ТЕХНОЛОГИИ ПОСТРОЕНИЯ ОНТОЛОГИЙ ИЗ ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

Солдатов Александр Иванович

студент, Сибирский государственный университет путей сообщения, РФ, г. Новосибирск

TECHNOLOGIES FOR CONSTRUCTING ONTOLOGIES FROM NATURAL LANGUAGE TEXTS

Alexander Soldatov

Student, Siberian State University of Railway Transport, Russia, Novosibirsk

Аннотация. В данном проекте рассмотрены основные построения онтологий на основе естественных языков. Показано, что онтологии пусть и вошли в нашу жизнь и недавно, но существенно облегчили нашу жизнь. На основе исследования выделены дальнейшие перспективы развития.

Abstract. In this project, the main constructions of ontologies based on natural languages are considered. It is shown that ontologies, although they have entered our lives recently, have made our lives much easier. On the basis of the study, further development prospects are identified.

Ключевые слова: филологические науки, онтология, художественная литература, языкознание, прикладное языкознание, лингвистические вопросы информационных и машинных языков, лингвистика, методы построения онтологий

Keywords: philological sciences, ontology, fiction, linguistics, applied linguistics, linguistic issues of information and machine languages, linguistics, methods of constructing ontologies

В современном развитие общества, все острее стоит необходимость обработки гигантских объемов знаний, информационных ресурсов, которые отличаются слабой структуризацией, плохой систематизацией, а также находятся на различных источниках в процессе накопления различных отраслей человеческой деятельности. Для решения данных проблем необходимо использовать онтологии. Онтологией является совокупность описаний предметной области, включающей в себя термины, взаимосвязи объектов, процессы, а также их значения.

Онтология обеспечивает словари для представления и обмена знаниями о некоторой предметной области и множество связей, установленных между терминами в этих словарях.

Онтология предоставляет словари для представления и обмена накопленными знаниями в определенной предметной области, а также набор отношений, установленных между терминами в этих словарях.

Онтологический анализ основан на описании системы в терминах сущностей, взаимосвязи друг с другом и преобразовании сущностей, которое выполняется в процессе решения

конкретной проблемы. Онтологическая инженерия подразумевает глубокий структурный анализ дисциплинарной области. Основным преимуществом онтологического инжиниринга является целостный подход к автоматизации предприятия.

Наиболее распространенные онтологические сервисы предназначены для тестирования следующего:

1) согласованность. Онтология является согласованной, если модель выполняет все аксиомы онтологии;

2) непротиворечивость. Онтология O является непротиворечивой, если любая модель онтологии удовлетворяет аксиоматике O;

3) реализуемость. Концепция C реализуема в онтологии O, если она не интерпретируется как пустое множество для некоторой модели онтологии O;

4) категоризация. Концепция D включает концепцию C для онтологии O, если C интерпретируется как подмножество D в каждой модели онтологии O;

5) классификация. Для онтологии O должна быть задана иерархия понятий на основе отношения подчиненности

На современном этапе развиваются, в основном, ниже приведенные типы онтологий

- Предметно-ориентированные (Domain-oriented)

- Ориентированные на прикладную задачу (Task-oriented)

- Базовая техническая онтология. (Basic technical ontology)

- Общие онтологии (Generic ontologies)

Основной задачей является возможность указывать дополнительную машинно-интерпретируемую семантику ресурсов, сделать машинное представление данных более соответствующим положению вещей в действующем мире, значительно увеличить выразительные возможности концептуального моделирования слабоструктурированных Web-данных.

Цель исследования в данной работе заключаются в том, чтобы разработать систему автоматического построения онтологий из текста.

В рамках данной работы ставятся следующие задачи:

- Изучение существующих подходов к построению онтологий

- Разработка формального языка описания онтологии, максимально похожего на естественный русский язык

- Программная реализация основных семантических отношений

1. Аналитический обзор лингвистических технологий

1.1 Аналитический обзор методов обработки информации на естественном языке

Текст является основной формой обмена информационной нагрузки, а также представляет существенную часть ресурсов информационных систем.
С целью обработки данных в большом объеме не обойтись без онтологий, для обеспечения построения комбинаций фактов, для дальнейшего получения выводов.

Обработка естественного языка — область, взаимодействия на стыке различных

компьютерных технологий, искусственного интеллекта и лингвистики. Главная цель состоит в обработке и “понимании” естественного языка для перевода текста и ответа на вопросы.

