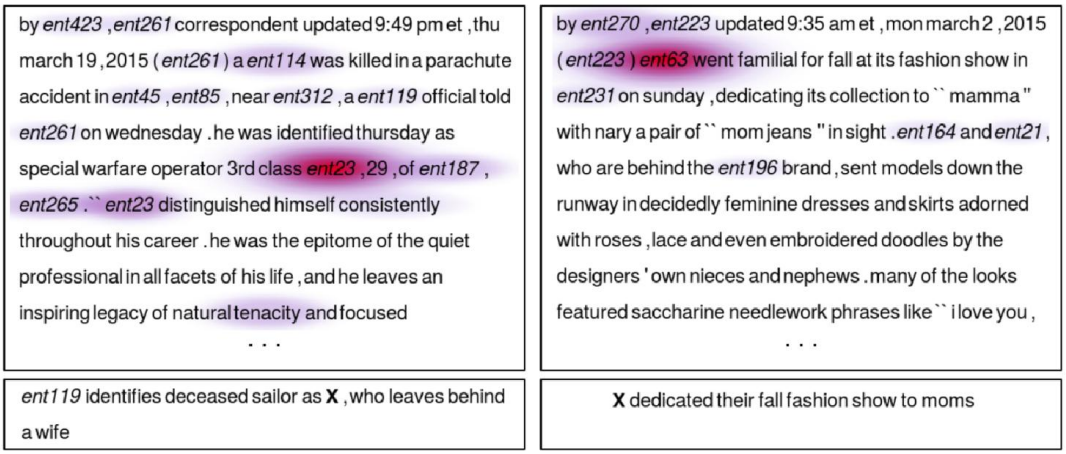


Рисунок 12 – Пример выделения текста систем с вниманием

Уже долгое время ученые опробуют новые подходы, основанные на механизме «внимания» (взвешенное среднее, примененное к LSTM-слою сети). Изыскания в данной области продолжаются. Предлагаются новые методы, например RWA (Recurrent Weighted Average) [40].

Также ряд статей посвящен более структурному и иерархическому обучению с использованием механизмов внимания [41], определению ремы [42], анализу последовательных данных [43], внедрению механизмов внимания в нейронные сети, отличные от рекуррентных [44]. Следует отметить, что внимание в нейронных сетях лишь вдохновлено человеческим мозгом, а само по себе представляет процедуру перемножения матриц с последующим softmax слоем. Но даже данная простая модель позволяет достичь впечатляющей интерпретируемости.

5.3 One-shot learning

Нейронные сети обучаются обычно довольно продолжительное время (неделями в случае глубоких сетей для обработки изображений). Однако это не соответствует человеческим представлениям об обучении. Человек способен обучаться на минимальном количестве примеров. Кроме того, глубина нынешних нейронных сетей позволяет подстраиваться под любой шум и вбросы в результатах. Поэтому ученые рассматривают новые методы обучения, одним из них является one-shot learning (рис. 13).

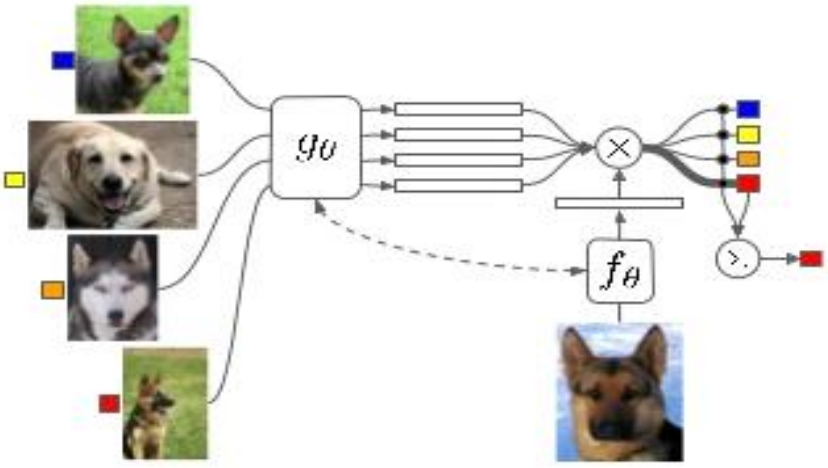


Рисунок 13 – Архитектура сети для One-shot Learning

One-shot learning позволяет обучаться концепту с использованием всего одного или небольшого числа тренировочных примеров. Так, в ходе одного из экспериментов тренировался классификатор на основе метода ближайших соседей, который включал в себя характеристики как параметрических, так и непараметрических моделей. Это позволило улучшить результаты для задачи Omniglot с 88% до 93.2%, по сравнению с другими подходами (классификация символов из 50 алфавитов с малым количеством тренировочных примеров для каждого класса – 1623 класса и 20 примеров на каждый). Это исследование было проведено командой DeepMind [45], которая в 2016 году произвела фурор в игре Го, неожиданно обыграв чемпиона мира в этой игре с помощью нейросетевой архитектуры.

Стоит отметить, что исследователи в этом направлении также используют другие методы – например, структуры памяти [46]. Под one-shot learning часто понимают transfer learning, т.е. когда модель предобучалась на одних данных (часто неразмеченных), а потом дообучается на новом небольшом датасете.

5.4 Модульные нейронные сети

Знакомство с современными чат-ботами показывает, что они недалеко ушли от «Элизы». И проблема глубокого обучения заключается не в объеме обучающей выборки, а в принципиальной неоднородности алгоритмов интеллектуальных систем (будь то человек или искусственная нейронная сеть), решающих поведенческие задачи, каковой, в том числе, является диалог.

Произошла смена парадигмы: вместо одной нейронной сети на задачу приходится одна нейронная сеть на функцию. Сразу несколько нейронных сетей обмениваются информацией для решения высокоуровневых сложных задач. Кроме того, при данной парадигме не только тренируются нейронные сети, но используются и фиксированные модули (рис. 14).

Поэтому архитектура интеллектуальной системы, решающей задачу ведения интеллектуального диалога, должна включать в себя, по крайней мере, три основных модуля: (1) модуль, формирующий и хранящий модель мира, включающую, в свою очередь, языковой компонент; (2) модуль, формирующий и хранящий модели отдельных ситуаций; а также (3) модуль, формирующий план целенаправленного поведения, и также контролирующий выполнение этого плана (рис. 15).

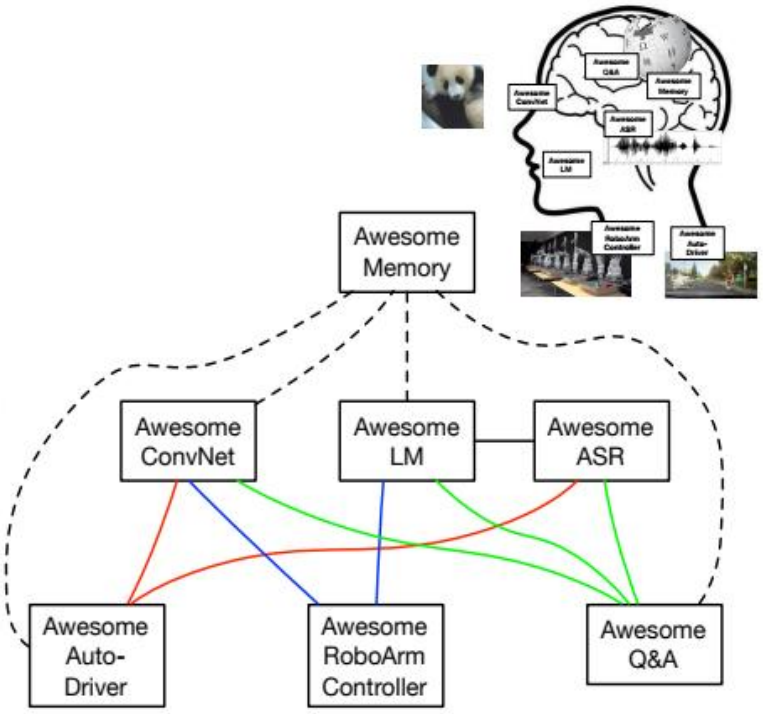


Рисунок 14 – Гипотетическая структура ИИ, состоящего из множества доменных нейронных сетей

