Оценки и визуализация уровня дискриминантной силы слов

Д.В. Ландэ, Институт проблем регистрации информации НАН Украины

Рассмотрены подходы к оценке дискриминантной силы слов. Подходы проверены на литературном произведении и массиве новостных сообщений. Предложен и реализован метод визуализации уровня неравномерности вхождения слов в тексты.

Ключевые слова для поиска в тексте, опорные слова для экстрагирования сниппетов или формирования автоматических рефератов, выбираются с учетом такого свойства слов, как «различительная» или дискриминантная сила. Ведь если слово относительно равномерно распределено по тексту, то оно вряд ли может использоваться для эффективного содержательного поиска или служить основой выбора чем-либо примечательного фрагмента, который может рассматриваться как некоторое сверфразовое единство [1].

Одна из первых технологий оценки качества ключевых слов была «материализована» Солтоном в векторно-пространственной модели поиска [2], в которой именно для учета дискриминантной силы слов было введено понятие инверсной частоты появления слова в отдельных документах массива. Предложенный метод взвешивания слов имеет сегодня стандартное обозначение – TF IDF, где TF указывает на частоту появления слов в документе, а IDF – на величину, обратную количеству документов в массиве, содержащих данное слово (чуть позднее, логарифм, монотонную функцию от этой величины):

$$w_i = tf_i \cdot \log \frac{N}{n_i},$$

где w_i – вес слова t_i, tf_i – частота слова t_i в документе, n_i – количество документов в информационном массиве, в которых используется слово t_i, слова N – общее количество документов в информационном массиве.

Оценка неравномерности вхождения слов возможна и на основе чисто статистических, дисперсионных оценок. В работе [3] предложена такая оценка дескриминантной силы слова:

$$\sigma_i = \frac{\sqrt{\langle d^2 \rangle - \langle d \rangle^2}}{\langle d \rangle},$$

где $\langle d \rangle$ – среднее значение последовательности d_1, d_2,..., d_n, n – количество появлений слова t_i в информационном массиве. Если обозначить координаты (номера) вхождения слова t_i в информационный массив как e_1, e_2,..., e_n, то $d_k = e_{k+1} - e_k \quad (e_0 = 0)$.

Для визуализации неравномерности вхождения слов в тексты в [3] была предложена технология спектограмм, которые внешне напоминают штрих-коды товаров [4], вместе с тем не позволяют рассматривать вхождения слов в разных масштабах измерений, как это делается, например, в вейвлет-анализе [5].

Автором были реализованы инструментальные средства позволяющие визуализировать плотность встречаемости слова в тексте в зависимости от ширины окна наблюдения. Через веб-интерфейс соответствующей программы (http://ling.infostream.ua/jag/jag.html) вводится текст и слово для анализа. В результирующей спектограмме по горизонтали откладываются номера вхождения слов в тексте, а по вертикали – ширина окна наблюдения. Одному вхождению слова соответствует светло-серый цвет. Если в соответствующее окно наблюдения попадает несколько целевых слов, то оно закрашивается более темным оттенком. Эксперт-лингвист по внешнему виду сразу может определить степень равномерности вхождения в текст анализируемого слова [6].

Рассчитанные автором коэффициенты неравномерности вхождения отдельных слов в повести Аркадия и Бориса Стругацких «Малыш» представлены в табл. 1, а соответствующие спектограммы – на рис. 1-5. При расчете коэффициента w_i использовался искусственный прием, исходный текст разбивался на фрагменты фиксированной длины по 500 слов, которые при расчетах рассматриваются как отдельные документы. Как видно, неравномерность вхождения отдельных слов, точно выражающаяся в коэффициентах w_i и σ_i, вполне может быть выявлена визуально в спектограммах. Однако монотонность возрастания значений w_i нарушается в одном случае (слово «Комов»), что объясняется разными подходами, применяемыми для расчета w_i и σ_i и аномально частым появлением данного слова.

Табл. 1. Значения коэффициентов неравномерности для отдельных слов в рассказе Стругацких «Малыш»

Слово	Вхождений	w_i (TF IDF)	σ_i
Видео	8	20,82	1,05
Комов	346	70,65	2,18
Семенов	16	32,96	2,19
Горбовский	33	63.01	2.76
Малыш	216	166,14	8,36

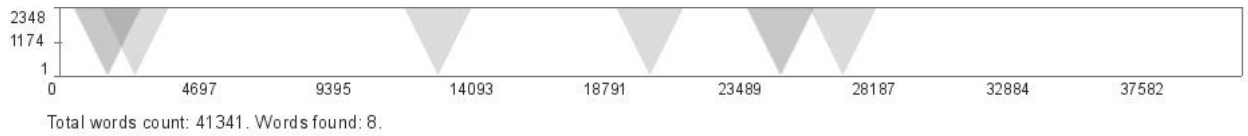

Рис. 1. Спектограмма вхождения слова «видео»

Рис. 2. Спектограмма вхождения слова «Семенов»

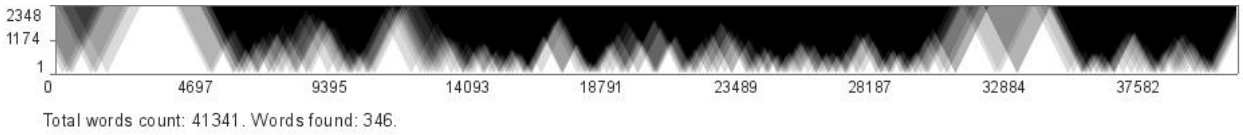

Рис. 3. Спектограмма вхождения слова «Комов»

