Секция «Вычислительная математика и кибернетика»

Поиск семантически близких терминов при оценки близости текстов

Карпов Илья Андреевич
Аспирант
Московский государственный открытый университет, Факультет информатики и радиоэлектроники, Москва, Россия
E-mail: karilan@yandex.ru

Большинство систем автоматической обработки текстов построены на векторной модели документа (VSM). Для оценки близости двух текстов производится сравнение векторов, состоящих из термов, входящих в эти тексты. Один из недостатков модели - отсутствие информации о синонимии (одно значение выражается множеством термов) и лексической многозначности (один терм имеет множество значений) при сравнении двух векторов. [2]

Задача снятия лексической многозначности, как правило, решается в два этапа:

1. для каждого слова, относящегося к тексту, определить, какие оно может иметь значения;

2. на основании контекста, в котором встретилось слово, выбрать наиболее подходящее значение.

Таким образом, у терма появляется дополнительная информация о контексте, отличающая его от аналогичного по написанию терма. Дальнейшее сравнение производится с учетом контекста, в котором существует терм.

Для решения задачи разработано большое число методов, дающих высокие результаты (F-мера определения значения 54-86% в зависимости от коллекции), например на основе сетей документов или алгоритм Леска. [1]

В данной работе предлагается использовать описанный выше подход для поиска семантически близких термов при сравнении текстов. Для выдвижения гипотезы о синонимии выполняется поиск среди устойчивых n-грамм слов. Тогда, если n-1 термов словосочетания одинаковы, но при этом контекст n-граммы остается неизменным, то можно сделать вывод о семантической близости двух слов в данном контексте. При этом выполняется допущение "One sence per collocation"о том, что одной n-грамме соответствует только 1 смысл, что справедливо для 90-99% биграмм. [3].

При сравнении n-грамм предлагается использовать соответствующие термам словарные статьи Wikipedia. При этом необходимо учитывать степень гранулярности значений терма: использование тонко различающихся значений оправдано только при обработке узко-специализированных текстов.

Литература

1. Турдаков Д. Ю. "Методы и программные средства разрешения лексической многозначности терминов на основе сетей документов"// Автореф. дисс. канд. ф-м. наук. Москва, 2010.

2. Tsatsaronis George, Panagiotopoulou Vicky. A generalized vector space model for text retrieval based on semantic relatedness // EACL '09 Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop

3. Yarowsky David. One sense per collocation // HLT '93 Proceedings of the workshop on Human Language Technology