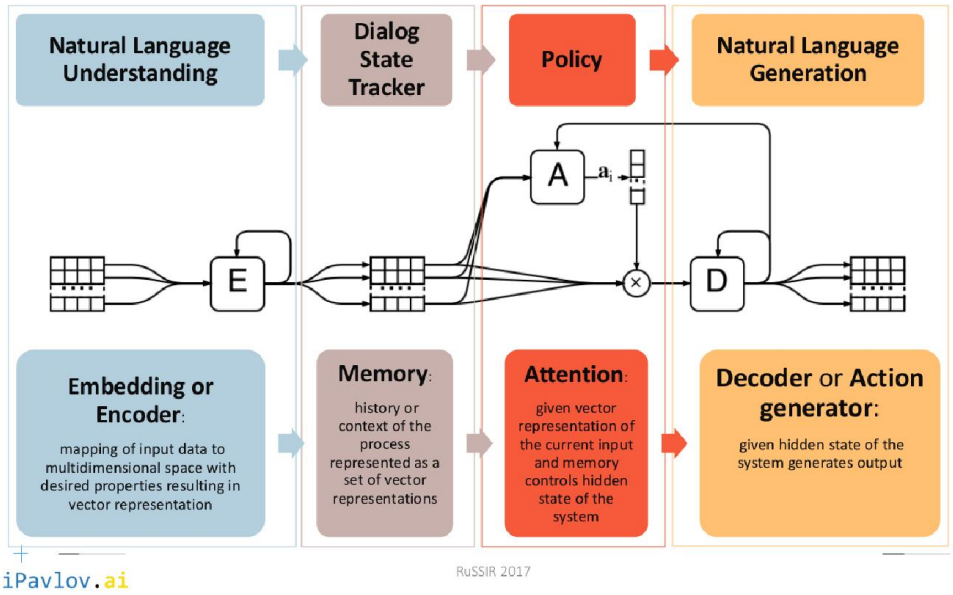


Рисунок 15 – Структура метадиалоговой системы из многих нейронных сетей

В случае диалога, инспирированного человеком-пользователем, инициатива в ведении диалога исходит от пользователя, поэтому в диалоговой системе не требуются модули, вносящие в систему потребности, которые бы стимулировали развитие диалога: направление диалога определяется пользователем. В этом случае необходимо только отслеживать цели диалога, которые возникают в процессе диалога у пользователя.

Для эффективного ведения диалога интеллектуальной системе необходимо выявлять цели, которые ставит перед системой пользователь в процессе диалога, и достигать этих целей.

Выявление целей пользователя осуществляется в процессе анализа обращений пользователя к системе. Цели пользователя могут быть представлены в обращении к системе в явном виде («Я хочу открыть счет»). Но цели могут быть затемнены привходящими обстоятельствами. В этом случае необходим перехват инициативы с последующей попыткой ввести диалог в один из возможных (доступных) сервисов.

Достижение целей диалога возможно с использованием модели предметной области, опосредующей выбранный сервис. В этом случае диалог разбивается на поддиалоги, каждый из которых имеет свою подцель, и поддиалог сводится к достижению этой подцели.

Модель предметной области в этом случае удобно представить в виде графа (например, семантической сети). Тогда отдельные поддиалоги могут рассматриваться как фрагменты текста, проецирующиеся на этот граф (высекающие на нем фрагмент модели предметной области). Этот фрагмент может быть однозначно интерпретируемым: «делай раз», «делай два», …. Он может интерпретироваться неоднозначно, и тогда потребуется дальнейшее углубление в структуру поддиалога.

Отдельные решения в рамках этого направления касаются в основном анализа текстовой информации.

5.4.1 Диалог

Подход к анализу текстов через обучение с подкреплением, так называемый «Reinforcementlearning», позволяет дообучать модели и улучшать работу использующих их чат-ботов. Среди таких работ можно отметить [47]. Также многие работы призваны заменить методы оценки диалогов [48].

Применение обучения с подкреплением к формированию диалога позволяет незначительно улучшить его качество за счет увеличения длины содержательной части диалога. Однако смысловое содержание диалога остается за рамками подхода: в процессе его применения формируются формальные шаблоны.

Этот подход может быть применен для решения поставленных задач именно с точки зрения автоматического формирования перечня формальных шаблонов.

5.4.2 Вопросно-ответные системы

Выявление цели диалога (глобальной или локальной) связано с распознаванием типов вопросов. Наличие выявленной цели диалога позволяет достичь ее умением синтезировать ответ на поставленный вопрос. Распознавание типа вопроса (intent), таким образом, является одним из ключевых моментов в процессе ведения диалога [49].Часть работ уделяет большее внимание ответам на вопросы [50], другие же работы фокусируются на ответах на вопросы по тому или иному тексту [51].

5.4.3 Дифференцируемый нейрокомпьютер

Исследователи Google DeepMind разработали так называемый дифференцируемый нейронный компьютер, DNC, который сочетает обучаемость нейросетей с дедуктивными способностями естественного ИИ [52].

В способности обучаться на данных без прямого программирования человеком заключается главное преимущество нейросетей. Однако простейшие нерекуррентные нейросети не являются полными по Тьюрингу, т.е. они не могут делать всех тех

вещей, на которые способны традиционные алгоритмические программы. Одна из причин этого – отсутствие у нейросетей внешней памяти, с помощью которой можно оперировать входными данными, хранить локальные переменные, и их повторно использовать.

Даже рекуррентные нейросети, реализующие ассоциативную память, имеют важный недостаток, который заключается в неявном характере использования памяти: размерность и характер обращения с памятью определяется архитектурой самой нейросети.

DNCDeepMind можно обучать на примерах (с помощью метода обратного распространения ошибки), а не программировать в явном виде. Но в нем общение вычислителя с памятью организовано существенно более гибким образом: реализуются концепции не только запоминания, но и контекстного узнавания и забывания.

Упрощенно работу DNC можно представить следующим образом. Система состоит из вычислителя, в роли которого может быть использована практически любая рекуррентная нейросеть и любая память. У вычислителя есть специальные модули для обращения к памяти, а над памятью есть особая надстройка в виде матрицы, хранящей историю ее (памяти) использования. Эта история хранится в матрице размерности N×N, которая позволяет системе последовательно «вспоминать» блоки данных, если они часто встречаются в контексте друг друга.

Эта система была испытана в нескольких тестовых задачах. Одной из них был тест на понимание графов. Обучающая выборка была представлена как последовательность предложений, которые описывали структуру некоторой сети (представленной в графическом виде) – например, реальной сети лондонского метро.

После обучения на миллионе примеров DNC компьютер научился отвечать на вопросы по схеме метро с точностью в 98,8 процентов, при этом система на базе LSTM почти совсем не справилась с задачей – она давала только 37 процентов правильных ответов (цифры приведены для самой простой задачи, наподобие «где я окажусь, если проеду столько-то станций по такой-то линии, пересяду там-то и проеду еще столько-то станций». Задача о кратчайшем расстоянии между двумя станциями оказалась более сложной, но с ней DNC тоже справилась).

