Diachronic Word Sense Change Identification

Ankit Singh[1]; T. Raghuveer[2];
[1]Department of Computer Science & Engineering; [2]Department of Electrical Engineering
Supervisor: Amitabha Mukerjee

ABSTRACT

Language has continuously changed over time. New senses for some words took birth while some old senses demised.

For an example the word 'instant' which referred to the current month about a century ago but is now used to refer to a very small brief time; 'economy' referred to management of resources in the past and is now being referred to state of country in terms of productions.

Diachronic word meaning change identification has been done by making vectors of words using Word2Vec over different time epochs; aligning the vectors across the epochs and finally used cosine similarity to identify the words.

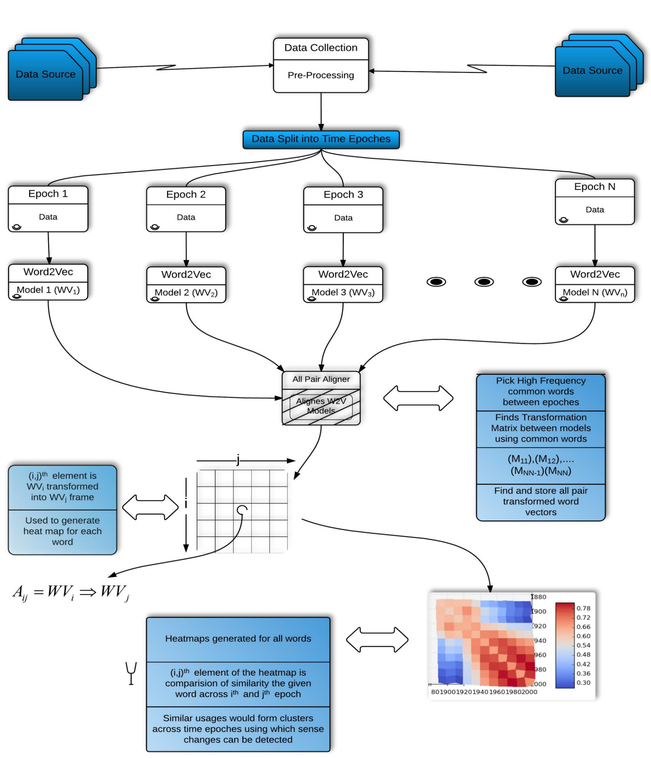
INTRODUCTION

Diachronic means dealing with phenomena (as of language or culture) as they occur or change over a period of time. Language changes over time; words acquired new meanings and cease to be used according to the old meanings.

The time sense disambiguation is also 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.

Adam Jatowt et al. identified word sense changes by measuring cosine similarity of co-occurrence matrix vectors created using 5-grams[1]. Sunny Mitra et al. compared polysemantic clusters over different time epochs to identify words with multiple senses and their changes[2].

METHODOLOGY



IMPLEMENTATION

Collection of Dataset:

English:
British Parliament debate corpus was extracted from Hansard archive in zip format. The 1500 zip contained debates from 1892-2005 in the form of XML files. XML file was pre-processed to extract year and text of use.
Hindi:
Articles, essays and poems were extracted from hindisamay.com. Large HTML pages were processed to extract year and imp text.

Procedure:

This work is based on Shashwat Chandra's M.Tech thesis of 'Aligned Word Vector Spaces and Document Vectors'[3]. Word2Vec[4] is trained on all epochs. Then we use the following formula for the calculation of the transformation matrix.

$$W_i = M_{ij}.W_j + b$$

Converting the formula in matrix form we get:

$$\begin{bmatrix} W_i \\ 1\ldots1 \end{bmatrix} = \begin{bmatrix} M_{ij} & b \\ 0\ldots0 & 1 \end{bmatrix} . \begin{bmatrix} W_j \\ 1\ldots1 \end{bmatrix}$$

$$[M_{ij} \mid b] = W_j.(W_i)^+; \quad (W)^+ = W^T(WW^T)^{-1}$$

W_i - Transformed word vector matrix, M_{ij}- Transformation matrix
W_j - Word vector matrix from initial set, b- Bias term for translation

In evaluation of M_{ij} and b, we pick 600 most frequent words across all the epochs; concatenate their vectors as columns to form W_i and W_j matrices and use the LS solution to find an estimate of M_{ij}, b.
The transformation matrix is calculated for all epoch pairs ($^NC_2$).

$$A_{ij} = [M_{ij} \mid b].W_i \qquad [M_{jj} \mid b] = I$$

A_{ij} here corresponds to WV_i transformed into WV_j. Now, given a word we construct a heat-map using the transformed models as:

$$HeatMap(i, j \mid word) = cos\_sim(A_{ij}[word], A_{jj}[word])$$

In a heat-map, similar senses of a word are clusters over Epoch axes and hence change in sense of a word can be detected. We only consider words with max. deviation was greater than 0.6 and std. deviation greater than 0.3; plot and identify sense changes.

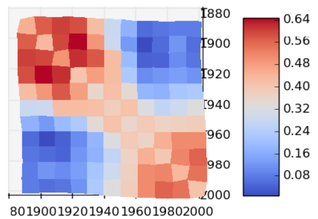


Fig1: Heat Diagram for 'instant'

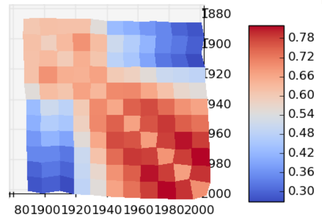


Fig2:Heat Diagram for 'economy'

RESULTS

Some words were found to have two different similarity clusters in their heat map indicating change in their sense over time. The intersection point of these two clusters indicates the epoch where the change in sense appeared.

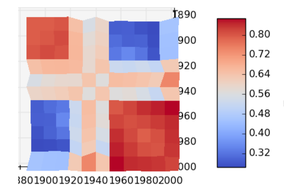


Fig3: Heat Diagram for 'east'

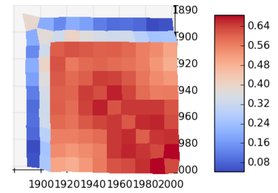


Fig4: Heat Diagram for 'deflected'

Words in Fig 1, Fig 2 i.e. 'instant', 'commissioned' have changed/added senses in the corpus in the two clusters formed as shown.
Instant – month to brief instant
'Economy' – management to financial market
'east' in Fig 3 corresponds to a negative example, because of varied contexts that appeared in the corpus.
Fig 4 corresponds to 'deflected' which kept same sense over epochs. The row and column corresponding to 1890 epochs shows low similarity with other epochs because the word 'deflected' occurred with low frequency.

FUTURE WORK

- Handle polysemy within one epoch and appropriately train word2vec separately for different words within an time epoch ; propose birth and death; and merge and split of the senses.
- Use of Distributed thesaurus to train Word2Vec over Google 5-gram model.

REFERENCES

[1] A Framework for Analyzing Semantic Change of Words across Time, Adam Jatowt, and Kevin Duh.

[2].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.

[3].Aligned Word Vector Spaces and Document Vectors, Shashwat Chandra.

[4]. Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean; Efficient Estimation of Word Representations in Vector Space.