Доказательство или опровержение данной гипотезы будет являться предметом научного исследования.

Литература

1. Об эффективности обмена информацией между САПР // Universum: Технические науки : электрон. научн. журн. Райкин Л.И. [и др.]. 2014. № 2 (3). — Режим доступа: URL: http://7universum.com/en/tech/archive/item/1034 (дата обращения: 15.02.2018).

2. Dr. Arnulf Frohlich Сравнение 3D-форматов. Исследование компании PROSTEP / Frohlich Dr. Arnulf // CAD/CAM/CAE Observer. - №4, 2011. – С. 53-62.

3. Малюх, В. Proficiency — параметрические инструменты для трансляции данных. Isicad, № 73 (8), 2010 / [Электронный ресурс]. – Режим доступа: URL: http://isicad.ru/ru/articles.php?article_num=13930 (дата обращения: 15.02.2018).

4. Малюх, В. Форматы данных: кто виноват и как с этим бороться? Isicad, № 79 (2), 2011 / [Электронный ресурс]. – Режим доступа: URL: http://isicad.ru/ru/articles.php?article_num=14227 (дата обращения: 15.02.2018).

5. Системы автоматизации производства и их интеграция. Представление данных об изделии. Часть 1. Общие представления и основополагающие принципы: ГОСТ Р ИСО 10301-1-99. – Введ. 1999-09-22. – М.: ИПК Издательство стандартов, 1999. - 16 с.

6. Крайнов, В.В. Анализ формата передачи данных STEP. / В.В. Крайнов, М.В. Пономарёв, И.Н. Фролова // Труды Нижегородского государственного технического университета им. Р.Е. Алексеева. - №5, 2013. – С. 129-134.

М.А. Ситникова

ИССЛЕДОВАНИЕ ВОЗМОЖНОСТИ ВСТРАИВАНИЯ ДОПОЛНЕННОЙ РЕАЛЬНОСТИ В ВЕБ-ПРИЛОЖЕНИЯ БЕЗ ИСПОЛЬЗОВАНИЯ СПЕЦИАЛЬНЫХ МАРКЕРОВ

(Самарский национальный исследовательский университет
им. академика С.П. Королева)

Что такое дополненная реальность? Каким образом используется дополненная реальность в онлайн приложениях? Возможно ли встраивание дополненной реальности в веб-приложения? При помощи какого инструмента это можно реализовать? Именно этот круг вопросов рассматривается в настоящей работе.

Сам термин «дополненная реальность» был предложен исследователем корпорации Boeing Томом Коделом в 1990 году. Существует несколько опреде-

лений дополненной реальности. Исследователь Рональд Азума в 1997 году определил её как систему, которая:

- совмещает виртуальное и реальное;
- взаимодействует в реальном времени;
- работает в 3D.

Данная работа очень актуальна в настоящее время, т.к. сервисы дополненной реальности используются в таких областях, как медицина, военная техника, компьютерные игры, кинематография и телевидение.

Каким же образом используются сервисы дополненной реальности? Ответом на этот вопрос могут послужить такие примеры как:

- сотрудничество Microsoft с автопроизводителем Volvo (технология HoloLens помогает клиентам выбирать подходящую конфигурацию автомобиля);
- проект Holoroom сети магазинов товаров для дома Lowe (технология позволяет оценить будущий дизайн кухни или ванной комнаты);
- всеми известная виртуальная примерочная компании IKEA (технология позволяет «примерить» мебель не выходя из дома).

Фрагмент виртуальной примерочной компании IKEA изображен на рисунке 1.

Рис. 1. Фрагмент виртуальной примерочной IKEA

Для построения сервиса дополненной реальности используют различные технологии, облегчающие его создание. Например, Vuforia, ARToolKit, Kudan, Catchoom, Aurasma, InfinityAR. Все эти инструменты реализуют дополненную реальность на основе обнаруженного маркера Aruco.

Однако, перед нами стоит следующая задача: встраивание дополненной реальности в веб-приложение без использования специальных маркеров.

Рассмотрим два различных способа решения данной задачи и реализуем один из них. В обоих способах подразумевается, что в качестве маркера будет использован белый лист размера А4 – подручный материал, который есть у каждого.

Первый способ заключается в написании программы на высокоуровневом языке программирования – Python. Данное приложение должно распознавать как белый лист, так и маркер Aruco. Далее при помощи веб-фрейморка выполняется привязка написанной программы к веб-странице. Это значит, что приложения дополненной реальности будут работать без установки и на устрой-

стве любого формата. Одним из таких является Django – чрезвычайно популярный и полнофункциональный серверный веб-фреймворк. Затем производится наложение картинки на исходное изображение, получая таким образом дополненную реальность. Обнаружение белого листа при этом происходит следующим образом:

- загружается изображение, цвет меняют на оттенки серого и уменьшают резкость;
- выполняется преобразование BGR в HSV, определяется диапазон синего цвета в HSV;
- производится распознавание контуров, выполняется преобразование Хафа;
- оставляем линии, формирующие прямоугольник (принимаем его стороны равными 210 и 297 соответственно).

