УДК 004.9

АЛГОРИТМ ФИЛЬТРАЦИИ И ОЦЕНКИ ВЕЛИЧИНЫ СЕМАНТИЧЕСКОГО ШУМА В ТЕМАТИЧЕСКИХ ОБРАЗОВАТЕЛЬНЫХ ТЕКСТАХ НА ПРИМЕРЕ АНГЛИЙСКОГО ЯЗЫКА

Барыкин Е.С.
Научный руководитель — к.т.н. Личаргин Д.В.
Сибирский федеральный университет

В статье предложен алгоритм фильтрации и оценки величины семантического шума в тематических образовательных текстах на примере английского языка на основе существующих методов анализа и представления многомерных данных. Приводится пример расчета с использованием предложенного алгоритма.

Ключевые слова: автоматическое реферирование, величина семантического шума, представление многомерных данных, анализ тематических текстов.

Одним из направлений автоматического анализа текстов на иностранном языке является автоматическое реферирование. Алгоритм работы многих существующих в этом направлении программных систем основывается на введение весовых коэффициентов для небольшого набора ключевых слов. В результате программа автоматически сохранит предложения с указанными словами и проигнорирует те, в которых ключевые слова не встретились. Предложения, отобранные в итоговый реферат, остаются без каких-либо изменений.

Основной идеей данной статьи является предложение алгоритма фильтрации семантического шума в тематических текстах на английском языке на основе методов анализа и представления многомерных данных для решения задачи автоматического реферирования и генерации осмысленных конструкций на естественном языке. Под семантическим шумом в данной статье следует понимать одну из его разновидностей – hesitation phenomenon (явление неопределенности/нерешительности).

Проблема создания алгоритма фильтрации семантического шума является актуальной и тесно связана с проблемой генерации осмысленных конструкций на естественном языке, которые определены потребностями лингвистического программного обеспечения.

Цель работы состоит в создании алгоритма фильтрации и оценки величины семантического шума в тематических образовательных текстах на английском языке. Для достижения этой цели были решены следующие задачи: анализ грамматической структуры предложений на английском языке, анализ методов представления многомерных данных, создание расчетной формулы для определения величины семантического шума.

Научная новизна состоит в предложении оригинального алгоритма фильтрации и оценки величины семантического шума в тематических текстах.

Представим множество тематических текстов на английском языке в виде многомерного пространства (Рис. 1.):

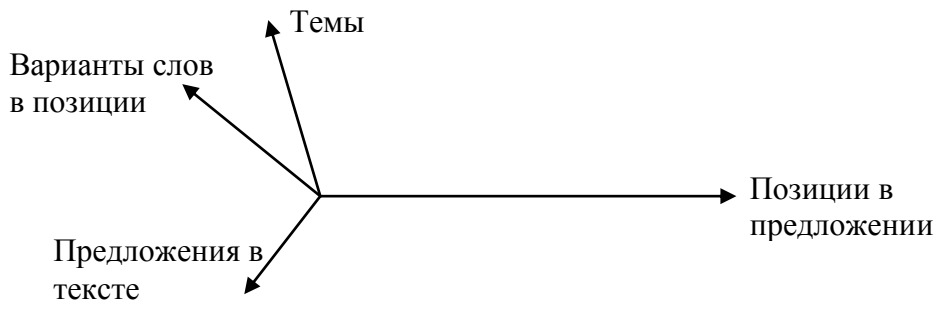

Рис. 1. Множество тематических текстов на английском языке

Алгоритм фильтрации и оценки величины семантического шума, предложенный в данной статье, применим для среза этого пространства – отдельно взятой темы.

В английском языке существует строгий порядок слов в предложении, который не зависит от его тематики. К примеру, после вопросительного слова (what, where, …) может следовать вспомогательный глагол (are, do, …) или артикли (a, the), но никогда не будет непосредственно следовать наречие степени (too, absolutely, …) или притяжательные местоимения (her, his…). Общая структура предложения насчитывает порядка 10^2 позиций в предложении с учетом именных групп:

<Вводное слово, Обстоятельство, Субъект <Определитель, Определение <Наречие степени, Группа прилагательного>, Именная часть>, Предикат <Модальность, Обстоятельство, Глагольная часть>, Объект <Определитель, Определение <Наречие степени, Группа прилагательного>, Именная часть>, Именная группа места <Связка, Определитель, Определение, Именная часть>, Именная группа времени <Связка, Определитель, Определение, Именная часть>, Именная группа цели <Связка, Определитель, Определение, Именная часть>, Именная группа инструмента <Связка, Определитель, Определение, Именная часть>, …, Обстоятельство>.

Как правило, большинство из этих позиций являются факультативными членами предложения и при составлении текстов многие из них пропущены, однако, знание общей структуры позволяет формировать структуру частного, упрощенного предложения.

Исходными данными для достижения поставленной цели является тематический образовательный текст на английском языке. Представим его в виде таблицы размером $m \times n$, где строкам будут соответствовать предложения (объекты) – $y(y_{1,\dots,} y_m)$, а столбцы членам предложения (признаки) – $x(x_1, \dots, x_n)$ в обобщенной структуре. Значение ячейки – $f(x_i, y_j)$.

