ОНЛАЙНОВЫЕ СЕМАНТИЧЕСКИЕ ВЫЧИСЛЕНИЯ НА ПЛАТФОРМЕ RUSVECTŌRĒS В ПРЕПОДАВАНИИ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ

Концевой М.Р.

Брестский государственный университет имени А.С. Пушкина, г. Брест, Беларусь, kmp@brsu.by

Abstract. The didactic potential of semantic computations in the formation of students' linguistic and computational competences is analyzed. The use of semantic calculators in the construction of word context vectors is considered.

Преподавание компьютерной лингвистики предполагает формирование у учащихся как лингвистических, так и вычислительных компетенций. Вычислительные компетенции могут быть поняты как способность эффективного применения вычислений в решении любых возникающих проблем, что предполагает сформированность важнейших математических компетенций на высоком уровне абстракции. В дистрибутивной семантике вычисление понимается наиболее абстрактно, как математическое преобразование входящий потока данных в выходной, с отличной от первого структурой. С точки зрения теории информации, вычисление – это процесс получения из входных данных нового знания.

Семантические вычисления (Semantic computing) – направление информатики, реализующее программы формального анализа и обработки текстовых данных на основе вычисления их семантической близости. Вычислением степени семантической близости между лингвистическими единицами на основании их распределения (дистрибуции) в массивах лингвистических данных занимается дистрибутивная семантика. Практические приложения имеет распределение лингвистических единиц в больших массивах данных (корпусах). Семантические вычисления обеспечивают технологии обработки больших данных. Ядром таких вычислений является использование естественного языка для представления и использования знаний, заданных онтологиями на основе булевой и предикатной (модельной) семантики дескриптивной логики. Дистрибутивная семантика наглядно демонстрирует, что вычисления не ограничиваются только числовыми приложениями, но могут быть использованы в работе с любыми конструкциями, в том числе, языковыми. На семантических вычислениях основана важнейшая для современного машинного обучения нейронных сетей концепция вложений (embeddings). Таким образом, семантические вычисления лежат в основе основных нейросетевых сервисов автоматической обработки текста (перевода, распознавания и синтеза речи, диалоговых систем, автореферирования, компьютерной корректуры и др.).

В семантических вычислениях каждой языковой единице (слову, терму, токену, n-грамме) присваивается свой контекстный вектор. Множество таких векторов формирует словесное векторное пространство. Семантическое расстояние между понятиями, выраженными словами естественного языка, вычисляется, как правило, как косинусное расстояние между векторами словесного пространства. Таким образом, в семантических вычислениях на новый уровень абстракции возводится и определение вектора, который понимается более обобщенно, как произвольный математический объект, характеризующийся величиной и направлением в специальном конфигурационном пространстве.

Важнейшим инструментом для современных семантический вычислений является Word2Wec [1]. Большинство современных приложений автоматической обработки языка и речи основываются на алгоритмах Word2vec. В качестве входных данных word2vec принимает текст и сопоставляет каждому слову вектор, выдавая координаты слов на выходе. Сначала он генерирует словарь корпуса, а затем вычисляет векторное представление слов. Векторное представление основывается на контекстной близости. Слова, встречающиеся в тексте рядом, будут иметь векторы с высоким косинусным сходством (cosine similarity).

Для образовательных целей в контексте преподавания компьютерной лингвистики удобно использовать сервис RusVectōrēs, который вычисляет семантические отношения между словами русского языка и предоставляет доступ к предобученным дистрибутивно-семантическим моделям (word embeddings) [2]. RusVectōrēs фактически является «семантическим калькулятором» с уже подготовленными моделями, с помощью которых пользователи могут вычислять семантические сходства между парами слов; находить слова, ближайшие к данному (с возможностью фильтрации по части речи и частотности); решать аналогии вида «найти слово X, которое так относится к слову Y, как слово A относится к слову B»; выполнять над векторами слов алгебраические операции (сложение, вычитание, поиск центра лексического кластера и расстояний до этого центра). RusVectōrēs также позволяет рисовать семантические карты отношений между словами; получать, в виде массива чисел, вектор и его визуальное представление для выбранного слова; генерировать контекстно-зависимые лексические подстановки для контекстуализированных дистрибутивных моделей. Знакомство учащихся с семантическими вычислениями и их практическое использование может стать значимым фактором формирования лингвистических и вычислительных компетенций в контексте преподавания компьютерной лингвистики.

Литература

1. Word2vec [Electronic resource] – Mode of access: https://code.google.com/archive/p/word2vec/ – Date of access: 02.05.2022.

2. RusVectōrēs [Электронный ресурс]. – Режим доступа: https://rusvectores.org/ru/ – Дата доступа: 02.05.2022.