Человеческий язык — специально сконструированная система передачи смысла изложенного устно или символьно.

Больший пласт технологий выполняется благодаря глубинному обучению , представляющему собой, применение многослойных сетей для принятия окончательных решений на основе неполной или неточной информации.

Внимание при работе с символьной информацией диффундирования от разработки подмножества признаков и поиска внешних баз знаний к нахождению источников данных и разметке текстов для дальнейшего обучения нейронной сети, для которого требуется существенно значимо больше данных по относительно стандартных методов. Поэтому появляется необходимость использования относительно больших объемов данных и из-за незначительной интерпретируемости и нестабильности нейронные сети в настоящий момент не востребованы в применяемых приложениях промышленного масштаба, в отличие от уже имеющихся и достойно зарекомендовавших себя алгоритмов обучения, таких как случайный лес и машины опорных векторов.

Векторное представление — систематизированная методика интерпретации строк, в векторном виде, имеющем значения. Таким образом выполняется построение плотного вектора (dense vector) применительно к каждому слову так, чтобы встречающиеся в приближенно одинаковых смысловых нагрузках слова имели идентичные вектора.

Векторное представление принято считать начальной точкой для большинства NLP задач и определяет глубокое обучение результативным на маленьких датасетах.

В отличие от традиционных представлений слов, здесь применяется нейровероятностная модель языка (отсюда и связь с глубоким обучением), где каждое слово представляется в векторном виде из вещественных чисел в маленьком пространстве. Изначально векторам присваиваются случайные значения. Далее в процессе обучения для слова подбирается вектор, максимально похожий на векторы других слов, которые встречаются в похожих контекстах. В качестве контекста берется небольшое окно предшествующих и последующих слов. Этот достаточно простой подход дает интересные результаты.

Разработанная в 2013 году группой исследователей совокупности моделей на базе искусственных нейронных сетей Word2vec, которые были предназначены для получения векторных представлений слов на естественном языке. Эти вектора позволяют результативней передать семантическую близость слов.

Word2vec принимает большой объем текстовой информации в качестве входных данных, который будет сформирован в векторном виде. Изначально происходит генерация словаря корпуса. На следующем этапе алгоритм переходит по каждой позиции t в тексте, которая является центральным словом с и контекстное слово о. На следующем этапе применяется оценка похожести векторов слов для с и о, с целью вычисления вероятности о при заданном с (или наоборот), и продолжается регулировка вектор слов для максимизации этой вероятности.

Word2vec представлен в 2 модельных вариациях:

1. Skip-Gram: рассматривается контекстное окно, содержащее k последовательных слов. На следующем этапе пропускается одно слово и обучается нейронная сеть, содержащая все слова, кроме пропускаемого, которое алгоритм пытается предсказать. Из этого следует, если 2 слова периодически делят схожий контекст в корпусе, эти слова будут иметь близкие векторы.