Обнаружение маркера Aruco происходит при помощи библиотеки OpenCV, а именно с использованием getPredefinedDictionary(cv2.aruco.DICT_4X4_50), где DICT_4X4_50 – необходимый маркер.

Второй способ основывается на создании веб-приложения при помощи инструментария JSARToolKit. За основу был взят ARToolKit v5.3.1, скомпилированный с C++ на JavaScript через EMscriptem — инструмента от Mozilla, который позволяет компилировать код на C/C++ и запускать его в браузерах. Он способен выводить чистый JavaScript. JSARToolKit используют в связке с three.js – библиотекой для отображения дополненной реальности с 3D графикой. В первую очередь, JSARToolKit нужен для того, чтобы запускать ARToolKit в браузере. Далее необходимо реализовать алгоритм обнаружения белого листа на языке Python, наложить на обнаруженный объект маркер Aruco, а затем выполнить привязку данного алгоритма с веб-приложением на основе JSARToolKit.

Таким образом, было рассмотрено использование дополненной реальности в онлайн приложениях. Были разобраны два способа встраивания дополненной реальности в веб-приложение, рассмотрены инструменты для осуществления каждого, а так же реализован один из них.

А.И. Соловьев

ОБРАБОТКА ЕСТЕСТВЕННЫХ ЯЗЫКОВ

(Самарский университет)

В настоящее время с бурным ростом информации и интернета значительное развитие получила автоматическая обработка текстов. Имеются существенные различия в обработке естественного языка компьютером и человеком - скорость вычислений компьютера значительно выше, и в то же время компьютер значительно хуже понимает естественный язык. Для того, чтобы компьютер лучше понимал естественный язык, необходимо сделать семантический

анализ текста. [8]

Существует множество идей для автоматизации обработки текстов на естественных языках (ЕЯ). Появляются новые сферы применения этой области, реализованы прикладные программные решения задач автоматической обработки текстов ЕЯ.

Наиболее распространенные прикладные задачи обработки ЕЯ:

Машинный перевод. Еще в середине XX века, были разработаны простые программы, в основе которых лежал пословный перевод. В наши дни существует большое количество систем автоматического перевода с разным качеством, применяющих сложные технологии.

Информационный поиск. Наиболее часто используемый функционал поисковых систем. При индексировании текстов, требуется предварительная лингвистическая обработка и создание индексных структур. Современные поисковики в интернете используют векторную модель, при котором запрос состоит из набора слов, а результат представляет выборку документов по индексированию текстов с использованием этих слов. С задачами информационного поиска связаны: реферирование, индексирование, аннотирование, рубрикация и классификация текстовых документов.

Реферирование. Для автоматического реферирования в настоящее время применяется отбор наиболее важных (значимых) предложений текста на основе лингвистических и структурных особенностей текста и используется статистика слов и словосочетаний.

Аннотирование. Для составления аннотации используют перечень ключевых тем с применением лингвистических и статистических критериев.

Классификация. При классификации каждый документ относят к заранее определенному классу с известными параметрами.

Кластеризация. При кластеризации документы разбиваются на кластеры (близкие по тематике документы).

Рубрицирование. Отнесение документа к определенной тематической рубрике.

Формирование ответов на вопросы. Решается путем поиска текстов, потенциально содержащих ответ на задаваемый вопрос.

Анализ тональности текстов и выделение мнений. Широкое применение получил в коммерческих целях и в вопросах анализа общественного мнения.

Поддержка диалога на естественном языке. Как правило, применяется в специализированных базах данных.

Редактирование текстов. Выявление орфографических и синтаксических ошибок в тексте.

Обучение естественному языку. Разработаны программы обучения морфологии, лексики, словари и т.д.

Автоматическая генерация текстов. Спецификой этого направления является автоматический перевод на несколько языков исходя из специфики документа.

Распознавание и синтез речи. Возникающие неизбежно при этом ошибки

автоматически исправляются на основе морфологических моделей и словарей.

Для решения перечисленных задач часто применяют методы машинного обучения.

При обработке текстов естественного языка выполняется:

1. Сегментация- в тексте выделяются предложения и токенов;

2. Морфологический анализ – токены (словоформы) переводятся к леммам или основам слов;

3. Синтаксический анализ - выявляются синтаксические связи и грамматические структуры предложений.