Что очень важно в архитектуре DNC для решения задачи поиска цели диалога и достижения этой цели: любой диалог с клиентом базируется на использовании некой инструкции. Так, в банке это инструкции по использованию конкретных сервисов. Автоматическое восстановление **модели предметной области**, которая формируется при обучении DNC (модели предметной области, представленной в виде графа, как схема метро, или представленной в виде инструкции) позволяет говорить о возможности ведения осмысленного диалога, в отличие от бессмысленной имитации диалога чат-ботами. Кроме того, новые исследования в области диалогов в какой-либо предметной области отходят от слот-ориентированных диалогов, где разговор представляется как задача по заполнению заранее предопределенных позиций, и используют нейронные сети с разной степенью успеха для ведения диалога целеориентированного [53].

5.4.4 Однородная семантическая сеть как модель предметной области

Другим примером модели предметной области является ассоциативная (однородная семантическая) сеть [32].

Ассоциативная (однородная семантическая) сеть строится с использованием механизмов, которые реализуются в головном мозге человека. Однородная обработка специфической информации в головном мозге человека осуществляется в основном

в двух структурах: в колонках коры большого мозга и в гиппокампе. В колонках информация о событиях, упорядоченная по ассоциации и схожая по форме хранится вместе. Кроме того, она упорядочена по иерархии: чем выше информация общего характера, тем оптимальнее она хранится и обрабатывается. На каждом уровне иерархии создаются словари событий своего уровня. Они связаны так, что слова более высокого уровня являются грамматиками для слов более низкого уровня.

Колонки коры, помимо нейронов других типов, состоят в основном из пирамидных нейронов третьего слоя, которые, будучи электронекомпактными, осуществляют временную суммацию сигналов. Искусственные нейронные сети на основе нейронов с временной суммацией сигналов моделируют колонки коры. Они реализуют многоуровневую структурную обработку информации на основе ассоциативного преобразования, в результате которой формируется иерархическое представление в виде множества автоматически выявляемых словарей событий различной частоты встречаемости.

Гиппокамп, имеющий структуру, состоящую из множества независимых образований – ламелей, моделируемых искусственной нейронной сетью Хопфилда – нейронной структурой поля CA_3, хранит в каждой такой структуре связи событий колонок коры в рамках более крупного события – ситуации.

Использование парадигматического представления информации, характерного для колонок коры, для хранения текстовой информации морфологического, лексического и синтаксического уровней, и формирование на семантическом уровне ассоциативной сети ключевых понятий с последующей перенормировкой весов понятий в соответствии с их смысловой значимостью в тексте, позволяет реализовать технологию автоматического смыслового анализа текстов, с помощью которой можно автоматически извлекать ключевые понятия текста (слова и устойчивые словосочетания), формировать семантическую сеть ключевых понятий со взвешенными понятиями и связями, автоматически реферировать текст, сравнивать тексты по смыслу (следовательно, классифицировать их), кластеризовать корпус текстов по темам.

Указанный подход позволяет автоматически, на основе анализа статистики слов и их связей в тексте, реконструировать внутреннюю структуру текста. Важной особенностью используемого подхода является возможность автоматически устанавливать взаимосвязи между выявленными элементами текста.

При выявлении связей учитывается статистика попарного появления слов в фрагментах исследуемого материала. Далее статистические показатели пересчитываются в семантические с помощью итеративной процедуры [27], идея которой заключается в том, что при расчете весовой характеристики элемента сети учитываются весовые характеристики элементов с ним связанных, а также учитываются численные показатели связей. После пересчета статистические характеристики понятий, которые слабо связаны с другими понятиями в тексте, получают малый вес, а наиболее взаимосвязанные наделяются высокими показателями. Полученная семантическая сеть отражает внутреннюю структуру текста (корпуса текстов), значимость выделенных понятий, а также показывает степень связанности понятий в тексте.

Семантические веса элементов сети используются при расчете смысловой близости (релевантности) текстов. На их основе возможно выделение наиболее информативных участков текста. Использование ассоциативных связей элементов сети позволяет расширять поле поиска информации. Ответ на запрос пользователя в этом случае может содержать информацию, явно не указанную в запросе, но связанную с ней по смыслу. Примером реализации данного подхода может служить программа TextAnalyst [54].

5.4.5 Прагматика предметной области как основа для ведения диалога

Наличие модели предметной области позволяет говорить содержательно о выявлении сценарной части текста – сюжетных линий текста [32].

Если спроецировать текст из конкретной предметной области на семантическую сеть – модель этой предметной области, то мы получим множество цепочек понятий в их взаимосвязях, пробегающих по вершинам семантической сети.

Под прагматическим анализом в данной работе будем понимать выявление сценария текста (корпуса текстов), представленного в виде цепочки (цепочек) расширенных предикатных структур, соответствующих оставшимся после удаления несущественной части предложений предложениям текста (корпуса текстов описывающих предметную область). Сценарий описывает динамику развития представленной в тексте (корпусе текстов) ситуации. Такая цепочка может быть описательной или алгоритмической. В первом случае сценарий характеризует восприятие, во втором – действие. Такое разделение мы наблюдаем в функциях коры полушарий большого мозга человека: задняя кора реализует описательное представление мира, фронтальная кора – деятельное. Прагматическому анализу обязательно предшествует семантический анализ: до выявления прагматики текста необходимо сформировать семантическую модель предметной области, или семантическую модель текста, на ключевые понятия которой в дальнейшем проецируется входной текст.

Прагматический анализ текста заключается в выделении цепочек предикатных структур предложений, которые на этапе семантического анализа целого текста оказались наиболее весомыми в рамках предметной области, к которой относится текст. Степень важности предложений текста определяется с учетом степени важности ключевых слов, которая определяется их ранжированием в рамках семантической сети предметной области на этапе семантического анализа. Количество этих предикатных структур зависит от порога, примененного к смысловому весу предикатных структур предложений, содержащих предикатные структуры, а порядок этих предикатных структур в цепочках – от порядка следования оставшихся после ранжирования и порогового преобразования предложений в тексте. Такие цепочки полностью характеризуют смысловое содержание текста (корпуса текстов – предметной области).

Понимание конкретного текста связано с выявлением расширенных предикатных структур $P_l = (S, O, < O_i >, < A_j >)$, где P_l – это расширенная предикатная структура, в составе S – субъекта, O – главного объекта, $< O_i >$– других – второстепенных – объектов, и $< A_j >$– атрибутов, характеризующих смысл предложений этого текста, а также – цепочек этих предикатных структур W_k=($P_l | l = 1..L_k$), которые опосредуют смысл отдельных последовательностей предложений текста [57]. Любой текст данной предметной области, порождающий цепочку предикатных структур W_k, таким образом, может быть проинтерпретирован как последовательность предложений текста, их содержащих. Под пониманием текста в данном случае понимается проекция цепочек предикатных структур текста на множество соответствующих цепочек предикатных структур предметной области, и поименование этих цепочек соответствующими им предложениями.

Множество таких прагматических цепочек $\{W_k | k = 1..K\}$, извлеченное из полного, в некотором смысле корпуса текстов$\{T_m\}_n$, описывающих предметную

область M_o, может быть подвергнуто процедуре кластеризации по степени их похожести. В процессе такой кластеризации некоторые цепочки могут быть разорваны на подцепочки, или, наоборот, рекомбинировать в более крупные цепочки. Однако, в любом случае, получается множество классов $\{CW_p\}$ цепочек, которые (классы) в совокупности описывают эту предметную область M_o.

