Е. И. Ильина

ВЕКТОРНОЕ ПРЕДСТАВЛЕНИЕ СЛОВ

Для многих автоматических операций с текстом компьютер должен «понимать» значение слова. Но как компьютер узнает, что слово *педагог* ближе к *наставнику*, чем к *инженеру*.

На сегодняшний день данная задача легко решается с помощью семантических моделей для естественных языков, в которых используется такое понятие, как *векторное представление слов*. Векторные представления слов позволяют хранить семантику слов в понятном для компьютера виде. Напомним, что семантическая модель данных (SDM) – это высокоуровневый, основанный на семантике, формализм описания баз данных и структурирования для баз данных.

Семантика – раздел лингвистики, изучающий смысловое значение единиц языка. В качестве инструмента изучения применяют семантический анализ. Другими словами, семантика – искусство выражения сложных смыслов через более простые. Например *идти* и *бежать*. И то, и другое обозначает ‘двигаться, попеременно передвигая ноги’, но в одном случае добавляется смысл ‘быстро’. То есть, семантика – искусство создания объяснений (дефиниций, как это называется в словарях).

Но бывает все гораздо сложнее. Например, что такое *проливной*? Это слово обозначает ‘очень сильный’, но сказать *проливной человек* будет неправильно (хотя человек на самом деле является очень сильным), потому что проливным бывает только дождь. Т.е. слово в определенных контекстах может иметь не собственное значение, а выражать некоторое общее значение в зависимости от того, к какому другому слову оно прилагается. Например *проливной дождь* это то же самое, что *жгучий брюнет*. Мы говорим о признаке, и о нем надо сказать в высокой степени. Получается, что в зависимости от ключевого слова функция слова становится совершенно другая (по аналогии с математической функцией). Значит, один из способов заниматься компьютерной семантикой – описывать лексические функции, т.е. выражать смысл слов математически.

Известный британский лингвист Джон Руперт Ферс всегда говорил о том, что значение слова определяется его контекстом. Он неоднократно писал: «Вы узнаете слово по окружающим его словам». Эта идея, которую в лингвистике называют дистрибутивной гипотезой, была взята на вооружение, и позднее была разработана теоретическая база для векторных представлений слов.

Одним из ключевых моментов в этой теории является понятие контекстного окна, т.к. контекстом слова принято называть слова, находящиеся на одинаковом расстоянии слева и справа от него. Традиционно используется окно размером 4, т. е берутся четыре слова слева от целевого слова (target) и четыре слова справа.

Самым простым видом векторного представления слова является частота, с которой данное слово встречается рядом с другим словом в тексте. Эти частотности принято хранить в матрице «слово/слово». Предварительно при этом исследуемый текст подвергается процессу лемматизации, т. е. процессу приведения словоформы к ее нормальной (словарной) форме.

Возьмем, например, скороговорку: *Ехал грека через реку, видит грека – в реке рак. Сунул грека руку в реку, рак за руку греку – цап!*. Составим для нее матрицу совместной встречаемости слов (табл. 1):

Таблица 1

Матрица совместной встречаемости слов

	в	видеть	грек	ехать	за	рак	река	рука	сунуть	цап	через
в	0	1	2	0	1	2	3	2	1	0	1
видеть	1	0	2	1	0	1	2	0	0	0	1
грек	2	2	1	1	1	3	5	2	1	1	2
ехать	0	1	1	0	0	0	1	0	0	0	1
за	1	0	1	0	0	1	1	2	0	1	0
рак	2	1	3	0	1	0	2	2	0	1	0
река	3	2	5	1	1	2	1	2	1	0	1
рука	2	0	2	0	2	2	2	0	1	1	0
сунуть	1	0	1	0	0	0	1	1	0	0	0
цап	0	0	1	0	1	1	0	1	0	0	0
через	1	1	2	1	0	0	1	0	0	0	0

Рассмотрим строку, содержащую слово *рак*. Совокупность чисел, находящихся в этой строке, и будет является векторным представлением искомого слова *рак* (табл. 2).

Таблица 2

Векторное представления слова *рак*

рак	2	1	3	0	1	0	2	2	0	1	0

Но существуют и определенные недостатки этого метода т.к. он не дает нам понимания некоторых важных закономерностей, которые происходят в нашей повседневной речи. Мы очень много вещей пропускаем. Например: *Звонила Таня. У нее заболела няня*. Тут важно понять, что второе предложение это содержание того, что Таня сказала. Это то значение, слова *звонить*, которое мы можем узнать, если укажем, что же называется словом *звонить*.