В первом случае, среднее значение ошибки составило ~36%, во втором случае при тех же условиях ~28%.

Рисунок 2 – средняя ошибка.

В результате работы, был написан алгоритм обучения нейронной сети на основе ограниченных машин Больцмана для работы с изображениями. Сравнительный анализ показал, что данный подход является более эффективным, нежели обучение с сигмоидальной функцией активации.

***

1. Alex Krizhevsky - Learning Multiple Layers of Features from Tiny Images, 2009г.
2. Geoffrey Hinton, Department of Computer Science, University of Toronto - A Practical Guide to Training Restricted Boltzmann Machines, Version 1, 2010г.
3. Басов С. А., Юркин В. А., Финогенов А. Э., Сараджишвили С. Э. Система биометрической идентификации личности по 3d модели лица, получаемой с помощью стереокамер, использующая математическую модель движения мимических мышц лица // Сборник докладов студенческой научной конференции Института компьютерных наук и технологий. Издательство: Федеральное государственное автономное образовательное учреждение высшего образования "Санкт-Петербургский политехнический университет Петра Великого" (Санкт-Петербург), 2016г.

Ворожцова Н.А.

Исследование методов и систем семантического анализа текстов

ФГБОУ ВО «ИжГТУ имени М.Т. Калашникова»

(Россия, Ижевск)

doi:10.18411/spc-22-12-2017-17

idsp: 000001:spc-22-12-2017-17

Аннотация

В связи с быстрым ростом количества текстовой информации все более актуальной становится проблема автоматической обработки текстов. Одной из основных задач в рамках данной проблемы является задача семантического анализа. В работе рассмотрены методы и существующие системы семантического анализа текстов, отмечены их основные достоинства и недостатки, сделан вывод по результатам исследования.

Семантический анализ – процесс выявления смыслового содержания слов и словосочетаний в предложении. Большинство методов семантического анализа работают со смыслом слов.

Существует множество методов представления смысла высказываний, но они не являются универсальными. Например, Мельчук И.А. [1] ввел понятие лексической функции, развил понятия синтаксических и семантических валентностей и рассмотрел их в контексте толково-комбинаторного словаря, который представляет собой языковую модель. Он показал, что значения слов соотносятся не непосредственно с окружающей действительностью, а с представлениями носителя языка об этой действительности.

Многие исследователи придерживаются идеи, что семантический анализ должен выполняться после синтаксического. Рубашкин В.Ш. и Лахути Д.Г. [2] ввели иерархию синтаксических связей для более эффективной работы семантического анализатора.

Лингвист Падучева Е.В. [3] рассматривает тематические классы слов, в частности глаголов, которые несут основную смысловую нагрузку в текстах. Проблема такого подхода заключается в том, что выделение тематических классов и составление семантических словарей чрезвычайно трудоемкий процесс, сильно зависящий от индивидуального восприятия и интерпретации понятий конкретным человеком.

Высокая сложность и разнообразие форм объектов, широкий круг взаимосвязанных разноплановых задач, решаемых в мире, приводит к большим объемам перерабатываемой информации, порой непреодолимым трудностям выбора отправных точек и направлений для синтеза решений, позволяющих получить нужный результат [4].

Для работы системы семантического поиска необходимо провести комплексный анализ описания каждого документа. Комплексный анализ состоит из следующих этапов:

- 1) морфологический и морфемный анализ слов и словосочетаний в документе;
- 2) синтаксический анализ текста, который позволяет определить предметы, действия, отношения и роли слов;
- 3) семантический анализ текста, который позволяет решить проблему однозначности истолкования слов текста, анафоры, устанавливает гиперсвязи с понятиями базы знаний и общими понятиями;
- 4) построение семантических моделей текста с целью выделения значимых конструкций документов: предметы, действия, участники действия, атрибуты действия, атрибуты участников действий [5].

Рассмотрим существующие системы семантического анализа и их возможности.

АОТ (автоматическая обработка текста) [6] – сайт, представляющий разработки группы специалистов в области автоматической обработки текстов. Среди предлагаемых продуктов представлен модуль семантического анализа текстов. Модуль выделяет закономерности языка, что приводит к регулярности грамматики именных групп. Компоненты, составляющие языковую модель, – лингвистические процессоры, которые друг за другом обрабатывают входной текст. Вход одного процессора является выходом другого. Выделяются следующие компоненты: графематический анализ, морфологический анализ, синтаксический анализ, семантический анализ; построение семантического графа текста. Для каждого уровня разработан свой язык представления.