Если мы будем анализировать некоторый текст из указанной предметной области, не обязательно входящий в исходный корпус текстов, то выделенные в результате анализа прагматического уровня из этого текста цепочки должны в большей или меньшей степени совпадать с цепочками сформированных классов. То есть, можно использовать прагматическое представление текста для его классификации соотнесением с прагматическими представлениями корпусов текстов, описывающих разные предметные области M_o.

Такие классы цепочек и характеризуют конкретные фрагменты диалога, сотнесенные с моделями предметных областей. Эти фрагменты диалога и используются в конкретных случаях достижения цели диалога. Если таких цепочек в классе несколько, они могут испытываться в диалоге последовательно, с ранжированием их по степени вероятности появления в модели предметной области.

Заключение

Анализ развития направления применения искусственных нейронных сетей для создания систем, реализующих интеллектуальный диалог, в том числе для анализа текстов, показывает, что от простых однородных искусственных нейронных сетей для решения простых задач анализа текстов (классификация, кластеризация) исследователи стали переходить к применению неоднородных систем из нейронных сетей. От векторного представления (дистрибутивная семантика) единиц текста разных языковых уровней – к сетевому представлению содержания целого текста (корпуса текстов), все чаще обращаясь к попыткам следовать архитектуре естественных нейронных сетей мозга человека.

Кроме того, стоит отметить, что исследователи предлагают все более и более сложные генерализованные предобученные модели и перешли от Word2Vec, который можно обучить на недорогом домашнем компьютере, к большим сетям типа Bert и Elmo, которые помещаются далеко не на каждую видеокарту. А успешное обучение Bert с нуля на видеокартах требует многие месяцы чистого времени и стоит десятки тысяч долларов при аренде оборудования. Поэтому исследователи предлагают специализированные устройства (TPU) для обучения нейросетевых моделей.

Однако область применения нейросетевых моделей для задач автоматического распознавания языка сейчас бурно развивается. И исследователи почти каждую неделю предлагают что-то новое, однако мы пока все еще далеки от сильных диалоговых агентов. Так как существующие подходы машинного обучения не способны выйти за рамки данных из тренировочной выборки. Кроме того, до сих пор в большинстве задач не учитывается прагматика текстов. А там, где она учитывается, это ограничивается предсказанием малого количества прагматических классов (интентов). Данный подход позволяет решать ряд задач для бизнеса, однако, несмотря на утверждения многих исследователей и популяризаторов в области автоматической обработки языка, свой ImageNet еще не создан для NLP (в отличие от компьютерного зрения), а существующие сложные модели типа Bert или Elmo – это, скорее, аналог предобученного вероятностного автокодировщика, GAN’а или сети Больцмана для

изображений. Это, несомненно, важный этап в становлении области, однако в NLP всё ещё нерешенных задач гораздо больше, чем решенных. Однако, возможно, что какие-либо из рассмотренных в данной главе методов помогут частично разрешить существующие проблемы в области автоматической обработки языка.

Список литературы

1. Антонова А. Ю. Метод условных случайных полей в задачах обработки русскоязычных текстов [Текст] / А.Ю. Антонова, А.Н. Соловьев // «Информационные технологии и системы – 2013». – Калининград, 2013.
2. McKinsey. Global Institute, Artificial Intelligence the Next Digital Frontier? [Электронный ресурс] / McKinsey. – Режим доступа : URL: http://www.mckinsey.com/~/media/mckinsey/industries/advanced electronics/our insights/how artificial intelligence can deliver real value to companies/mgi-artificial-intelligence-discussionpaper.ashx (дата обращения: 04.12.2017).
3. Wang Tim. Mizuho Industry Focus. A Survey of U.S. Healthcare IT Industry Landscape [Электронный ресурс] / WangTim. – Режим доступа : URL: https://www.mizuhobank.com/fin_info/industry/pdf/mif_183.pdf (дата обращения: 28.07.2016)
4. Tractica. Natural Language Processing: Enterprise Applications for Natural Language Technologies (Processing, Understanding, Generation) Software and Systems: Market Analysis and Forecasts [Электронный ресурс] / Tractica. –
   Режим доступа : URL: https://www.reportlinker.com/p05069690/Natural-Language-Processing-Enterprise-Applications-for-Natural-Language-Technologies-Processing-Understanding-Generation-Software-and-Systems-Market-Analysis-and-Forecasts.html (дата обращения: 04.12.2017).
5. Ruder Sebastian. Word embeddings in 2017: Trends and future directions [Электронный ресурс] / Ruder Sebastian. – Режим доступа : URL: http://www.ruder.io/word-embeddings-2017/index.html (дата обращения: 04.12.2017).
6. Facebook Research, Fast Text [Электронный ресурс] // Сайт: fastText. Library for efficient text classification and representation learning. –
   Режим доступа : URL: https://www.fasttext.cc/ (дата обращения: 04.12.2017).
7. Lifelong Word Embedding via Meta-Learning [Электронный ресурс]. – Режим доступа : URL: https://openreview.net/forum?id=H1BO9M-0Z (дата обращения: 04.12.2017).
8. Word2net: Deep Representations of Language [Электронный ресурс]. – Режим доступа : URL: https://openreview.net/forum?id=SkJd_y-Cb&noteId=SkJd_y-Cb (дата обращения: 04.12.2017).
9. Харламов А. А. Нейросетевые подходы к классификации текстов на основе морфологического анализа [Электронный ресурс] / А.А. Харламов, Ле Мань Ха. – Режим доступа: URL: https://mipt.ru/science/trudy/soderzhanie-zhurnala-trudy-mfti-tom-9-2-34-2017.php (дата обращения: 04.12.2017).
10. Semi-supervised recursiveautoencoders for predicting sentiment distributions / R. Socher, J. Pennington, E.H. Huang, A.Y. Ng, C.D. Manning // Proceedings of the conferenceon empirical methods in natural language processing 2011 Jul 27. Association for Computational Linguistics. – P. 151–161.
11. Kiros Ryan. Skip-Thought Vectors [Электронный ресурс] / Kiros Ryan, Zhu Yukun et al. – Режим доступа : URL: https://arxiv.org/abs/1506.06726 (дата обращения: 04.12.2017).
12. Le Q. Distributed representations of sentences and documents [Текст] / Q. Le, T. Mikolov // International conference on machine learning. – 2014. – С. 1188–1196.
13. Kouloumpis Efthymios. Twitter Sentiment Analysis: The Good the Bad and the OMG! [Электронный ресурс] / Kouloumpis Efthymios, Wilson Theresa et al. –
   Режим допуска :
   URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/download/2857/3251
   (дата обращения: 04.12.2017).
14. Zhao Rui. Automatic detection of cyberbullying on social networks based on bullying features [Электронный ресурс] / Zhou Anna, Mao Kezhi. – Режим допуска : URL: http://dl.acm.org/citation.cfm?id=2849567 (дата обращения: 04.12.2017).
15. Bolotova V. V. Which IR Model Has a Better Sense of Humor? Search over a Large Collection of Jokes [Электронный ресурс] / V.V. Bolotova, V.A. Blinov et al. – Режим доступа: URL: http://www.dialog-21.ru/media/3905/bolotovavvetal.pdf (дата обращения: 04.12.2017).
16. Poria Soujanya. A Deeper Look into Sarcastic Tweets Using Deep Convolutional Neural Networks [Электронный ресурс] / Poria Soujanya, Cambria Erik, Hazarika Devamanyu, Vij Prateek – Режим доступа : URL: https://arxiv.org/abs/1610.08815 (дата обращения: 04.12.2017).