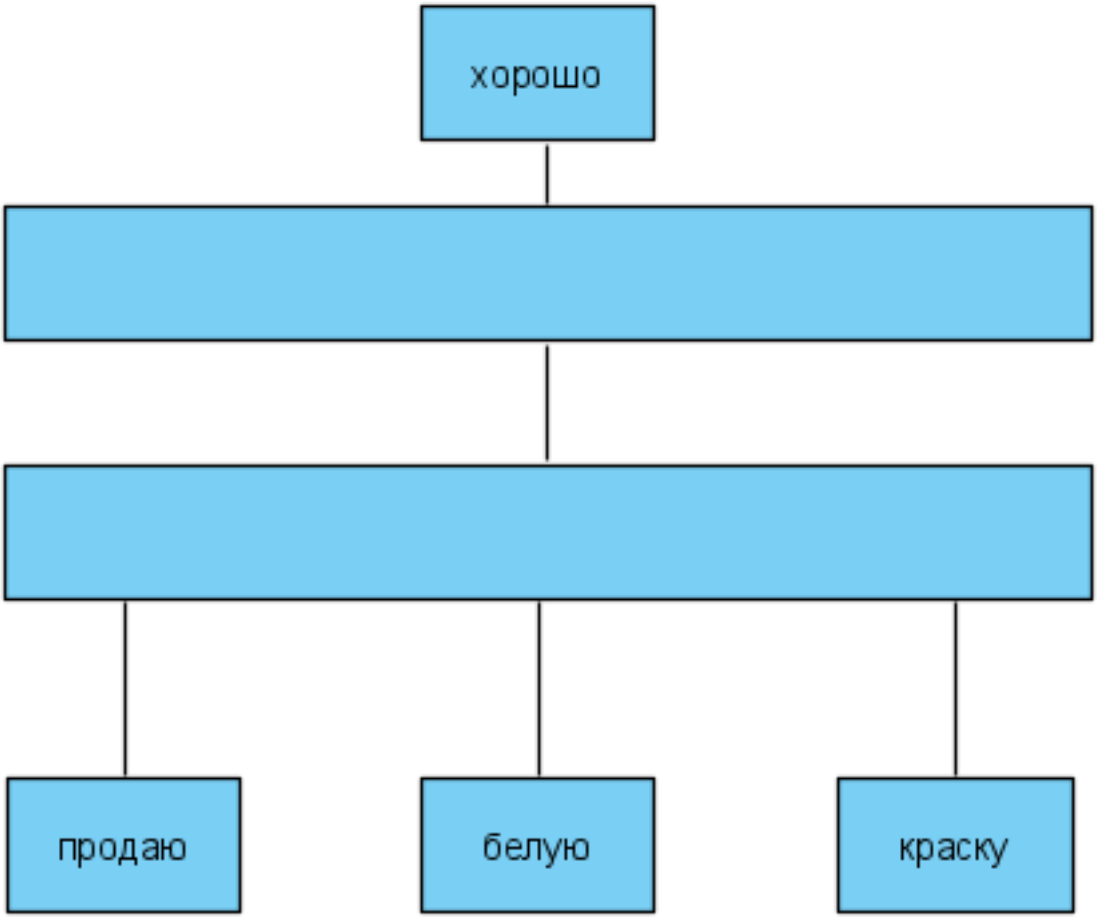

Рисунок 1. Skip-Gram модель

2. Continuous Bag of Words: берется много предложений в корпусе. Каждый раз, когда алгоритм видим слово, берется близлежащие слово. На следующем этапе на вход нейросети направляются контекстные слова и предсказыватся слово в центре этого контекста. В случае тысяч таких контекстных слов и центрального слова, получаем один экземпляр датасета для рассматриваемой нейросети. Нейросеть обучается и ,наконец, выход закодированного скрытого слоя выдает вложение (embedding) для определенного слова. То же происходит, если нейросеть обучается на большом числе предложений и словам в схожем контексте приписываются схожие вектора.

