Influence of Orthographic Frequency of Words on the HAL Model of Semantic Space

Cyrus Shaoul & Chris Westbury
Department of Psychology, University of Alberta, Edmonton, AB, T6G 2E9 Canada.

INTRODUCTION

HAL (Hyperspace Analogue to Language) is a well known statistical model of lexical semantics that is based on word co-occurrence frequency in corpora of text. For HAL to be a general model of lexical semantics, it must be assumed that the co-occurrence distances it produces are not significantly altered by the choice of corpus that is the basis of these measures. In this study we test this assumption and demonstrate that it does not hold: co-occurrence distances are sensitive to some features of the corpus. We compared word frequencies in two very large (over 400 million word) corpora of English text (wire news articles versus web texts), and found that there were many words that had divergent orthographic frequencies. It has been pointed out by multiple researchers (Rodhe, Gonnerman and Plaut, 2004; Lowe, 2001) that there was a highly correlated relationship between HAL co-occurrence distances and orthographic frequency of the target word. Taken together these facts lead to the conclusion that the original HAL is not a suitable model for semantic spaces. We propose alternative models that are free from orthographic frequency dependencies.

META-QUESTIONS

- What is HAL?:
- Is it possible to improve on HAL?:
- Are large corpora truly different in terms of word frequencies?
- Will these differences in the corpora make a difference in the size and and direction of the HAL vectors?

Word frequency differences between two corpora

METHOD

We used the ACQUAINT corpus of news wire service text (435 million words) and a corpus that we call the MIX corpus that is 76% personal writings (blog entries) and 24% other text (literature, manuals, and other types of text). We then calculated the frequencies of 47,266 words in our lexicon for both corpora.

RESULTS

We calculated the correlations between the two orthographic frequency measures. The correlation of all 47,622 words was $r = 0.97$. This statistic hides some interesting information in the correlations of certain subsets of the lexicon. Table 1 reports the correlations for these subsets. To confirm the validity of these results, a second comparison was carried out between the frequencies in the CELEX database and the our web corpus frequencies. It was found that a similar pattern emerged: the top 100 most frequent words correlated at $r = 0.9$, the high frequency words at $r = 0.78$, the medium-high words at $r = 0.35$, the medium-low words at $r = 0.01$ and the low frequency words at $r = 0.07$. These results suggest that depending on the frequency range of the word, the correlation between the orthographic frequency of that word in the two corpora can be highly correlated or uncorrelated.

Frequency Category	r	r^2	Freq Range (words/million)	Percent of Lexicon
Top 100 words	0.97	0.93	53,000 > f > 864	0.2%
High Freq Words	0.30	0.09	864 > f > 100	2.1%
Med-High Freq Words	0.36	0.13	100 > f > 10	11.9%
Med-Low Freq Words	0.03	0.001	10 > f > 2	20.5%
Low Freq Words	0.48	0.23	2 > f > 0	65.2%
Top Quartile	0.97	0.93	53,000 > f > 3.8	25%
Second Quartile	-0.06	0.003	3.7 > f > 0.85	25%
Third Quartile	-0.12	0.01	0.84 > f > 0.22	25%
Bottom Quartile	0.25	0.06	0.21 > f > 0	25%

Table 1: Correlation of orthographic frequency in two large corpora

Also, contrary to assumptions made in much existing research, corpora may have very low correlations for all but the most frequently used words.

The influence of orthographic frequency on HAL

To test the influence of orthographic frequency on HAL, we created our own implementation of HAL that was slightly different from the one described by Burgess (1998). The only change that we made was that we modified the vector normalization process. In HAL, all word vectors are normalized by dividing all the values in each vector by the sum of the all the values in that vector. This was described as a way to mitigate the influence of orthographic frequency on the size of vectors. In our vector normalization process, each value in each vector is divided by the orthographic frequency of the word itself. By doing this, we remove the influence of orthographic frequency on the results of the HAL calculations. We call the average distance between a word and its 20 nearest neighbors ARC (Average Radius of Co-occurrence).

METHOD

We received BSD (Burgess SD) measures for 4218 words (mean orthographic frequency = 120 words/mil, mode = 2.8 words/mil, $\sigma = 548$) from Curt Burgess, and calculated the ARC for all of these words. We then used mathematical methods (Hollis and Westbury, in press) to search for the best non-linear relationship between BSD and OFREQ and ARC and OFREQ that it could find.

RESULTS

The algorithm found a relationship of the form: BSD = 1/SQRT(OFREQ) that correlated at $r = 0.71$. The same relationship applied to ARC, ARC = 1/SQRT(OFREQ) correlated at $r = 0.04$. The best fit that could be found using mathematical methods for predicting ARC from OFREQ was $r = 0.022$.

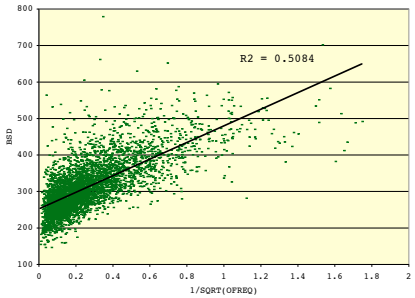

Figure 1: BSD predicted by 1/SQRT(OFREQ)

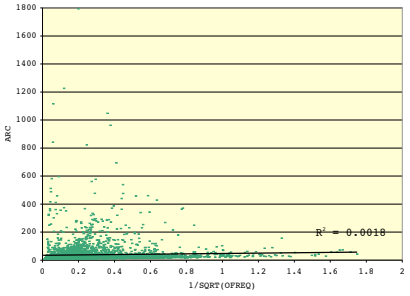

Figure 2: ARC predicted by 1/SQRT(OFREQ)

CONCLUSION

We found that different corpora can have very different frequencies for the same words. We also found that the HAL model's measure of neighborhood size is influenced strongly by orthographic frequency. There is a need to find a better model than HAL to capture the semantic information in lexical co-occurrence. We have described one method for doing this. Future work will focus on comparing this method to other methods, such as the use the conditional probability of two words co-occurring as the measure of semantic information (Rohde, Gonnerman and Plaut, 2004).

REFERENCES:
Burgess, C. (1998). From simple associations to the building blocks of language: Modeling meaning in memory with the HAL model. Behavior Research Methods, Instruments, & Computers, 30, 188–198.

Lowe, W. (2001) 'Towards a theory of semantic space' Presented at Twenty-first Annual Conference of the Cognitive Science Society.

Rohde, D. L. T., Gonnerman, L. M., & Plaut, D. C. (2004). An improved method for deriving word meaning from lexical co-occurrence. Massachusetts Institute of Technology, Cambridge, MA. (Retrieved September 20th, 2004, from http://tedlab.mit.edu/~dr/).

Song, D., Bruza, P.D., Cole, R.J. (2004) Concept learning and information inferencing on a high-dimensional semantic space. ACM SIGIR 2004 Workshop on Mathematical/Formal Methods in Information Retrieval (MF/IR'2004), 29 July 2004, Sheffield, UK.

Hollis, G. & Westbury, C. (In press). NUANCE: Naturalistic University of Alberta Nonlinear Correlation Explorer. Behavior Research Methods, Instruments, and Computers.

Acknowledgements: This work was supported by the National Science and Engineering Research Council of Canada.
Geoff Hollis and Emilio Gagliardi also contributed to this project.