4. Семантический анализ;

WORD2VEC

В настоящее время выделяют следующие виды машинного обучения: деревья решений, наивный байесовский классификатор, логистическая регрессия, метод условных случайных полей, скрытые модели Маркова и нейронные сети. При извлечении информации часто используют методы обучения с учителем (строится модель на обучающей выборке и затем применяется для новых текстов) [2]

В связи с развитием социальных сетей, веб-порталов обработка текстовых данных играет большую роль для разных коммерческих и социологических целей. Повсеместно активно развиваются модели и алгоритмы NLP (Natural Language Processing).

Для NLP основной задачей является исследование моделей представления текстов удобной для автоматической компьютерной обработки. В связи с вышеизложенным, происходит постоянное совершенствование моделей, которые приближены к умственной деятельности человека.

В 2013 году благодаря работам Томаса Миколова возникла идея векторного представления слов с помощью нейронных сетей. Т.Миколов разработал инструмент Word2vec для создания моделей нейронных сетей (предсказательных моделей) [5]

В Word2Vec с помощью алгоритма Skip-Gram(SG), на основании текущего слова предсказываются близлежащие (слова, встречающиеся в похожем контексте) слова.

Word2vec — набор алгоритмов для расчета векторных представлений слов, реализущий две основные архитектуры — «непрерывный мешок со словами» (Skip-gram и Continuous Bag of Words-CBOW).

Обучение Word2vec происходит на большом массиве данных (корпусе). Каждое слово представлено в виде специального вектора (координаты слова). Делается предположение, что слова, близкие по смыслу расположены рядом. CBOW предназначена для предугадывания слов исходя из окружающих его слов.

На основе Word2vec можно построить дистрибутивно-семантическую модель(модель семантических связей между словами). [11]

Различают два основных подходов к моделированию семантики:

1. «Сверху вниз» - подход, построенный на знаниях, очень трудоемкий, требу-

ющий огромных затрат человеческих ресурсов;

2. «Снизу-вверх» - при таком подходе значение извлекается из особенностей употребления слов в тексте;

Для машинного обучения характерен второй подход. При использовании технологии Word2vec задается размерность векторов, заполненная случайными величинами. Во время обучения эти значения будут изменяться. Вектора близких слов будут максимально близки по значению.

Этапы алгоритма векторизации:

1. Синтаксический анализ предложений;
2. Предварительная обработка текста (например, удаление стоп-слов);
3. Процедура связывания слов в предложении (gram-skip);
4. Обработка нейронной сетью;

Данный подход хорошо работает для кластеризации формальных документов (патенты, статьи, новости), но для художественных произведений не подходит. [4]

Принцип работы заключается в поиске связей между контекстами слов. Близкие по контексту слова могут быть семантически близкими. При обучении нейронной сети каждое слово заменяется номером семантической группы – таким образом происходит предсказание контекста данного слова. Для каждого слова векторы фиксированной длины, объединяются, используя кластеризацию.

КЛАСТЕРИЗАЦИЯ И TF-IDF

При компьютерном анализе теста при наличии неоднозначных смысловых конструкций затруднителен анализ смысла текста. При кластеризации слов происходит устранение морфологической неоднозначности и машинный анализ текста по семантическим характеристикам, с учетом всевозможных выражений, которые ранее были доступны лишь человеку. Word2vec создает кластеры сходных по смыслу слов. Используя алгоритмы кластеризации создаются центры кластеров. Количество кластеров определяется эмпирическим путем. [8]

В процессе кластеризации текст переводится в векторное представление и в последующем к ним применяются методы кластеризации, основанные на расстоянии между словами.

Для снижения влияния длины текста на его признаки (признаку соответствует одна n-грамма), используется нормализация количества вхождений на единицу размера текста. При этом каждый признак принимает вид TF (term frequency), который считается как отношение количества вхождений соответствующей n-граммы к общему количеству слов текста. В TF-IDF предполагается, что значимость n-граммы прямо пропорциональна частоте ее появления в документе и обратно пропорциональна в наборе документов, в которых она встречается. Наибольший вес получает n-грамма, чаще всего встречающаяся в данном документе. Признаки документов представляют собой произведение частоты n-граммы и обратной частоты документа (inverse document frequency-IDF). [10]

ЗАКЛЮЧЕНИЕ

В статье рассмотрены вопросы практического применения компьютерной обработки естественного языка. Выявлены широкие возможности применения различных технологий обработки текстов. Рассмотрена технология Word2vec для решения поставленных задач. Преимуществом применения Word2vec является снижение вычислительных затрат при обучении. К недостаткам относится невозможность изменения векторной величины при возникновении такой необходимости для компьютерной обработки.