17. Kim Yoon. Convolutional Neural Networks for Sentence Classification [Электронный ресурс]. – Режим доступа: URL: https://arxiv.org/abs/1610.08815 (дата обращения: 04.12.2017).
18. Tang Duyu. Document Modeling with Gated Recurrent Neural Networkfor Sentiment Classification [Электронный ресурс] / Tang Duyu, Qin Bing, Liu Ting. – Режим доступа : URL: https://arxiv.org/abs/1610.08815 (датаобращения: 04.12.2017).
19. Википедия: Вероятностный латентно-семантический анализ [Электронный ресурс]. – Режим доступа : URL: https://ru.wikipedia.org/wiki/Вероятностный_латентно-семантический_анализ (дата обращения: 04.12.2017).
20. Википедия: Распределение Дирихле [Электронный ресурс]. – Режим доступа : URL: https://ru.wikipedia.org/wiki/Распределение_Дирихле (дата обращения: 04.12.2017).
21. Википедия: Служебные слова [Электронный ресурс]. – Режим доступа : URL: https://ru.wikipedia.org/wiki/Служебные_слова (дата обращения: 04.12.2017).
22. BigARTM: State-of-the-art Topic Modeling [Электронный ресурс]. – Режим доступа : URL: http://bigartm.org/ (дата обращения: 04.12.2017).
23. Воронцов К. В. Вероятностное тематическое моделирование: обзор моделей и регуляризационный подход [Электронный ресурс] / К.В. Воонцов. – Режим доступа : URL: http://www.machinelearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf (дата обращения: 04.12.2017).
24. Lin Henry W. Criticality in Formal Languages and Statistical Physics [Электронный ресурс] / Lin Henry W. and Tegmark Max. – Режим доступа : URL: https://arxiv.org/pdf/1606.06737.pdf(дата обращения: 04.12.2017).
25. Levy Omer. Neural Word Embedding as Implicit Matrix Factorization [Электронныйресурс] / Levy Omer, Goldberg Yoav. – Режим доступа : URL: http://papers.nips.cc/paper/5477-neural-word-embedding-as-implicitmatrix-factorization (дата обращения: 04.12.2017).
26. Холоденко А. Б. О построении статистических языковых моделей для систем распознавания русской речи [Электронный ресурс] / А.Б. Холоденко. – Режим доступа : URL: http://intsys.msu.ru/invest/speech/articles/rus_lm.htm (дата обращения: 04.12.2017).
27. Харламов А. А. Формирование n-граммной тематической модели текста [Электронный ресурс] / А.А. Харламов // Речевые технологии. – 2016. – № 1–2. – М. – Режим доступа : http://speechtechnology.ru/files/1-2016.pdf
28. The Reactor: Afastandsample-efficientActor-CriticAgentforReinforcementLearning [Электронный ресурс]. – Режим доступа : URL: https://openreview.net/forum?id=rkHVZWZAZ&noteId=rkHVZWZAZ (дата обращения: 04.12.2017).
29. An inference-based policy gradient method for learning options [Электронный ресурс]. – Режим доступа : URL: https://openreview.net/forum?id=rJIgf7bAZ&noteId=rJIgf7bAZ (дата обращения: 04.12.2017).
30. Automatic Goal Generation for Reinforcement Learning Agents [Электронный ресурс]. – Режим доступа : URL: https://openreview.net/forum?id=SyhRVm-Rb&noteId=SyhRVm-Rb (дата обращения: 04.12.2017).
31. Time Limits in Reinforcement Learning [Электронный ресурс]. – Режим доступа : URL: https://openreview.net/forum?id=HyDAQl-AW&noteId=HyDAQl-AW (дата обращения: 04.12.2017).
32. Харламов А. А. Ассоциативная память – среда для формирования пространства знаний: От биологии к приложениям (RussianEdition) [Электронный ресурс] / А.А. Харламов. – Режим доступа : URL: https://www.amazon.com/Ассоциативная-память-формирования-пространства-приложениям/dp/363 9645499 (дата обращения: 04.12.2017).
33. Акопов Р. Н. Теория мозга: формирование высших функций головного мозга человека [Электронный ресурс] / Р.Н. Акопов. – Режим доступа : URL: https://books.google.ru/books?id=HhEtDwAAQBAJ&pg=PA76&lpg=PA76&dq=Структурная+обработка+информации+Кора+полушарий+головного+мозга+человека&source=bl&ots=pHKkXKxJPy&sig=DTTHhr0OVEYUx-adoAms-_YQfV8&hl=ru&sa=X&ved=0ahUKEwjbr_f6jKPXAhWmNJoKHTf2CEYQ6AEITTAJ#v=onepage&q=Структурная обработка информации Кора полушарий головного мозга человека&f=false (дата обращения: 04.12.2017)
34. Stachenfeld Kimberly L. The hippocampus as a predictive map [Электронный ресурс] / L. Stachenfeld Kimberly, M. Botvinick Matthew & J. Gershman Samuel // Nature Neuroscience. – 2017. – № 20. – P. 1643–1653. – Режим доступа : URL: https://www.nature.com/articles/nn.4650 (дата обращения: 04.12.2017)

35. Now I Remember! Episodic Memory for Reinforcement Learning [Электронный ресурс]. – Режим доступа : URL: https://openreview.net/forum?id=SJxE3jlA-&noteId=SJxE3jlA- (дата обращения: 04.12.2017).
36. Memory Architectures in Recurrent Neural Network Language Models [Электронный ресурс]. – Режим доступа : URL: https://openreview.net/forum?id=SkFqf0lAZ&noteId=SkFqf0lAZ (дата обращения: 04.12.2017).
37. Benefits of Depth for Long-Term Memory of Recurrent Networks [Электронный ресурс]. – Режим доступа : URL: https://openreview.net/forum?id=HJ3d2Ax0-&noteId=HJ3d2Ax0- (дата обращения: 04.12.2017).
38. Neural Map: Structured Memory for Deep Reinforcement Learning [Электронный ресурс]. – Режим доступа : URL: https://openreview.net/forum?id=Bk9zbyZCZ&noteId=Bk9zbyZCZ (дата обращения: 04.12.2017).
39. Deep Generative Dual Memory Network for Continual Learning [Электронный ресурс]. – Режим доступа : URL: https://openreview.net/forum?id=BkVsWbbAW&noteId=BkVsWbbAW (дата обращения: 04.12.2017).
40. Ostmeyer Jared. Machine Learning on Sequential Data Using a Recurrent Weighted Average [Электронный ресурс] / Ostmeyer Jared, Cowell Lindsay. – Режим доступа : URL: https://arxiv.org/pdf/1703.01253.pdf (дата обращения: 04.12.2017).
41. Compositional Attention Networks for Machine Reasoning [Электронный ресурс]. – Режим доступа : URL: https://openreview.net/forum?id=S1Euwz-Rb&noteId=S1Euwz-Rb (дата обращения: 04.12.2017).
42. Novelty Detection with GAN [Электронный ресурс]. – Режим доступа : URL: https://openreview.net/forum?id=Hy7EPh10W&noteId=Hy7EPh10W (дата обращения: 04.12.2017).
43. Efficiently applying attention to sequential data with the Recurrent Discounted Attention unit [Электронный ресурс]. – Режим доступа : URL: https://openreview.net/forum?id=BJ78bJZCZ&noteId=BJ78bJZCZ (дата обращения: 04.12.2017).
44. A Painless Attention Mechanism for Convolutional Neural Networks [Электронный ресурс]. – Режим доступа : URL: https://openreview.net/forum?id=rJe7FW-Cb&noteId=rJe7FW-Cb (дата обращения: 04.12.2017).
45. Vinyals Oriol. Matching Networks for One Shot Learning [Электронный ресурс] / Vinyals Oriol, Blundell Charles et al. – Режим доступа : URL: https://deepmind.com/research/publications/matching-networks-one-shot-learning/ (дата обращения: 04.12.2017).
46. Santoro Adam. One-shot Learning with Memory-Augmented Neural Networks [Электронный ресурс] / Santoro Adam, Bartunov Sergey et al. – Режим доступа : URL: https://arxiv.org/abs/1605.06065 (дата обращения: 04.12.2017).
47. Li Jiwei. Deep Reinforcement Learning for Dialogue Generation [Электронный ресурс] / Li Jiwei, MonroeWill et al. – Режим доступа : URL: https://aclweb.org/anthology/D16-1127 (дата обращения: 04.12.2017).
48. Li Jiwei. Dialogue Learning with Human-in-the-Loop [Электронный ресурс] / Li Jiwei, Miller Alexander H. et al. – Режим доступа : URL:https://arxiv.org/pdf/1611.09823.pdf (дата обращения: 04.12.2017).
49. Liu B. Attention-based recurrent neural network models for joint intent detection and slot filling [Текст] / B. Liu, I. Lane // arXiv preprint arXiv:1609.01454. – 2016.
50. Li Jiwei. Learning through Dialogue Interactions by Asking Questions [Электронный ресурс] / Li Jiwei, Miller Alexander H. et al. – Режим доступа : URL: https://arxiv.org/pdf/1612.04936.pdf (дата обращения: 04.12.2017).
51. Hewlett Eunsol. Coarse-to-Fine Question Answering for Long Documents [Электронный ресурс] / Hewlett Eunsol, Choi Daniel, Lacost Alexandre et al. – Режим доступа : URL: https://arxiv.org/pdf/1611.01839.pdf (дата обращения: 04.12.2017).
52. Graves Alex. Hybrid computing using a neural network with dynamic external memory [Электронный ресурс] / Graves Alex, Wayne Greg et al. – Режим доступа : URL: https://www.nature.com/articles/nature20101 (дата обращения: 04.12.2017).
53. Bordes Antoine. Learning End-to-End Goal-Oriented Dialog [Электронный ресурс] / Bordes Antoine, Boureau Y-Lan & Weston Jason. – Режим доступа : URL:https://arxiv.org/pdf/1605.07683.pdf (дата обращения: 04.12.2017).
54. Text Analyst 2.0 [Электронный ресурс]. – Режим доступа : URL:http://www.analyst.ru/index.php?lang=eng&dir=content/products/&id=ta (дата обращения: 04.12.2017).