Для оцифровки значений признаков используем дихотомическую шкалу:

$$f(x_i, y_j) = \begin{cases} -1, \text{ если } y_j \notin x_i \text{ (}j\text{-й признак отсутствует в } i\text{-м предложении);} \\ 0, \text{ если невозможно точно определить наличие / отсутствие признака;} \\ 1, \text{ если } y_j \in x_i. \end{cases}$$

Положим, что все признаки имеют семантические весовые коэффициенты:

$\omega(\omega_1, \omega_2, \dots, \omega_n)$, где n – количество признаков.

Основная цель использования этих коэффициентов – присвоить степень влияния конкретного признака на смысл предложения. Введение коэффициентов обусловлено тем, что в речи часто используются слова, которые незначительно влияют на общий смысл предложения (текста). Например:

These new red sport cars are usually very expensive. – Эти новые красные спортивные автомобили обычно очень дорогие.

Обозначим семантические весовые коэффициенты слов:

$\omega(these), \omega(new), \omega(red), \omega(sport), \omega(cars), \omega(are), \omega(usually), \omega(very), \omega(expensive).$

Очевидно, что очень дорогими обычно являются не только эти новые красные, а вообще спортивные автомобили любого цвета. Таким образом, семантические веса:

$\omega(these) = \omega(new) = \omega(red) \approx 0.$

Учитывая последнее утверждение, запишем исходное предложение:

Sport cars are usually very expensive. – Спортивные автомобили обычно очень дорогие.

В данном предложении наименьшим (почти нулевым) семантическом весом обладают определитель и определение субъекта.

Рассмотрим другой случай. Исходный текст посвящен кулинарии и способу приготовления некоторого блюда:

In fact you need to cut two sweet green peppers and add some olive oil. – В действительности вам необходимо порезать 2 сладких зеленых перца и добавить немного оливкового масла.

В этом случае важно сохранить смысл, что необходимо использовать именно 2 сладких зеленых перца, а не 3 красных острых перца. Вводное слово In fact и комплимент need to, напротив, никакой смысловой нагрузки не несут и их семантические веса равны:

$\omega(in\_fact) = \omega(need\_to) \approx 0.$

Смысл приведенных примеров в том, что в зависимости от тематики исходного текста, можно заранее предположить с некоторой вероятностью, какие из признаков будут более значимы для общего смысла текста, а какие менее значимы.

Рассмотрим упрощенный пример расчета (6 признаков):

1-й – обстоятельство времени и места;

2-й – определитель субъекта;

3-й – именная часть субъекта;

4-й – вспомогательный глагол;

5-й – наречие степени объекта;

6-й – определение объекта.

Всем признакам присваиваются семантические весовые коэффициенты, и вводится пороговое значение:

$\omega(0.25, 0.07, 0.25, 0.13, 0.2, 0.1), \xi = 0.1$

Признаки, для которых $\omega_j < \xi$ почти не влияют на смысл текста, считаются шумом и для составления реферата не учитываются. При этом должны выполняться условия:

$$\sum_{j=1}^{n} \omega_j = 1; \omega_j > 0.$$

Исходный текст:

This woman is absolutely lonely. Today this woman is very happy. These kids are nice.

Представим его в табличном виде:

	y_1	y_2	y_3	y_4	y_5	y_6
x_1	-1	0	0	1	0	1
x_2	1	1	1	0	0	1
x_3	0	1	-1	1	-1	0

Исходный текст в матричной форме:

$$A = \begin{pmatrix} -1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & -1 & 1 & -1 & 0 \end{pmatrix}.$$

Умножим матрицу A на вектор семантических коэффициентов:

$$B = \begin{pmatrix} -0.25 & 0 & 0 & 0.13 & 0 & 0.1 \\ 0.25 & 0.07 & 0.25 & 0 & 0 & 0.1 \\ 0 & 0.07 & -0.25 & 0.13 & -0.2 & 0 \end{pmatrix}.$$

Величина семантического шума i-го предложения находится в интервале:

$$S_i \in [0; \left(\frac{a_i}{b_i} + \frac{c_i}{b_i + c_i} \right) \cdot 100\%],$$

где a_i – количество признаков i-го предложения, для которых $f(x_i, y_j) = 1$ и $\omega_j < \xi$,
b_i – количество признаков i-го предложения, для которых $f(x_i, y_j) = 1$ и $\omega_j \geq \xi$,
c_i – количество признаков i-го предложения, для которых $f(x_i, y_j) = 0$ и $\omega_j < \xi$.

$S_1 \in [0;33\%]$, $S_2 \in [0;25\%]$, $S_2 \in [0;50\%]$ – интервалы величины семантического шума предложений.

Величина семантического шума текста находится в интервале:

$$S \in [0; \frac{1}{m} \sum_{i=1}^{m} \left(\frac{a_i}{b_i} + \frac{c_i}{b_i + c_i} \right) \cdot 100\%].$$

$S \in [0;36\%]$ – интервал величины семантического шума всего текста.

Исходный текст после фильтрации семантического шума:

Woman is absolutely lonely. Today woman is very happy. Kids are nice.

Предложенный алгоритм фильтрации и оценки величины семантического шума в тематических образовательных текстах на примере английского языка положен в основу программной системы, созданной автором в рамках выполнения магистерской диссертации. Полученные результаты предполагается использовать для дальнейших исследований, например, с целью трехмерной визуализации текстов в семантическом пространстве.