SMR-Cmp: Square-Mean-Root Approach to Comparison of Monolingual Contrastive Corpora

ZHANG HuaRui [1,2] *HUANG Chu-Ren*[1] *Francesca QUATTRI* [1]
(1) The Hong Kong Polytechnic University, Hong Kong
(2) MOE Key Lab of Computational Linguistics, Peking University, Beijing, China
hrzhang@pku.edu.cn, churen.huang@polyu.edu.hk,
quattri.francesca@gmail.com

ABSTRACT

The basic statistic tools used in computational and corpus linguistics to capture distributional information have not changed much in the past 20 years even though many standard tools have been proved to be inadequate. In this demo (SMR-Cmp), we adopt the new tool of Square-Mean-Root (SMR) similarity, which measures the evenness of distribution between contrastive corpora, to extract lexical variations. The result based on one case study shows that the novel approach outperforms traditional statistical measures, including chi-square (χ^2) and log-likelihood ratio (LLR).

KEYWORDS : Square-Mean-Root evenness, SMR similarity, corpus comparison, chi-square, log-likelihood ratio.

1 Motivation

Tools for detection and analysis of language variations are of foundational importance in computational/corpus linguistics. In general, most, if not all, NLP tasks (e.g. name entity recognition, disambiguation, information retrieval), are carried out based on distributional variations within the same text genre. On the other hand, distributional properties can be used to describe and account for language variations, such as the difference between two or more contrastive corpora. Such studies typically aim to locate and account for different lexical items in these contrastive corpora; but no satisfying quantitative ranking on the difference between the contrastive corpora is typically provided. In other words, there are no existent criteria to define what a meaningful ranking list of divergent words between contrastive corpora should look like. The ranking lists resulting from previous statistical comparisons have often been in conflict with intuition.

The same problem arises in the case of language learners who desire to learn significant words in a particular field. These categorical words are generally listed alphabetically and the list generated is often very long. We may ask - how could we assign a rank to the list so as to help foreign language beginners? In other words, how can we divide domain words into different levels of usefulness? Our research will also try to answer this question.

In the following, we first propose our solution based on Square-Mean-Root (SMR) evenness, then compare it with common statistical methods via a case study on American and British English.

2 Methodology

Our demo utilizes the novel statistical measure from Zhang et.al. (2004) and Zhang (2010).

2.1 Square-Mean-Root Evenness (DC)

The Distributional Consistency (DC) measure was proposed by Zhang et.al. (2004), and renamed as Square-Mean-Root evenness (Even_{SMR}) in Zhang (2010). SMR is the direct opposite of RMS (Root-Mean-Square) which is usually used in statistics. Gries (2010) provided a comprehensive comparison of dispersion measures, including DC.

SMR evenness captures the fact that if a word is commonly used in a language, it will appear in different parts of a corpus, and if it is common enough, it will be evenly distributed.

When a corpus is divided into *n* equally sized parts, SMR evenness is calculated by

$$\text{Even}_{\text{SMR}} = DC = \left(\sum_{i=1}^{n} \left(\sqrt{f_i} \right) \Big/ n \right)^2 \Big/ \left(\sum_{i=1}^{n} f_i \Big/ n \right)$$

where

f_i: the occurrence frequency of the specified word in the i$^{\text{th}}$ part of the corpus

n: the number of equally sized parts into which the corpus is divided

Σ: the sum of

When the whole corpus is divided into unequally sized parts, the formula becomes:

$$\text{Even}_{\text{SMR}} = DC = \left(\sum_{i=1}^{n}\sqrt{f_i C_i}\right)^2 \Bigg/ \left(\sum_{i=1}^{n} f_i\right)\left(\sum_{i=1}^{n} C_i\right)$$

with C_i denoting the occurrence frequency of all words that appears in the i$^{\text{th}}$ part of the corpus

The SMR evenness will decrease when some parts are further divided (*n* increases), however, this will not affect the effectiveness of comparison with the fixed number *n*.

2.2 Square-Mean-Root Similarity (bDC & mDC)

When comparing two contrastive corpora, there are two distributions *f* and *g*. The SMR similarity is calculated by the following formula (Zhang, 2010):

$$\text{Sim}_{\text{SMR}} = bDC = \sum_{i=1}^{n}\left(\left(\sqrt{f_i}+\sqrt{g_i}\right)\big/2\right)^2 \Bigg/ \left(\sum_{i=1}^{n}(f_i+g_i)\big/2\right)$$

When comparing three or more contrastive corpora, the formula becomes (Zhang, 2010):

$$\text{Sim}_{\text{SMR}} = mDC = \sum_{i=1}^{n}\left(\sum_{f}^{m}\left(\sqrt{f_i}\right)\Big/m\right)^2 \Bigg/ \sum_{i=1}^{n}\left(\sum_{f}^{m} f_i \Big/m\right)$$

where $\sum_{f}^{m}\sqrt{f_i}$ means sum over *f*, if there are three distributions called *f, g, h*, then it expands to be $\sqrt{f_i}+\sqrt{g_i}+\sqrt{h_i}$.