55. Graves Alex. Differentiable neural computers [Электронный ресурс] / Graves Alex, Wayne Greg et al. – Режим доступа : URL: https://deepmind.com/blog/differentiable-neural-computers/ (дата обращения: 04.12.2017).
56. Kowalke Peter. Five CRM Innovations You'll See in 2018 / Kowalke Peter ; пер. с англ. «Пять инноваций CRM, которые вы увидите в 2018 году» [Электронный ресурс] // Сайт Хабрахабр. – 8 ноября 2017 в 18:50. – Режим доступа : URL: https://habrahabr.ru/post/341968/ (дата обращения: 5.12.2017).
57. Alexander A. Kharlamov. The Language Model of the World and Purposeful Human Behavior [Текст] / Alexander A. Kharlamov // Journal of Brain, Behaviour and Cognitive Sciences. – 2018. – Vol. 1, № 2:11. – P. 1–5.
58. Харламов А. А. Семантическая сеть как модель мира и целенаправленное поведение [Текст] / А.А. Харламов // Проблемы искусственного интеллекта. – Донецк : ГУ ИПИИ. – 2019. – № 1 (12). – С. 71–85.

References

1. Antonov A.Yu., Soloviev A.N. *Metod uslovnykh sluchaynykh poley v zadachakh obrabotki russkoyazychnykh tekstov. «Informatsionnyye tekhnologii i sistemy* [The method of conditional random fields in the tasks of processing Russian texts. “Information Technologies and Systems – 2013”], Kaliningrad, 2013.
2. McKinsey Global *Institute, Artificial Intelligence The Next Digital Frontier?* [Electronic resource] // URL: http://www.mckinsey.com/~/media/mckinsey/industries/advanced electronics / our insights / mg-artificial-intelligence-discussion-paper .ashx (dated April 12, 2017).
3. WangTim, Mizuho *Industry Focus. A Survey of U.S. Healthcare IT Industry Landscape* [Electronic Resource] // 07.28.2016 URL: https://www.mizuhobank.com/fin_info/industry/pdf/mif_183.pdf
4. *Software and Systems Analysis and Forecasts* [Electronic resource] // URL: https://www.reportlinker.com/p05069690/Natural- Language-Processing-Enterprise-Applications-for-Natural-Language-Technologies-Processing-Understanding-Generation-Software-and-Systems-Market-Analysis-and-Forecasts.html (data circulation: 04.12.2017).
5. Ruder Sebastian, *Word embeddings in 2017: Trends and future directions* [Electronic resource] // URL: http://www.ruder.io/word-embeddings-2017/index.html (data circulation: 04.12.2017).
6. *Facebook Research, Fast Text* [Electronic resource] // Site: fastText. Library for efficient text classification and representation learning, URL: https://www.fasttext.cc/ (referral date: 12/04/2017).
7. *Lifelong Word Embedding via Meta-Learning* [Electronic resource] // URL: https://openreview.net/forum?id=H1BO9M-0Z (data address: 12/04/2017).
8. *Word2net: Deep Representations of Language* [Electronic Resource] // URL: https://openreview.net/forum?id=SkJd_y-Cb&noteId=SkJd_y-Cb (data circulation: 12/04/2017).
9. Kharlamov AA, Le Manh Ha; *Neyrosetevyye podkhody k klassifikatsii tekstov na osnove morfologicheskogo analiza* [Neural network approaches to the classification of texts based on morphological analysis] [Electronic resource] // URL: https://mipt.ru/science/trudy/soderzhanie-zhurnala-trudy-mfti-tom-9-2-34-2017.php (circulation date: 12/04/2017).
10. Socher R, Pennington J, Huang EH, Ng AY, Manning CD. *Semi-supervised recursiveautoencoders for predicting sentiment distributions*. Jul 27. Association for Computational Linguistics. P. 151–161.
11. Kiros Ryan, Zhu Yukun et al .; *Skip-Thought Vectors* [Electronic resource] // URL: https://arxiv.org/abs/1506.06726 (data circulation: 12/04/2017).
12. Le Q., Mikolov T. *Distributed by International Conference on Machine Learning*, 2014, p. 1188-1196.
13. Kouloumpis Efthymios, WilsonTheresa et al.; *Twitter Sentiment Analysis: The Electronic Resource*] // URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/ download / 2857/3251 (data circulation: 12/04/2017).
14. Zhao Rui, Zhou Anna, Mao Kezhi; Automatic URL: http://dl.acm.org/citation.cfm?id=2849567 (data circulation: 04.12.2017).
15. Bolotova V. V., Blinov V. A. et al.; Which IR Model Has a Better Sense of Humor? Search over a Large Collection of Jokes [Electronic resource] // URL: http://www.dialog-21.ru/media/3905/bolotovavvetal.pdf (appeal data: 12/04/2017).
16. Poria Soujanya, Cambria Erik, Hazarika Devamanyu, Vij Prateek, *A Deeper Look into Sarcastic Tweets Using Deep Convolutional Neural Networks* [Electronic resource] // URL: https://arxiv.org/abs/1610.08815 (data: 04.12.2017).
17. Kim Yoon *Convolutional Neural Networks for Sentence Classification* [Electronic Resource] // URL: https://arxiv.org/abs/1610.08815 (data: 04.12.2017).