Данная технология позволяет применять ее для решения очень широкого круга новых задач, например, разведочного информационного поиска, тематического моделирования.

Литература

1. Беляков Д.Е., Кантор В.В. Исследование эффекта добавления негативного сэмплирования при обучении факторизационных машин в задачах построения рекомендательных систем. Информационные процессы. 2017. Т. 17. № 2. С. 159-163.

2. Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С Автоматическая обработка текстов на естественном языке и анализ данных: учеб. пособие.—М.: Изд-во НИУВШЭ,2017.—269с.

3. Буденков С.С. Семантические векторные модели текстов для анализа тональности. Научно-технический вестник Поволжья. 2017. № 2. С. 75-78.

4. Иванов Н.Н. Синтаксический разбор предложения для векторизации текста. Вопросы науки и образования. 2017. № 11 (12). С. 45-46.

5. Кириллов А.Н., Крижановский А.А. Модель геометрической фигуры синсета. Труды Карельского научного центра Российской академии наук. 2016. № 8. С. 45-54.

6. Левченко С.В. Разработка метода кластеризации слов по смысловым характеристикам с использованием алгоритмов WORD2VEC. Новые информационные технологии в автоматизированных системах. 2017. № 20. С. 44-46.

7. Машкин Д.О., Котельников Е.В. Извлечение аспектных терминов на основе условных случайных полей и векторных представлений слов. Труды Института системного программирования РАН. 2016. Т. 28. № 6. С. 223-240.

8. Мишенин А.Н., Нефедова Е.А. Анализ тональности текстов с использованием технологии WORD2VEC.. Естественные и математические науки в современном мире. 2016. № 7 (42). С. 89-97.

9. Науменко А.М., Шелудько С.Д., Юлдашев Р.Ю., Хлебников Н.О., Радыгин В.Ю. Разработка вопросно-ответной системы с нейросетевым обучением на базе современных свободных технологий. Иннов: электронный научный журнал. 2017. № 2 (31). С. 7.

10. Пархоменко П.А., Григорьев А.А., Астраханцев Н.А. Обзор и экспериментальное сравнение методов кластеризации текстов. Труды Института системного программирования РАН. 2017. Т. 29. № -2. С. 161-200.

11. Пескишева Т.А. Анализ применения дистрибутивно-семантических

моделей для пополнения словаря оценочной лексики. Научно-исследовательские публикации. 2017. № 3. С. 6-13.

12. Сбоев А.Г., Воронина И.Е., Гудовских Д.В., Селиванов А.А. Продвинутые нейросетевые модели для решения задачи определения тональности. Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии. 2016. № 4. С. 178-183.

13. Смелик Н.Д., Фильченков А.А. Мультимодальная тематическая модель текстом и изображений на основе использования их векторного представления. Машинное обучение и анализ данных. 2016. Т. 2. № 4. С. 421-441.

14. Черноусов Е.О., Чикунов Н.С. Исследование и разработка интеллектуальной системы поддержки принятия решений для службы удаленной технической поддержки на основе методов WORDEMBEDDING. Инновационная наука. 2017. № 12. С. 66-70.

15. Щербаков Д.А. Синтез и ранжирование ответов в поисковых системах типа вопрос-ответ, основанных на онтологической реккурентной структуре связанных данных. Инновационная наука. 2015. № 12-2. С. 153-160.

Е.В. Старкова, С.А. Прохоров

ИССЛЕДОВАНИЕ МЕТОДОВ КЛАССИФИКАЦИИ ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

(Самарский университет)

На сегодняшний день в условиях стремительного роста текстовой информации в электронном виде и в связи с потребностью в ней ориентироваться, все более актуальной становится проблема построения универсального классификатора текстов, предоставляющего возможность распределения исходного набора статей по нескольким заранее установленным тематикам в соответствии с их смысловым содержанием [1]. Использование такого классификатора позволит сократить трудозатраты на поиск необходимой информации, представленной электронными текстами, а также ограничить поиск относительно небольшим подмножеством документов.

Различные решения данной задачи находят свое практическое применение в таких областях, как составление тематических каталогов, фильтрация спама, классификация сайтов по тематическим каталогам, обработка документооборота и т.д. В настоящее время примерами классификатов текстов являются такие системы как NNCS (Neural Network Classification & Search), TextAnalyistPro, TextCat, CBTReader, а также проект ДИАЛИНГ, который был разработан специалистами факультета лингвистики РГГУ. Однако все они имеют ряд недостатков: во-первых, это коммерческие проекты, стоимость которых достаточно высока, а во-вторых, эти проекты рассчитаны на профессионального пользователя, следовательно, только обучение использованию предлагаемых пакетов займет слишком много времени.