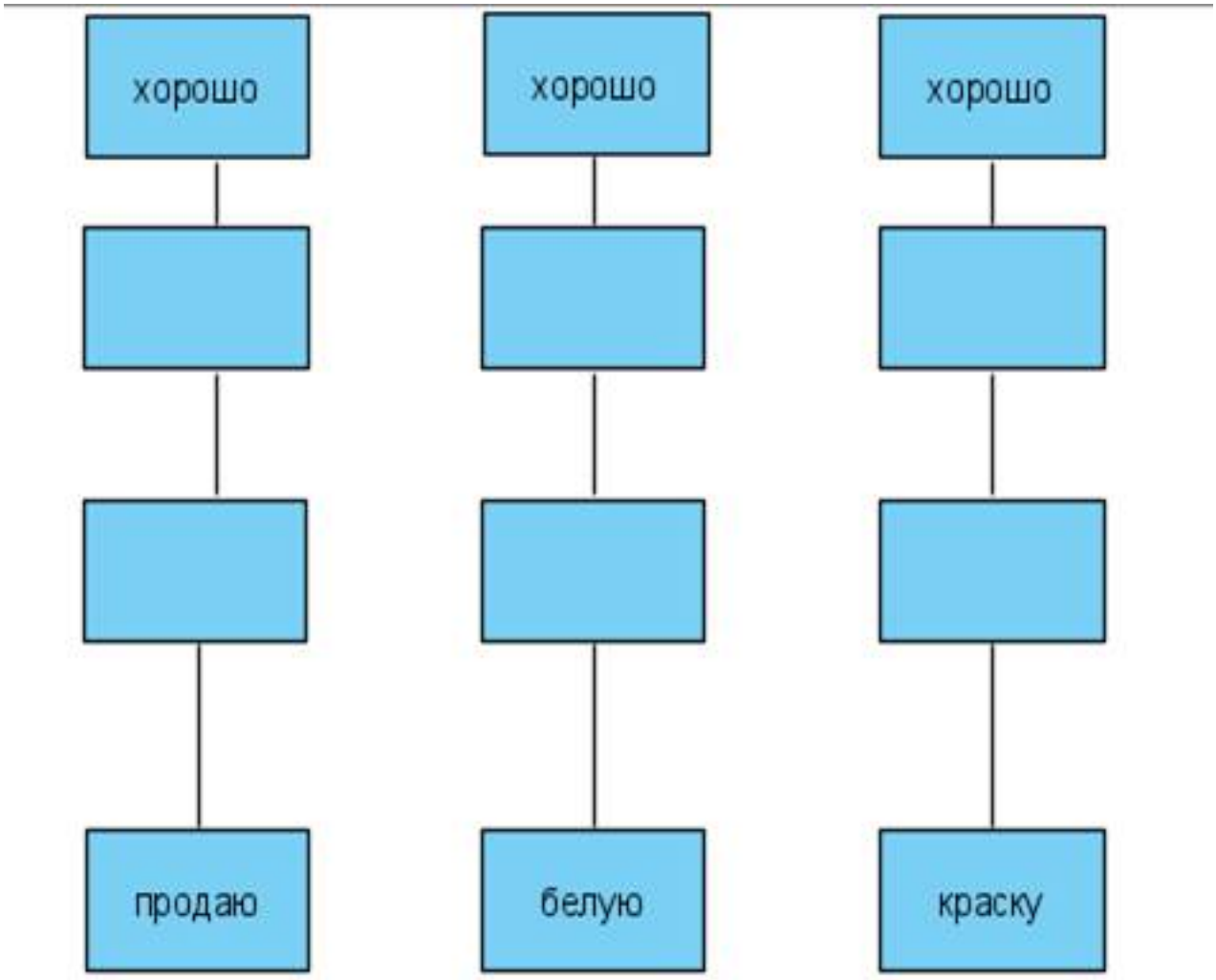

Рисунок 2. Continuous Bag of Words модель

Сегодня подобные векторы широко используются как для автоматического построения и расширения семантических ресурсов, так и в качестве числовых признаков в системах машинного обучения при решении задач системной классификации, кластеризации и проч.

1.2 Аналитический обзор методов обработки информации на естественном языке на основе машинного обучения

Машинное обучение – это класс методов автоматического создания прогнозных моделей на основе данных. Алгоритмы машинного обучения превращают набор данных в модель.

Типы задач машинного обучения:

1)задача регрессии – выводы на основе выборки объектов с отличительными характеристиками.

2) задача классификации – получение категориального ответа на основе набора характеристик.

3) задача кластеризации – распределение данных на кластеры.

4) задача уменьшения размерности – сведение большого числа характеристик к меньшему для удобства их представления в вивузальном виде по окончанию.

5) задача выявления аномалий – отделение нестандартных случаев от обычно встречающихся случаев.

Основные методы машинного обучения:

Обучение с учителем:

Данный способ позволяет изменять данные до тех пор, пока не получится намечанный результат. В последствие применяя полученные данные для будующих прогнозов, при использовании новых данных. Применяется для задач систематизации и предсказания. Например определение риска вложения средств в компанию.

Обучение без учителя:

Машина изучает набор данных и находит сокрытые закономерности и корреляции между различными переменными. Используется для группировки данных в кластеры. метод кластеризации, применяемый для вероятностного соединения записей. Ориентируются связи меж веществами данных, и на основании данных отношений обнаруживаются связи меж людьми и организациями в физиологическом или же виртуальном мире.

Например для анализа клиентской базы компании.

Обучение с частичным привлечением учителя:

Это гибрид обучения с учителем и без него.

Разметив малую часть данных, учитель дает понять машине , каким образом кластеризовать остальное.

Обучение с подкреплением :

При обучении с подкреплением машине позволяют вести взаимодействие с данными и «вознаграждают», когда она правильно выполняет задание. Автоматизировав подсчет вознаграждений, можно дать возможность машине обучаться самостоятельно. Одно из применений обучения с подкреплением — сортировка товаров в розничных магазинах.

Глубинное обучение:

Глубинное обучение может проходить как без учителя, так и с подкреплением. При глубинном обучении отчасти моделируется основы изучения людей — используются нейронные сети для все больше досконального уточнения данных комплекта данных.