Text Analyst 2.0 [7] – это система автоматического анализа текста, выполняющая смысловой поиск информации, формирование электронных архивов, автоматическое реферирование документов. Имеет следующие основные возможности: анализ содержания текста и автоматическое формирование семантической сети с гиперссылками; анализ содержания текста и автоматическое формирование тематического дерева с гиперссылками; смысловой поиск с учетом скрытых смысловых связей слов запроса со словами текста; автоматическое реферирование текста; кластеризация информации; автоматическая индексация текста с преобразованием в гипертекст; ранжирование всех видов информации о семантике текста по «степени значимости» с возможностью варьирования детальности ее исследования, автоматическое/автоматизированное формирование полнотекстовой базы знаний с гипертекстовой структурой и возможностями ассоциативного доступа к информации.

Галактика ZOOM[8] – это автоматизированная система поиска и аналитической обработки информации, позволяющая извлекать полезные сведения из огромного объема данных. Она позволяет строить информационные портреты объектов по текстовой информации. Такой портрет состоит из статистически значимых слов и выражений, сопровождающих упоминание объекта. Галактика ZOOM обеспечивает поиск в информационных массивах с применением языка запросов, а также контекстный или тематический поиск информации с учетом морфологии.

У рассмотренных систем существует общий недостаток: они не позволяют после проведения синтаксического анализа записать результаты в базу знаний, затем проверить полученный результат на непротиворечивость и уже после этого проводить поиск по тексту.

Заключение

Обзор методов и систем семантического анализа текстов на естественном языке, выполненный в статье, показывает необходимость и принципиальную возможность создания методов, объединяющих семантический и синтаксический анализ. Возможности большинства существующих инструментов ограничиваются морфологическим и синтаксическим анализом в сочетании с методами из теории вероятностей и математической статистики. Таким образом, лишь некоторая часть наиболее простых задач оказалась решенной. Остальные проблемы по-прежнему еще предстоит решить.

Универсальный язык представления знаний должен быть удобным инструментом для вывода новых знаний из уже имеющихся, а значит, необходимо создать аппарат для проверки правильности высказываний [9]. В качестве инструмента для накопления знаний и проверки их на непротиворечивость можно использовать онтологию. Например, можно провести синтаксический анализ текста, записать полученные результаты в онтологию и запустить проверку правильности построенных связей. Инструмент для такой проверки реализован в программном средстве Protégé [10].

Некоторые исследователи предполагают, что, имея базу знаний, содержащую общесмысловую информацию о понятиях и отношениях между ними, можно будет решить проблему семантического анализа текстов. Предполагается, что при обращении к базе можно будет определить подходящий контекст высказывания в автоматическом режиме. Она помогла бы учитывать накопленные знания о мире, которые в явном виде не присутствуют в конкретном высказывании, но непосредственно влияют на его смысл.

***

1. Мельчук И.А. Опыт теории лингвистических моделей «Смысл-Текст». М.: Языки русской культуры, 1999 – 346 с.
2. Лахути Д.Г., Рубашкин В.Ш. Семантический (концептуальный) словарь для информационных технологий // Научно-техническая информация. 2000. – № 7. – С. 1–9.
3. Падучева Е.В. Динамические модели в семантике лексики. М.: Языки славянской культуры, 2004. – 608 с.
4. Кучуганов В.Н. Семантикаграфической информации / В.Н. Кучуганов // Известия ТРТУ. Тематич. вып. "Интеллектуальные САПР. Материалы междунар. научн.-техн. конф. "Интеллектуальные САПР". – Таганрог: Изд-во ТРТУ. 2002. – № 3(26). – С. 157–166.
5. Кучуганов А. В. Распознавание изображений и семантический анализ текстов в задачах патентно-лицензионного поиска / А. В. Кучуганов, М. Н. Мокроусов // Интеллектуальные системы в производстве. 2010. Ижевск. Изд-во ИжГТУ. 2010. – № 1 – С. 292–299.
6. Автоматическая Обработка Текста (АОТ) [Электронный ресурс]. [2003] – Режим доступа: <http://www.aot.ru/> – (Дата обращения: 25.11.2017).
7. Microsystems, Ltd [Электронный ресурс]. – Режим доступа: <http://www.analyst.ru/> – (Дата обращения: 25.11.2017).
8. Компьютер пресс [Электронный ресурс]. [1999] – Режим доступа: <http://compress.ru/Article.aspx?id=19605#10> – (Дата обращения: 28.11.2017)
9. Электронный научный журнал: программные продукты, системы и алгоритмы [Электронный ресурс]. [2012] – Режим доступа: <http://swsys-web.ru/methods-and-systems-of-semantic-text-analysis.html> – (Дата обращения: 29.11.2017)
10. Protégé [Сайт]. – Режим доступа: <http://protege.stanford.edu/> – (Дата обращения: 29.11.2017).