18. Tang Duyu, Qin Bing, Liu Ting, Document *Modeling with Gated Recurrent Neural Network for Sentiment Classification* [Electronic Resource] // URL: https://arxiv.org/abs/1610.08815 (data circulation: 12/04/2017).
19. *Vikipediya: Veroyatnostnyy latentno-semanticheskiy analiz* [Wikipedia: Probabilistic latent-semantic analysis] [Electronic resource] // URL: https://ru.wikipedia.org/wiki/ Probability-latent-semantic_analysis (access date: 04.12.2017).
20. *Vikipediya: Raspredeleniye Dirikhle* [Wikipedia: Dirichlet distribution] [Electronic resource] // URL: https://ru.wikipedia.org/wiki/ Distribution_Dirichlet (access date: 04.12.2017).
21. *Vikipediya: Sluzhebnyye slova* [Wikipedia: Official words] [Electronic resource] // URL: https://ru.wikipedia.org/wiki/Servidables (reference date: 04.12.2017).
22. *BigARTM: State-of-the-art Topic Modeling* [Electronic resource] // URL: http://bigartm.org/ (data circulation: 12/04/2017).
23. Vorontsov K.V. *Veroyatnostnoye tematicheskoye modelirovaniye: obzor modeley i regulyarizatsionnyy podkhod* [Probabilistic thematic modeling: a review of models and a regularization approach] [Electronic resource] // URL: http://www.machinelearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf (access date: 04.12.2017).
24. Lin Henry W. and Tegmark Max, *Criticality in Formal Languages and Statistical Physics* [Electronic Resource] // URL: https://arxiv.org/pdf/1606.06737.pdf (return data: 04.12.2017).
25. Levy Omer. *Neural Word Embedding as Implicit Matrix Factorization* [Elektronnyy resurs] / Levy Omer, Goldberg Yoav, Rezhim dostupa : URL: http://papers.nips.cc/paper/5477-neural-word-embedding-as-implicit???matrix-factorization (data obrashcheniya: 04.12.2017).
26. Kholodenko A. B. *O postroyenii statisticheskikh yazykovykh modeley dlya sistem raspoznavaniya russkoy rechi* [On the construction of statistical language models for Russian speech recognition systems] [Elektronnyy resurs] Rezhim dostupa : URL: http://intsys.msu.ru/invest/speech/articles/rus_lm.htm (data obrashcheniya: 04.12.2017).
27. Kharlamov A. A. *Formirovaniye n-grammnoy tematicheskoy modeli teksta* [Elektronnyy resurs], *Rechevyye tekhnologii* [Speech technology], 2016, No. 1–2, M. Rezhim dostupa : http://speechtechnology.ru/files/1-2016.pdf
28. *The Reactor: Afastandsample-efficient Actor-Critic Agent for Reinforcement Learning* [Elektronnyy resurs]. –
Rezhim dostupa : URL: https://openreview.net/forum?id=rkHVZWZAZ&noteId=rkHVZWZAZ (data obrashcheniya: 04.12.2017).
29. *An inference-based policy gradient method for learning options* [Elektronnyy resurs]. Rezhim dostupa : URL: https://openreview.net/forum?id=rJIgf7bAZ&noteId=rJIgf7bAZ (data obrashcheniya: 04.12.2017).
30. *Automatic Goal Generation for Reinforcement Learning* Agents [Elektronnyy resurs]. Rezhim dostupa : URL: https://openreview.net/forum?id=SyhRVm-Rb&noteId=SyhRVm-Rb (data obrashcheniya: 04.12.2017).
31. *Time Limits in Reinforcement Learning* [Elektronnyy resurs].
Rezhim dostupa : URL: https://openreview.net/forum?id=HyDAQl-AW&noteId=HyDAQl-AW (data obrashcheniya: 04.12.2017).
32. Kharlamov A. A. *Assotsiativnaya pamyat' – sreda dlya formirovaniya prostranstva znaniy: Ot biologii k prilozheniyam* [Associative memory - an environment for the formation of a space of knowledge: From biology to applications] (Russian Edition) [Elektronnyy resurs],
Rezhim dostupa : URL: https://www.amazon.com/Assotsiativnaya-pamyat'-formirovaniya-prostranstva-prilozheniyam/dp/363 9645499 (data obrashcheniya: 04.12.2017).
33. Akopov R. N. *Teoriya mozga: formirovaniye vysshikh funktsiy golovnogo mozga cheloveka* [Theory of the brain: the formation of higher functions of the human brain] [Elektronnyy resurs] / Rezhim dostupa : URL: https://books.google.ru/books?id=HhEtDwAAQBAJ&pg=PA76&lpg=PA76&dq=Strukturnaya+obrabotka+informatsii+Kora+polushariy+golovnogo+mozga+cheloveka&source=bl&ots=pHKkXKxJPy&sig=DTTHhr0OVEYUx-adoAms-_YQfV8&hl=ru&sa=X&ved=0ahUKEwjbr_f6jKPXAhWmNJoKHTf2CEYQ6AEITTAJ#v=onepage&q=Strukturnaya obrabotka informatsii Kora polushariy golovnogo mozga cheloveka&f=false (data obrashcheniya: 04.12.2017)
34. Stachenfeld Kimberly L. The hippocampus as a predictive map [Elektronnyy resurs] / L. Stachenfeld Kimberly, M. Botvinick Matthew & J. Gershman Samuel. *Nature Neuroscience*, 2017, No. 20, R. 1643–1653, Rezhim dostupa : URL: https://www.nature.com/articles/nn.4650 (data obrashcheniya: 04.12.2017)
35. *Now I Remember! Episodic Memory for Reinforcement Learning* [Elektronnyy resurs], Rezhim dostupa : URL: https://openreview.net/forum?id=SJxE3jlA-&noteId=SJxE3jlA- (data obrashcheniya: 04.12.2017).