1.3 Аналитический обзор методов обработки информации на естественном языке на основе вопросно-ответных систем

Аналитический обзор методов обработки информации на естественном языке на основе вопросно-ответных систем

Вопросно-ответные системы- это вид информационно-поисковых систем, способных обрабатывать введенный пользователем вопрос на естественном языке и выдавать осмысленный ответ. Результатом является ответ сформированный системой в результате анализа данных.

Классификация вопросно-ответных систем:

Метапоисковая система. В качестве источника данных такая система использует классическую поисковую систему, то есть использует неструктурированные данные, которые делятся на две группы: традиционные неструктурированные документальные и неструктурированные семантические.

Система анализирует вопрос пользователя на естественном языке с целью выделить следующее:

–предположение о семантическом классе ответа;

–фокус вопроса (вопросительные слова: кто, где, в каком, когда, сколько и др.);

–опора вопроса

–остальные члены вопросительного предложения, которые описывают уникальные свойства искомого объекта

Поиск по аннотированному тексту. Эти системы имеют в собственном составе поисковый индекс документов в различие от метапоисковых.

Функционируют такие системы также с неструктурированными данными. Элементами индекса являются не отдельные слова текста, а объекты детализированного лингвистического анализа:

- именованные сути

- простые синтаксические связки

- предикативно-аргументные структуры предложения.

Экспертная система. В начале 1970-х годов начал активно развиваться подход разделения системы работы с правилами - системы вывода и системы хранения от самих правил. Информация теперь хранится не в виде данных, а в виде знаний - набора правил и простых фактов. А система вывода при поддержке объединения знаний из различных правил может образоваться новая информация, не хранящуюся в базе знаний системы непосредственно. Основными компонентами экспертной системы являются: база фактов, база правил, база автоматически сгенерированных знаний и машина вывода.

База фактов—это структурированная БД, которая может быть автоматизирована в результате анализа коллекции документов. Этот процесс схож с построением аннотированного индекса.

Поиск в коллекции вопросов и ответов.В социальных системах вопросно-ответного поиска одни пользователи отвечают на вопросы других. Пользователь открывает страницу Web-сайта и формулирует вопрос. Система ищет похожие запросы в коллекции вопросов и ответов и выдает необходимы раздел с нужным вопросом. В случае если аналогичный вопрос не существует, создается новый раздел для обсуждения вопроса. На этот вопрос отвечают желающие, а автору приходят уведомления по мере появления ответов. Данные в такой системе представлены в виде коллекции вопросов с ответами, которая имеет возможность пополняться другими пользователями или даже автоматически.

Тенденция развития вопросно-ответных систем. Одним из первых подходов к вопросно-ответным системам можно назвать систему BASEBALL начала 60-х годов прошлого века. Отличительной особенностью, позволяющей считать ее первой вопросно-ответной, являлась возможность задавать вопросы к системе на естественном языке, но базой знаний служила обычная структурированная база данных. Таким образом, можно считать ее системой естественного ввода.

1.4 Аналитический обзор методов обработки информации на естественном языке на основе формальной грамматики

Формальная грамматика(ФГ) является способом описания формального языка, путем выделения групп из конечного алфавита. ФГ необходима для представления синтаксиса в трансляторах и автоматизации синтаксического анализа. Данный способ служит связующим звеном вмежду синтаксисом языка и программной средой.

На основе формальной грамматики создается инструментарий синтаксического анализа.

Рисунок 3. Схема синтаксического анализатора

Грамматики по ограничениям, накладываемым на правила, образуют несколько классов:

- класс 0 – грамматики общего вида (или грамматики с фразовой структурой), на которые не накладывается никаких ограничений;
- класс 1 - контекстно-зависимые грамматики, Термин контекстно-зависимая характеризует частный случай правил в такой грамматике, имеющих вид xAy ::=xby, когда замена нетерминала A на цепочку b возможна на только в окружении некоторых символов, то есть в контексте.
- класс 2 – контекстно-свободные грамматики, имеющие в левой части любого правила единственный нетерминал.

Список литературы:

1. Bob Violino. Machine learning: When to use each method and technique. InfoWorld. SEP 6, 2018

2. Tomas Mikolov et. al. Efficient Estimation of Word Representations in Vector Space, SEP 7, 2016

3. Kai Sheng Tai et. al. Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks // Association for Computational Linguistics (ACL), 2015.