DiaView: Visualise Cultural Change in Diachronic Corpora

Introduction

This paper will introduce and demonstrate *DiaView*[1], a new tool to investigate and visualise word usage in diachronic corpora. DiaView highlights cultural change over time by exposing salient lexical items from each decade or year, and providing them to the user in an effortless visualisation. This is made possible by examining large quantities of diachronic textual data, in this case the Google Books corpus (Michel et al., 2010) of one million English books. This paper will introduce the methods and technologies at its core, perform a demonstration of the tool and discuss further possibilities.

Applicable Corpora

Key to the success of any large-scale cultural inspection is a large corpus of writing, both in terms of chronological span, and also depth of sampling from each year. Two corpora stand out as candidates for this approach: *The Corpus of Historical American English* (COHA) (Davies, 2010) and the Google Books corpus. The COHA dataset, while being 400 million words and having been more rigorously compiled, is restricted in terms of its availability. Google Books, on the other hand, is 155 billion words, spanning from the sixteenth century to the present, although the precise corpus make-up is less lucid.

Google have provided n-gram frequency lists for download, and consequently this visualisation has been based upon the Google Books *English One Million* data set. It should be stressed that the techniques DiaView are applicable to many other data sets, from very focused specialised corpora to large-scale newspaper archives. The choice of corpus used for this demonstration is mainly

driven by two factors: public availability and re-use in addition to the need for general content leading to a wider public appreciation of the results this tool provides.

Established Query Methods

The standard tools available to current corpus users revolve around two methods: word searches or the comparison of frequencies (either over time or between two focussed points). These tools are in widespread use, and while well understood and used, they do not present the whole linguistic picture to their users.

The *Google Books n-gram Viewer*[2], for instance is very capable at allowing users to chart the usage of lexical items over time. While illuminating, it implies that users already know what they are looking for (the lexical item) and also that frequency (the number of times that word is used) provides enough data to make assertions. Choosing the term of interest at the outset means this resource delivers a powerful search function, but it fails at browsing or exploration. In other terms you only find what you look for.

The alternative method, based upon comparing frequencies over time, is best exemplified (using the same corpus) by the *Top Words 1570-2008*[3]. This tool meticulously visualises the most frequent words in the Google Books corpus and charts their rise and fall across every decade the corpus covers. Word frequency alone does not tell us enough. Just because a lexical item is frequently occurring in a particular time span, it may not be of interest, particularly if you find it frequently elsewhere. That word soon downgrades from important to common and suddenly becomes much less interesting.

DiaView takes a different approach, one that promotes browsing (not searching) and transcends basic word frequency methods (it strives to deliver salience).

DiaView Concept

DiaView is designed to operate at a more opportunistic conceptual level than the examples above; in this case summarising and browsing data takes precedence over specific focussed searches. DiaView aims to complement other approaches, and as such will operate as a gateway to other methods, particularly lexical searches.

The tool divides its corpus into a number of chronological divisions, i.e. years or decades. In each, statistical measures are used to extract those lexical elements that are salient or focussed on that time-span. This does not necessarily imply frequently used words, but, rather, words which are found predominantly in that time, a possible key into the cultural issues experienced or explored at that moment in time.

DiaView Method

A relational database has been used to realise DiaView, however other technologies are equally applicable to producing similar results. Equally, any number of statistical measures or comparators can be used to ascertain the salient words for each year, *mutual information* or *log-likelihood* are good candidates for use.

For this demonstration DiaView requires word frequency lists for each year. In this case the corpus was trimmed to 1900-1950. Fundamentally this is a tabulated list of lexical items, followed by the number of occurrences in each year.

DiaView then creates an aggregated view of these frequency lists, gathering each *type* (individual word) and the total number of occurrences across the entire corpus (summing the data above). A further optional stage is to cull this list; either by disposing of the least frequently used types (e.g.

remove words not used more than ten times each year) or by concentrating on the most frequently used types (e.g. keep the 100,000 most often used words).

Extracting salient terms from each year is now performed. Taking each type (identified previously) in turn, its frequency in each year is inspected and compared to its global frequency. Here the statistical measure is used to gauge its connection to each particular year. This removes the dependence of raw frequencies: a word may occur few times or a great number, what is of prime importance is if its distribution is skewed or focussed on a particular chronological range. For each year a ranked list of salient types are created.

The visualisation is created by extracting the salient types from each year and displaying the top 25 in descending order. Each type in each year can be used to hyperlink other resources, such as the Google Book n-gram Viewer.

Future possibilities

While DiaView offers new ways to view large data sets, it is open to further enhancements. Access to corpora divided by genre would add valuable benefits to the visualisation, allowing users to narrow down the corpus to include material only of their choice. Linguistic stemming could also be used to gather words around their root form, e.g. to cluster run, running, runs, ran etc.

Notes

1 http://www.scottishcorpus.ac.uk/corpus/diaview/

2 http://books.google.com/ngrams

3 http://pages.cs.wisc.edu/~gleicher/Hacks/tops.html

References

Davies, M. (2010-). *The Corpus of Historical American English: 400 million words, 1810-2009*. Available online at http://corpus.byu.edu/coha/

Michel, J., Shen, Y., Aiden, A., Veres, A., Gray, M., Brockman, W., The Google Books Team, Pickett, J., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. and Aiden E. (2010). Quantitative Analysis of Culture Using Millions of Digitized Books. *Science* (Published online ahead of print: 12/16/2010)