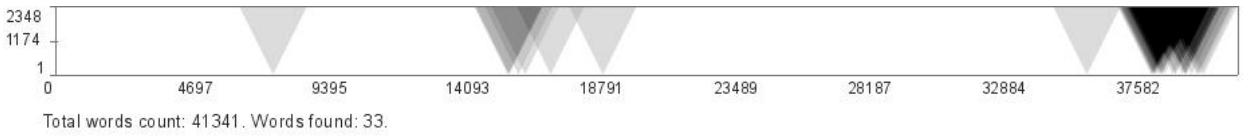

Рис. 4. Спектограмма вхождения слова «Гробовский»

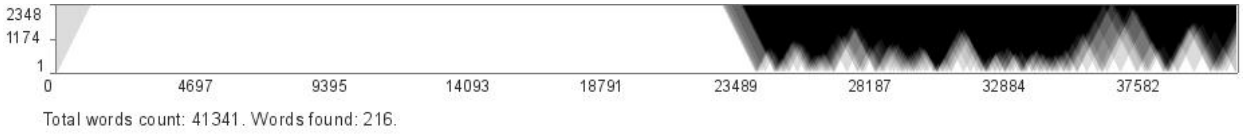

Рис. 5. Спектограмма вхождения слова «малыш»

Аналогичные расчеты были проведены для массива из 35 новостных веб-публикаций по тематике парламентских выборов в Украине в 2012 г. (табл. 2,

рис. 6-8). В этом случае монотонность возрастания значений w_i по отношению к σ_i не нарушается.

Табл. 2. Значения коэффициентов неравномерности для отдельных слов в массиве новостных сообщений

Слово	Вхождений	w_i (TF IDF)	σ_i
Америка	27	54,60	2,38
ЦИК	55	83,12	2,58
Тимошенко	86	115,63	2,87

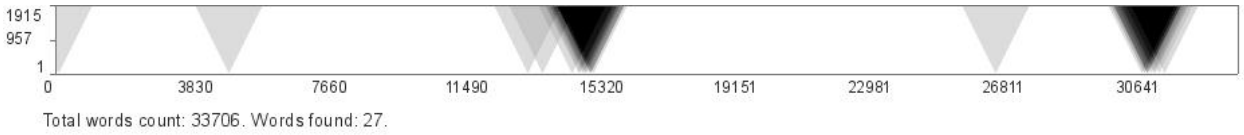

Рис. 6. Спектограмма вхождения слова «Америка»

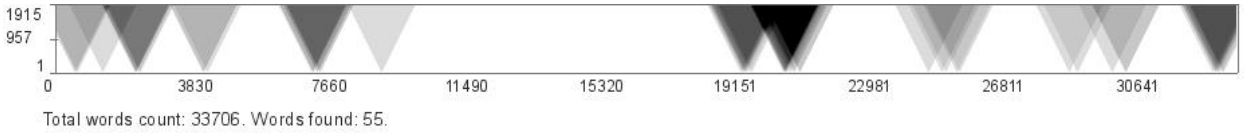

Рис. 7. Спектограмма вхождения слова «ЦИК»

Рис. 8. Спектограмма вхождения слова «Тимошенко»

Можно сделать вывод, что помимо традиционного подхода к оценке дискриминантной силы слов в текстах, предложенного Солтоном, дисперсионный анализ дает близкие по качеству результаты. Несмотря на то, что подход TF IDF за последнее время прошел ряд трансформаций, дополняется вспомогательными параметрами, в частности, получил популярность метод BM25, учитывающий длину документов, дисперсионный анализ оказывается волне перспективным.

Рассмотренные примеры показали, что искусственный прием, заключающийся в том, что исходный текст большого размера разбивался на фрагменты фиксированной длины, вполне оправдался, результаты во многом совпали с результатами, полученными другим методом.

Приведенные примеры показывают, неравномерность слов в массивах новостных сообщений и литературных произведениях имеет близкую, во

многом аналогичную природу, что выражается в близких значениях соответствующих коэффициентов.

И, наконец, предложенный метод визуализации неравномерности вхождения слов, по сравнению с существующими, добавил еще одно измерение – величину окна наблюдения, что оказалось удобным при рассмотрении текстовых массивов больших объемов. Техника спектограмм позволяет экспертам-лингвистам без дополнительных усилий качественно оценивать значения отдельных слов при формировании так называемых сверхфразовых единств.

Литература

1. *Ягунова Е.В., Ландэ Д.В.* Динамические частотные характеристики как основа для структурного описания разнородных лингвистических объектов // Труды 14-й Всероссийской научной конференции .Электронные библиотеки: перспективные методы и технологии, электронные коллекции. – RCDL-2012, Переславль-Залесский, Россия, 15-18 октября 2012 г. – С. 196-205.
2. *Salton G., McGill M. J.* Introduction to Modern Information Retrieval. – New York: McGraw-Hill, 1983. – 448 p.
3. *Ortuño M., Carpena P., Bernaola P., Muñoz E., Somoza A.M.* Keyword detection in natural languages and DNA // Europhys. Lett., 2002, 57. – P. 759-764.
4. *Carpena P., Bernaola-Galván P., Hackenberg M., Coronado A.V., Oliver J.L.* Level statistics of words: Finding keywords in literary texts and symbolic sequences // Phys Rev E Stat Nonlin Soft Matter Phys. 2009, E 79. – P. 035102-1 – 035102-4.
5. *Чуи К.* Введение в вэйвлеты. – М.: Мир, 2001. – 416 с.
6. *Ландэ Д.В.* Визуализация статистики вхождения слов // MegaLing'2009. Горизонты прикладной лингвистики и лингвистических технологий. Материалы международной конференции 21-26 сентября 2009 г., Украина, Киев / – К.: Довіра. - С. 63-64.