36. *Memory Architectures in Recurrent Neural Network Language Models* [Elektronnyy resurs], Rezhim dostupa : URL: https://openreview.net/forum?id=SkFqf0lAZ&noteId=SkFqf0lAZ (data obrashcheniya: 04.12.2017).
37. *Benefits of Depth for Long-Term Memory of Recurrent Networks* [Elektronnyy resurs], Rezhim dostupa : URL: https://openreview.net/forum?id=HJ3d2Ax0-&noteId=HJ3d2Ax0- (data obrashcheniya: 04.12.2017).
38. *Neural Map: Structured Memory for Deep Reinforcement Learning* [Elektronnyy resurs], Rezhim dostupa : URL: https://openreview.net/forum?id=Bk9zbyZCZ&noteId=Bk9zbyZCZ (data obrashcheniya: 04.12.2017).
39. *Deep Generative Dual Memory Network for Continual Learning* [Elektronnyy resurs] Rezhim dostupa : URL: https://openreview.net/forum?id=BkVsWbbAW&noteId=BkVsWbbAW (data obrashcheniya: 04.12.2017).
40. Ostmeyer Jared. *Machine Learning on Sequential Data Using a Recurrent Weighted Average* [Elektronnyy resurs] / Ostmeyer Jared, Cowell Lindsay, Rezhim dostupa : URL: https://arxiv.org/pdf/1703.01253.pdf (data obrashcheniya: 04.12.2017).
41. *Compositional Attention Networks for Machine Reasoning [*Elektronnyy resurs], Rezhim dostupa : URL: https://openreview.net/forum?id=S1Euwz-Rb&noteId=S1Euwz-Rb (data obrashcheniya: 04.12.2017).
42. *Novelty Detection with GAN* [Elektronnyy resurs], Rezhim dostupa : URL: https://openreview.net/forum?id=Hy7EPh10W&noteId=Hy7EPh10W (data obrashcheniya: 04.12.2017).
43. *Efficiently applying attention to sequential data with the Recurrent Discounted Attention unit* [Elektronnyy resurs]. Rezhim dostupa : URL: https://openreview.net/forum?id=BJ78bJZCZ&noteId=BJ78bJZCZ (data obrashcheniya: 04.12.2017).
44. *A Painless Attention Mechanism for Convolutional Neural Networks* [Elektronnyy resurs]. Rezhim dostupa : URL: https://openreview.net/forum?id=rJe7FW-Cb&noteId=rJe7FW-Cb (data obrashcheniya: 04.12.2017).
45. Vinyals Oriol. *Matching Networks for One Shot Learning* [Elektronnyy resurs] / Vinyals Oriol, Blundell Charles et al. – Rezhim dostupa : URL: https://deepmind.com/research/publications/matching-networks-one-shot-learning/ (data obrashcheniya: 04.12.2017).
46. Santoro Adam. *One-shot Learning with Memory-Augmented Neural Networks* [Elektronnyy resurs] / Santoro Adam, Bartunov Sergey et al. Rezhim dostupa : URL: https://arxiv.org/abs/1605.06065 (data obrashcheniya: 04.12.2017).
47. Li Jiwei. *Deep Reinforcement Learning for Dialogue Generation* [Elektronnyy resurs] / Li Jiwei, MonroeWill et al. Rezhim dostupa : URL: https://aclweb.org/anthology/D16-1127 (data obrashcheniya: 04.12.2017).
48. Li Jiwei. *Dialogue Learning with Human-in-the-Loop* [Elektronnyy resurs] / Li Jiwei, Miller Alexander H. et al. Rezhim dostupa : URL:https://arxiv.org/pdf/1611.09823.pdf (data obrashcheniya: 04.12.2017).
49. Liu B. *Attention-based recurrent neural network models for joint intent detection and slot filling [*Tekst] / B. Liu, I. Lane // arXiv preprint arXiv:1609.01454. – 2016.
50. Li Jiwei. *Learning through Dialogue Interactions by Asking Questions* [Elektronnyy resurs] / Li Jiwei, Miller Alexander H. et al. Rezhim dostupa : URL: https://arxiv.org/pdf/1612.04936.pdf (data obrashcheniya: 04.12.2017).
51. Hewlett Eunsol. *Coarse-to-Fine Question Answering for Long Documents* [Elektronnyy resurs] / Hewlett Eunsol, Choi Daniel, Lacost Alexandre et al. Rezhim dostupa : URL: https://arxiv.org/pdf/1611.01839.pdf (data obrashcheniya: 04.12.2017).
52. Graves Alex. *Hybrid computing using a neural network with dynamic external memory* [Elektronnyy resurs] / Graves Alex, Wayne Greg et al. – Rezhim dostupa : URL: https://www.nature.com/articles/nature20101 (data obrashcheniya: 04.12.2017).
53. Bordes Antoine. *Learning End-to-End Goal-Oriented Dialog* [Elektronnyy resurs] / Bordes Antoine, Boureau Y-Lan & Weston Jason. Rezhim dostupa : URL:https://arxiv.org/pdf/1605.07683.pdf (data obrashcheniya: 04.12.2017).
54. *Text Analyst 2.0* [Elektronnyy resurs]. Rezhim dostupa : URL:http://www.analyst.ru/index.php?lang=eng&dir=content/products/&id=ta (data obrashcheniya: 04.12.2017).
55. Graves Alex. *Differentiable neural computers* [Elektronnyy resurs] / Graves Alex, Wayne Greg et al. Rezhim dostupa : URL: https://deepmind.com/blog/differentiable-neural-computers/ (data obrashcheniya: 04.12.2017).
56. Kowalke Peter. *Five CRM Innovations You'll See in 2018* / Kowalke Peter ; per. s angl. «Pyat' innovatsiy CRM, kotoryye vy uvidite v 2018 godu» [Elektronnyy resurs] // Sayt Khabrakhabr. 8 noyabrya 2017 v 18:50. Rezhim dostupa : URL: https://habrahabr.ru/post/341968/ (data obrashcheniya: 5.12.2017).
57. Alexander A. Kharlamov. The Language Model of the World and Purposeful Human Behavior [Tekst] / *Journal of Brain, Behaviour and Cognitive Sciences*, 2018, Vol. 1, No. 2:11. P. 1–5.

58. Kharlamov A. A. Semanticheskaya set' kak model' mira i tselenapravlennoye povedeniye [The semantic network as a model of the world and purposeful behavior] *Problemy iskusstvennogo intellekta* [International Peer-Reviewed Scientific Journal «Problems of Artificial Intelligence], ISSN 2413-7383, 2019, No. 1 (12), pp. 71–85.

RESUME

A. A. Kharlamov, D. I. Gordeev

Distributive vs Network Semantics in Dialog Systems

In the last 8 years, we have witnessed an increased interest to the field of dialogue agents. This is largely due to the application of machine learning to the tasks of natural language processing. Distributional and network semantics (e.g. Word2Vec, FastText) make it possible to use generalized data from huge text corpora, which was problematic before with n-grams. Also new language models (BERT, ELMo, ULMFiT) fine-tuned with data from huge corpora can significantly reduce the cost of model training for new tasks (transfer learning), and in some cases they let get rid of it completely (zero-shot learning). In addition, this chapter discusses popular neural network architectures (LSTM, Transformer) and promising approaches to the use of neural networks for tasks of natural language processing and dialogue agents. Also this paper provides an overview of the modular approach to dialogue agents, where each model carries on with its own task and the main meta-module combines the results of the children models. The main types of modules are considered. This work also briefly discusses thematic text modelling (mostly unsupervised) and highlights the latest advances in syntactic and morphological language modelling, as well as outlines the latest ideas in machine learning that are at least partially inspired by human cognitive structures.

РЕЗЮМЕ

А. А. Харламов, Д. И. Гордеев

Дистрибутивная vs сетевая семантика в диалоговых системах

В последние 8 лет мы стали свидетелями возросшего интереса к сфере диалоговых агентов. Это во многом связано с применением машинного обучения к задачам обработки естественного языка. Дистрибутивная и сетевая семантика (например, Word2Vec, FastText) позволяют использовать обобщенные данные из огромных текстовых корпусов, что было проблематично раньше с n-grams. Также новые языковые модели (BERT, ELMo, ULMFiT), доработанные с учетом данных из огромных корпусов, могут существенно снизить стоимость обучения модели для новых задач (transferlearning), а в некоторых случаях и вовсе избавиться от неё (zero-shotlearning). Помимо этого, в данной главе рассматриваются популярные архитектуры нейронных сетей (LSTM, Transformer) и перспективные подходы к использованию нейронных сетей для задач обработки естественного языка и диалоговых агентов. Также в статье представлен обзор модульного подхода к диалоговым агентам, где каждая модель выполняет свою задачу, а главный мета-модуль объединяет результаты дочерних моделей. Рассмотрены основные типы модулей. В данной работе также кратко обсуждается тематическое моделирование текста (в основном неконтролируемое) и освещаются последние достижения в синтаксическом и морфологическом моделировании языка, а также излагаются последние идеи в области машинного обучения, которые, по крайней мере, частично вдохновлены когнитивными структурами человека.

Статья поступила в редакцию 15.02.2019.