Ф

ОПРЕДЕЛЕНИЕ БИГРАММ НА МАТЕРИАЛЕ НАУЧНЫХ ТЕКСТОВ ПО ИЗВЛЕЧЕНИЮ ДАННЫХ ИЗ ТЕКСТОВ

Сарры Николай Антонович[1]

В статье рассматривается извлечение информации о предметной области научных текстов, что является неотъемлемой частью задачи выделения важных терминов. В качестве предметной области была выбрана область, связанная с извлечением данных из текстов, большинство терминов которой являются не однословными. Не однословные термины характеризуются термином коллокация.

Коллокация – неслучайное сочетание двух и более лексических единиц, характерное как для языка в целом, так и для определенного типа текстов. Использование статистических мер позволяет выделять из текста коллокации и ранжировать их по степени устойчивости в соответствии со значениями выбираемых мер [1].

Множество выделяемых терминологических коллокаций в существенной степени характеризуют предметную область рассматриваемой коллекции. Чрезвычайно актуальным статистический метод становится в случае становления новой предметной области, изменения терминологии (особенно при сосуществовании разных научных парадигм, каждая из которых использовать свой терминологический аппарат). Для того, чтобы установить состав терминологических единиц, могут применяться статистические меры оценки не случайности совместной встречаемости единиц.

Для текстов научного стиля статистически определяются составные слова и устойчивые конструкции, характеризующие особенности стиля, смысловую и коммуникативную структуру текста.

В основу статьи следующие гипотезы [2]:

1. Использование меры *MI* позволяет выделить ключевые не однословные термины, которые характеризуют предметную область.

2. Использование меры *t-score* позволяет выделить устойчивые сочетания, устойчивые конструкции, характеризующиеся стилистическими особенностями научного текста.

Статистические мера MI – Mutual Information (коэффициент взаимной информации) [2] определяется по формуле (1):

$$MI = \log_2 \frac{f(n,c) \times N}{f(n) \times f(c)}, \qquad (1)$$

где n – ключевое слово; c – коллокат; $f(n,c)$ – абсолютная частота встречаемости ключевого слова n в паре с коллокатом c; $f(n)$, $f(c)$ – абсолютные частоты ключевого слова n и слова c в корпусе; N – объем корпуса (количество словоупотреблений) [2].

Мера t-score [2] определяется по формуле (2):

$$t - score = \frac{f(n,c) - \frac{f(n) \times f(c)}{N}}{\sqrt{f(n,c)}}. \qquad (2)$$

С точки зрения теории вероятности, мера MI является способом проверить независимость появления двух слов в тексте – если слова полностью независимы, то

[1] студент Института информатики и искусственного интеллекта Донецкого национального технического университета. Тел.: 0951293247

вероятность их совместного появления равна произведению вероятностей появления каждого из них.

Мера t-score используется гораздо реже, чем мера MI, поскольку она является лишь несколько модифицированным ранжированием коллокаций по частоте. Очевидно, что значение данной меры тем выше, чем выше частота коллокации в наборе текстов. Данная мера содержит коррекционный компонент, но эта поправка отражается лишь на самых частотных словах.

Был подобран набор текстов в области извлечения данных. На основании обработки этих текстов была получена предварительная информация о терминах, употребляемых в текстах, посвященных извлечению данных.

В табл. 1 представлен список биграмм, полученных с помощью меры MI.

Этого списка достаточно, чтобы получить предварительную информацию о наиболее важных не однословных терминах: объектах исследования, материале, методах, результатах.

Несмотря на то, что обычно мера t-score считается малопригодной для поиска терминологических словосочетаний, она оказывается полезна при решении задачи о выделении тех единиц, которые характеризуют все (или подавляющее большинство) текстов коллекции. Используя минимальный морфологический фильтр из списков t-score-коллокаций, можно выделить те сочетания, которые могут рассматриваться как терминологические.

Используя меру *t-score* можно выделить те сочетания, которые могут рассматриваться как терминологические. Таким образом, был получен список биграмм общий для всех текстов из набора (табл. 2).

Таблица 1. Биграммы (MI-score), выделяющиеся и для лексем, и для словоформ

№	биграмма	
1	лексическая	единица
2	математическая	лингвистика
3	семантический	анализатор
4	морфологическая	разметка
5	научная	статья
6	предметная	область
7	анализ	текста
8	выделение	сущностей
9	автоматическое	извлечение
10	извлечение	информации
11	профессиональный	словарь
12	целевой	фрейм
13	фильтрация	документа
14	обучающая	выборка
15	шаблоны	фраз

Данное исследование показывает, что:

– использование меры MI позволяет выделить «ключевые» не однословные термины, характеризующие предметную область набора текстов;
– использование меры t-score позволяет выделить: «устойчивые сочетания», «устойчивые конструкции», характеризующие стилистические особенности научных текстов, коллокации, общие для всех текстов из набора.

Результаты исследования являются основой для разработки алгоритмов определения принадлежности текстов к научной тематике по извлечению данных.

Таблица 2. Терминологические биграммы (t-score), выделяющиеся и для лексем, и для словоформ.

№	Лексемные биграммы	
1	лексическая	единица
2	математическая	лингвистика
3	семантический	анализатор
4	выделение	сущностей
5	извлечение	информации
6	фильтрация	документа
7	модель	текста

Список источников

1. Ягунова Е.В. Формальные и неформальные критерии вычленения ключевых слов из научных и новостных текстов / Е.В. Ягунова. – М. – 2010. – С. 340–355.
2. Ягунова Е.В., Пивоварова Л.М. Извлечение и классификация коллокаций на материале научных текстов. Предварительные наблюдения / Е.В. Ягунова, Л.М. Пивоварова. – СПб. – 2010. – С. 356–364.