2.3 Difference Measure

The difference measure is based on frequency and SMR similarity.

Here we propose the following formula:

$$\text{Diff} = \text{Freq} \times (1-\text{Sim}_{\text{SMR}})^2$$

This formula is comparable with chi-square in terms of dimension. But it is symmetric to both sides being compared while chi-square is not. Although there can be a symmetric version for chi-square, the result is not satisfying as our experiment shows.

3 Comparison with Chi-square (χ^2) and Log-Likelihood Ratio (LLR)

In order to test the validity of our method, we extract the lexical difference between American English and British English via the Google Books Ngram dataset (Michel et al, 2010), which is the largest such corpus to the best of our knowledge. This enormous database contains millions of digitalized books, which cover about 4 percent (over 5 million volumes) of all the books ever printed. We utilize the American part and the British part during time span of 1830-2009 (180 years, *n*=180).

There have been various approaches to corpus comparison (e.g. Dunning, 1993; Rose and Kilgariff, 1998; Cavaglia, 2002; McInnes, 2004). We compare our result with more common approaches, including chi-square (χ^2), as recommended by Kilgarriff (2001), and log-likelihood ratio (LLR) recommended by Rayson and Garside (2000).

In Table 1, the top 30 words by each criterion (SMR, χ^2 & LLR) are listed. Almost every word in the list by our SMR measure is interpretable in the sense of being American or British except the word *cent* which demands further explanation.

From Table 1 we can see that there are large difference between the ranking of most biased words in AmE and BrE. On the left, almost every word in the list ranked by our method is obviously an American-dominant word or British-dominant word. In the middle, words in the list ranked by chi-square presents a mixture of biased words (e.g. £, *labour*, *centre*, *colour*) and unbiased common words (e.g. *which*, *you*, *of*), and both these example words appear in the top dozen. On the right, we can see somewhat similar or slightly better result in the list ranked by LLR.

It is interesting that *which* is ranked the 1st and 2nd position by χ^2 and LLR, respectively. This suggests that *which* should be a very biased word. But from Figure 1 we can see the frequency distribution in AmE and BrE. The trend is so similar that we can hardly know whether *which* is more American or British.

In our approach, *which* is outside the top 100 words. Instead, *color* is ranked the second (as shown in Figure 2). This is clearly more reasonable by intuition.

Another example is *of* (as shown in Figure 3), whose frequency is almost the same (after smoothing) through 90 percent of the time span investigated, is yet ranked the 3rd position by both χ^2 and LLR. By contrast, in our ranking by SMR, *of* is outside the top 1000 words.

Proportions of positive (in bold), vague, and negative (underline) contrastive words in three columns of Table 1:

SMR: **90%**; 10%; <u>0%</u>. (vague: *“, ”, cent.*)

χ^2: **40%**; 10%; <u>50%</u>. (top 3: *which*, *you*, *of*: all negative.)

LLR: **50%**; 10%; <u>40%</u>. (top 3: *you*, *which*, *of*: all negative.)

The conclusion we draw is that SMR is more appropriate than χ^2 and LLR for lexical difference detection between contrastive corpora.

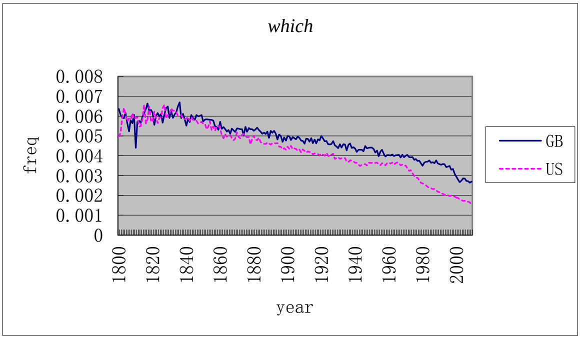

FIGURE 1 – *which*: with same trend in AmE and BrE, only different in quantity of use

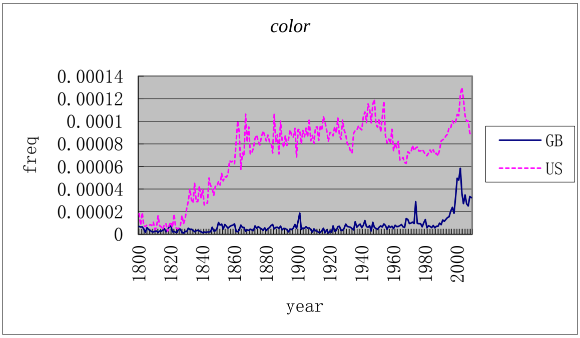

FIGURE 2 – *color* (AmE) is more frequent than *color* (BrE), although the latter experienced an increase in use around the year 2000

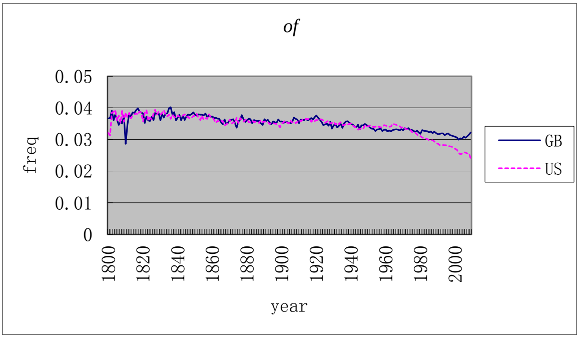

FIGURE 3 – *of*: Overlapping in AmE and BrE with a slight divergence from the 1980s on

Conclusion and Future Work

The Square-Mean-Root (SMR) approach clearly outperforms chi-square (χ^2) and LLR.

