Diachronic Embeddings

Modelling the semantic change over time

Author: Aymane Hachcham

Department of Statistics

Table of Contents

1. Motivation

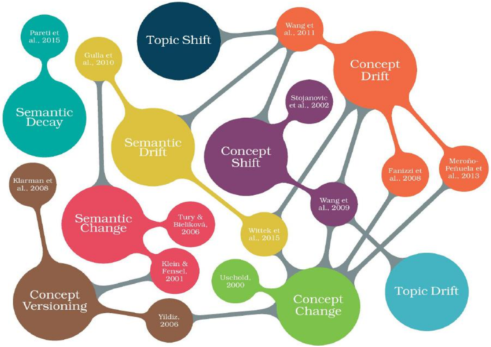
2. Methodology

Diachronic Analysis

- The term **Diachronic** refers to the study of how something changes over time. In linguistics, *Diachronic analysis* is concerned with the way that language changes over time.
- It Examines how to following evolves over time:
  - How forms and meanings of words
  - Grammatical structures
  - Pronunciation

Diachronic Sematic Shifts

- An active research Field.
- Research possible because of:**Existence of large corpora** and **development of computational semantics**.



Diachronic Semantic Shifts

First Research

- *(Bloomfield 1933)*: *"innovations which change the lexical meaning rather than the grammatical function of a form."*
- *(Bréal, 1899; Stern, 1931; Bloomfield, 1933)*: Found the 9 most prominent categories in semantic shift.
- *(Blank and Koch, 1999; Grzega and Schoener, 2007)*: Determined the driving forces for semantic change.
- *(Mikolov et al., 2013b)*: Used word embeddings to model Diachronic Semantic change.

Diachronic Sematic Shifts

Types of Semantic Shifts

Theoretical linguists have identified regularities in language change and have described various types of lexical semantic shifts:

- Narrowing/Broadening the sense
- Positive/Negative connotations
- Cultural Changes

Examples:

- *"mete"* (food, all kinds of food) *"meat"* (edible flesh)
- *"gay"* (joyful, cheerful, sweet) *"gay"* (Homosexual)
- *"Iraq"*, *"Syria"* (Cities in Middle East) *"Iraq"*, *"Syria"* (Synonyms of war).

Diachronic Sematic Shifts

Drivers

The drivers or factors that lead to Sematinc Shifts can be:

- Linguistic
- Psychological
- Sociocultural

Tasks and Methods

Objective: Study the evolution in the meanings of these words over time by comparing the representations of their meanings across different periods.

Modelling semantic shifts

- The research investigates the way in which the meanings of words change over time in a corpus of documents.
- The documents are divided into different time periods, and the meanings of target words are analyzed by examining how they are used in context during each time period.

Tasks and Methods

Tracing Semantic Shifts

From a collection of corpora $[C_1, C_2, ..., C_n]$ with periods of time of different granularity ranging from $[1, 2, ..., n]$, we can trace the words that change in meaning most often.

Other Methods

Other important tasks:

- Quantifying the degree of semantic change of each word in a corpus.
- Detecting whether words undergo semantic change or not in a corpus.
- Interpreting the change undergone by a word.

Tasks and Methods

Two main approaches

Word frequency

Analyzing the statistical distribution of words across different time periods. This involved calculating the frequency of words in each time period (Michel et al. 2011)

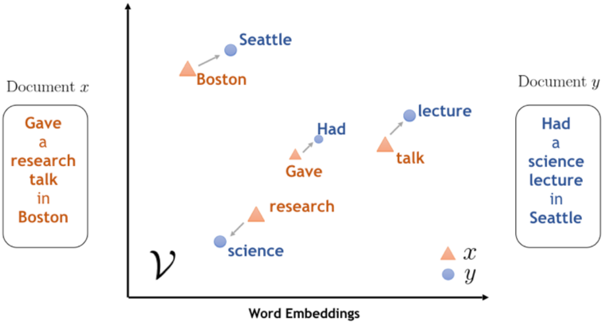
Word co-occurrences

Word collocations, or the words that tend to occur together, can be useful for understanding how the general context of a word changes over time:

- (Hilpert 2006): Compute statistical dependency between pairs of words at different times slices in a corpus.
- (Sagi et al. 2009): Words are represented using Singular Value Decomposition on a condensed version of a matrix of co-occurrences.

Neural word Embeddings

Latest research focuses on word embeddings, which are real-valued vectors that represent a word and its usage based on the contexts in which it appears.



Neural word Embeddings

Main idea: Word embeddings are an extension of distributional similarity methods, which are based on the idea that words with similar meanings tend to appear in similar contexts.

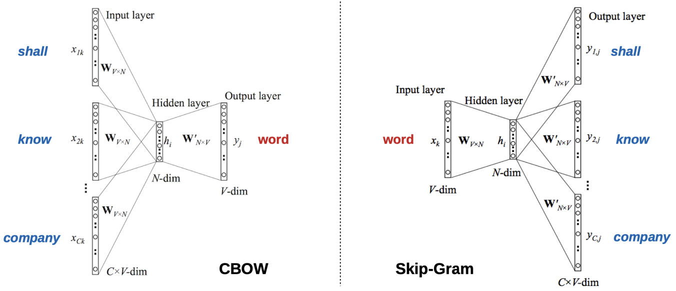
Methods for learning word embeddings

- Word2Vec (Mikolov, Chen, Corrado, and Dean, 2013), which has two algorithms called Continuous Bag of Words (CBOW) and Skip-Gram.
- Glove (Pennington, Socher, and Manning, 2014), which is based on the factorization of a word-context co-occurrence matrix.
- FastText (Bojanowski, Grave, Joulin, and Mikolov, 2017) is another algorithm for learning word embeddings that handles out-of-vocabulary words.

Focus on Word2Vec

The Word2Vec framework is made up of two models, called Continuous Bag of Words (CBOW) and Skip-Gram.

“ You *shall know* a **word** by the *company* it keeps”



Focus on Word2Vec

Continuous Bag of Words (CBOW) and Skip-Gram, are both two-layer neural networks that are designed to learn the linguistic contexts of words and represent them in a vector space.

- CBOW predicts which word is most likely to appear based on the context in which it appears.
- CBOW treats the full context of a word as a single observation.
- Skip-Gram uses the network to predict the context words around a given target word.
- Skip-Gram treats each context-target pair as a separate observation.