Diachronic Word Sense Change Identification

Ankit Singh 12128
T Raghuveer 12762

October 2015

Motivation

Language has continuosly changed over time. New senses for some words took birth while some old senses demised. For an example the word 'Artificial' had a positive sense in the previous centuries but now has a negative sense associated to it. Some other examples include 'gay' which meant a noble person in the previous century but is now used to refer to a sexual orientation; 'sick' referred to illness in the past and is now being reffered to as something crazy or cool;

It has been theorized that polysemy in language is simply a transitory phase of word evolution where a new sense evolves with time and competes with existing senses. This time sense disambiguation is also highly instrumental in *culturonomics*; to analyse the changes in human culture and historical phenomena by evaluating usage of various words. Understanding the changes in meaning and usage of words is highly important for people working with historical texts, such as librarians, historians and linguists. It is also helpful to lexicographers and design engineers in a variety of NLP/IR tasks.

Given the availability of Diachronic datasets and the computational ability, we in this current work, aim to look at evolution of semantics over time; identify and report words where change of sense has occured.

Problem Statement:

The task at hand is to devise an unsupervised approach to determine semantic change, i.e. transition in sense of words over time based on extensive analysis of diachronic text data available in the form of N-grams of digitized books.

To visually present the word sense changes identified in *part 1* over a time frame and to discuss and compare results of our approach with existing *state of arts.*

Methodology:

The existing arts in this field of diachronic word sense change identification disambiguate words by measuring cosine similarity of co-occurence matrix vectors created using 5-grams [2]. Some other arts compare polysemantic clusters over different time epoches to identify words with multiple senses and their changes [3].

In our approach, we align vectors of different epoches using methodology in [1]. We first split the dataset into multiple epoches based on their time period. We then individually train each epoch using Word2Vec. We then pick some selected words common to all the epoches whose meaning has not changed over time. For given two epoches, we find a transformation matrix which transforms the above selected words from one epoch to other. We then take all words from the first epoch, transform them using the transformation matrix and then compare the cosine distance with the same words in the second epoch. Words with low cosine distances are expected to have their sense changed. We do such evaluations for all pairs of epoches and correspondingly plot the changes over inter-epoch heat diagrams.

If time permits, we also wish to create clusters based on word context on each epoch and then train Word2Vec individually to incorporate polysemantic word meaning changes into our approach.

Dataset:

Google books N-gram dataset (Version 2009) will be used, as it is the largest historical corpus available. It provides the N-gram(1 to 5) count information by year. The corpus contains books from 1505 to 2008, which captures change in word senses over a long period of time. The original corpus contains around 200 billion words but we will use a limited segment of the corpus for our purpose.

References:

[1].Aligned Word Vector Spaces and Document Vectors, Shashwat Chandra.
http://172.28.64.70:8080/jspui/handle/123456789/15193

[2] A Framework for Analyzing Semantic Change of Words across Time, Adam Jatowt, and Kevin Duh.
http://cl.naist.jp/~kevinduh/papers/jatowt14change.pdf

[3].That's sick dude!: Automatic identification of word sense change across different timescales.Sunny Mitra, Ritwik Mitra, Martin Riedl, Chris Biemann, Animesh Mukherjee, Pawan Goyal.
https://www.aclweb.org/anthology/P/P14/P14-1096.pdf

[4].Word Epoch Disambiguation: Finding How Words Change Over Time. Rada Mihalcea,Vivi Nastase. http://www.aclweb.org/anthology/P12-2051

[5] Understanding Semantic Change of Words Over Centuries, Derry Tanti Wijaya and Reyyan Yeniterzi. http://rtw.ml.cmu.edu/papers/wijaya-detect11.pdf