# Finding Meaning in the Chaos

## *Establishing Corpus Size Limits for Unsupervised Semantic Linking*

**Gary Munnelly[1], Alexander O'Connor[3], Jennifer Edmond[2], and Séamus Lawless[1]**

1:ADAPT, KDEG, School of Computer Science & Statistics,
2:Long Room Hub,
Trinity College, Dublin, Ireland
3: ADAPT, School of Computing
Dublin City University, Dublin, Ireland
{munnellg, edmondj}@tcd.ie, Seamus.Lawless@scss.tcd.ie, alexander.oconnor@dcu.ie

## Abstract

The study and analysis of language as a subject in computer science is both an important and challenging field. Much emphasis has been placed on discovering methods of allowing the computer, not only to understand the meaning of the text that it is reading, but also to infer new meaning for words given the context in which they is mentioned. The first step in such a problem is to attempt to determine which words may have similar meaning, given various contexts in which they are used together. Two algorithms which have been applied to this task are Latent Dirichlet Allocation (LDA) [1] and Latent Semantic Indexing (LSI) [2]. Both attempt to group the words of a corpus under a predefined number of headings where words belonging to the same cluster are assumed to be related to the same topic [4].

In order to establish these relationships, the computer must have enough data (i.e. text) with which to work. This invariably begs the question how much is "enough"? Investigations have already been conducted to determine an appropriate number of topics to use given the size of the corpus [3]. However, the question that we propose is, how large must a corpus be in order to establish stable topics given that the number of topics is fixed?

Our approach to this problem has been to choose a corpus which is certainly big enough to build a model. In this case, the entire contents of Wikipedia. A model is built on this information and taken as a gold standard for the topics in the collection. We then reduce the size of the corpus by 10% and build a new model, hopefully getting similar topics to those we found previously. Note that the reduction is achieved by removing random sentences from the corpus. This is to ensure that we don't unintentionally remove an entire subject from the collection, say by removing all articles whose titles begin with "Z". This process of reducing and rebuilding continues until we have only a small percentage of the corpus left.

We now have a number of models, all built by the same method on reduced versions of the same collection. We can evaluate the quality of each model by comparing the contents of its topics to those of the gold standard model we originally built. This may be achieved using DICE [6] or the Jaccard Index [7] between the sets. The point at which we see a large deviation between the model of a reduced corpus and that of the gold standard is the point at which we say the models become unstable.

To date, we have generated the reduced collections and are in the process of building the models with Wikipedia. This is a time consuming and costly process, however. To facilitate faster testing, we have also begun an investigation using the works of H.P. Lovecraft as a corpus [5]. This collection is much smaller and hence much easier to work with, but as yet it is unclear as to how useful the results of that investigation will be.

## References

1. *David M. Blei, Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." J. Mach. Learn. Res. 3 (2003), 993-1022.*
2. *Thomas K. Landauer,  Peter W. Foltz, and Darrell Laham. "An introduction to latent semantic analysis." Discourse processes 25.2-3 (1998): 259-284.*
3. *Derek Greene, Derek O'Callaghan, Padraig Cunningham,. How Many Topics? Stability Analysis for Topic Models. Lecture Notes in Computer Science Vol 8724 (2014), 498-513*
4. *Matthew L. Jockers. "Topic Modeling" Text analysis with R for Students of Literature (2014) Ch. 3, 135-159*
5. *Donovan K. Loucks. "The H.P. Lovecraft Archive", http://www.hplovecraft.com/ (2015). Last Accessed: 28/10/2015*
6. *Lee R. Dice. "Measures of the amount of ecologic association between species." Journal of Ecology (1945) 26:297-302.*
7. *Real, Raimundo, and Juan M. Vargas. "The probabilistic basis of Jaccard's index of similarity." Systematic biology (1996): 380-385.*