Future work includes the following :

(1) Exploring the theoretical nature of Square-Mean-Root (SMR)

(2) Extending to detection of lexical variation in Chinese, e.g. Mainland versus Taiwan

(3) Possible application in other NLP tasks, e.g. term extraction and document analysis

Acknowledgments

Thanks to Adam Pease for his kind help. This work has been supported by RGC GRF(543512) and NSFC(91024009).

No.	SMR rank	ratio: GB/US	Chi-square rank	χ^2 (x10^6)	ratio	LLR rank	LLR (x10^6)	ratio
1	labor	0.1	which	9.65	1.14	you	9.43	0.87
2	color	0.1	you	8.43	0.87	which	8.87	1.14
3	program	0.16	of	7.93	1.01	of	7.71	1.01
4	behavior	0.17	£	7.6	3.22	t	7.19	0.71
5	center	0.11	behaviour	6.56	5.14	£	6.01	3.22
6	programs	0.18	cent	6.52	0.99	her	5.64	0.93
7	labour	3.88	labour	6.38	3.88	toward	5.23	0.21
8	toward	0.21	t	6.19	0.71	–	5.1	0.9
9	favor	0.1	centre	5.79	2.28	cent	4.99	0.99
10	centre	2.28	towards	5.61	1.87	labor	4.9	0.1
11	colour	4.04	her	5.1	0.93	labour	4.88	3.88
12	favour	3.43	colour	4.79	4.04	behaviour	4.84	5.14
13	£	3.22	–	4.54	0.9	towards	4.46	1.87
14	“	0.36	was	4.48	1.07	program	4.42	0.16
15	”	0.35	programme	4.41	5.21	was	4.34	1.07
16	cent	0.99	et	4.32	1.92	centre	4.3	2.28
17	percent	0.16	toward	3.97	0.21	she	4.23	0.9
18	honor	0.14	she	3.79	0.9	percent	4.01	0.16
19	colored	0.07	the	3.73	1	color	3.97	0.1
20	whilst	2.82	favour	3.66	3.43	et	3.94	1.92
21	towards	1.87	is	3.45	0.98	colour	3.84	4.04
22	defense	0.12	labor	3.41	0.1	the	3.68	1
23	honour	2.94	my	3.38	1.08	behavior	3.67	0.17
24	behaviour	5.14	your	3.25	0.9	your	3.65	0.9
25	neighborhood	0.08	had	3.18	1.09	my	3.53	1.08
26	colors	0.09	de	3.11	1.5	is	3.37	0.98
27	railroad	0.09	he	3.1	1.04	he	3.23	1.04
28	defence	1.9	program	3.07	0.16	programme	3.2	5.21
29	favorable	0.09	his	2.92	1.05	center	3.16	0.11
30	favorite	0.11	me	2.83	1.04	had	3.13	1.09

TABLE 1 – Comparison of ranking by our SMR(left), chi-square(χ^2, middle) and LLR(right)

References

Cavaglia, G. (2002). Measuring Corpus Homogeneity Using a Range of Measures for Inter-Document Distance. *LREC 2002.*

Dunning, T. (1993). Accurate methods for the Statistics of Surprise and Coincidence. *Computational Linguistics.* 19(1): 61-74.

Google Inc. (2009). Google Books Ngrams (20090715 version). http://books.google.com/ngrams

Gries, S. Th. (2010). Dispersions and adjusted frequencies in corpora: further explorations. In Stefan Th. Gries, Stefanie Wulff, & Mark Davies (eds.), *Corpus linguistic applications: current studies, new directions*, pages 197-212. Amsterdam: Rodopi.

Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics*, 6(1):1-37.

Michel, J.-B., Shen, Y. K. et al. (2010). Quantitative Analysis of Culture Using Millions of Digitalized Books. *Science*. 331(6014):176-182.
http://www.sciencemag.org/content/331/6014/176

McInnes, B. T. (2004). Extending the Log Likelihood Measure to Improve Collocation Identification. (Master of Science, Thesis). University of Minnesota.

Rayson, P. and Garside, R. (2000). Comparing Corpora using Frequency Profiling. *Proceeding of the workshop on Comparing Corpora. ACL 2000.*

Rose, T. and Kilgarriff, A. (1998). Measures of Corpus Similarity and Homogeneity between Corpora. *EMNLP 1998*.

Zhang, HR, Huang, C.-R. and Yu, SW. (2004). Distributional Consistency: As a General Method for Defining a Core Lexicon. *LREC 2004.*

Zhang, HR. (2010). Quantitative Measure of Language Information Concentrated on Square-Mean-Root Evenness. (PhD Thesis). Peking: Peking University.