Diachronic Word Sense Change Identification

Bass: fish

???

Bass: instrument

Ankit Singh 12034
T Raghuveer 12762
Problem Statement

• To devise an unsupervised approach to identify words which have semantically changed over time.

• Perform the task on multiple epochs in multiple languages. – English, Hindi and Mandarin if possible
Motivation

- Time sense disambiguation is highly instrumental in *culturonomics, etymology*.

- 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.
Examples

- ‘Gay’ - Noble person (early 20th century)
  - Homosexual (21st century)

- ‘Artificial’ - positive sense (early 20th century)
  - negative sense (current decade)

- ‘Sick’ - illness (20th & 21st century)
  - crazy or cool (21st century)

- ‘Bachelor’ - young knight (Long ago)
  - University affiliated (21st Century)

- ‘Flirt’ - jerky motion (500 years ago)
  - .................. (21st Century)
Existing Approaches

- **Haim Dubossarsky et al.**—Word2vec trained on Google books, initializing with word vectors of prev. epoch. Change in distance from centroid of K means clusters => sense change. [1]

- **Adam Jatowt et al.** - measured cosine similarity of word vectors created using frequency of co-occurring word on Google books 5-grams model. [2]

- **Sunny Mitra et al.** - Chinese Wispers over a co-occurrence graph vectors from distributed thesaurus (Riedl and Biemann, 2013), then compare clusters for birth, death, merge or split. [3]
Our Approach

- Adapted from work of Shashwat Chandra. [4]

\[ W_t = M \cdot W_s + b \]

- Used to transform and align vectors trained on two different datasets

  \( W_t \) - transformed word vector
  
  \( M \) - transformation matrix (train.set1 to train.set2)
  
  \( W_s \) - word vector from first set
  
  \( b \) - bias term for translation

\[
\begin{bmatrix}
  W_t \\
  1 \ldots 1
\end{bmatrix}
= 
\begin{bmatrix}
  M & b \\
  0 \ldots 0 & 1
\end{bmatrix} 
\cdot 
\begin{bmatrix}
  W_s \\
  1 \ldots 1
\end{bmatrix}
\]
Our Approach ..

- $M$ and $b$??
  - Use words whose meaning has not changed over the two datasets and find an estimate

- minimize $\| M \cdot W_s - W_t \|_F$

- Least square solution helps

\[
\begin{bmatrix}
M & b \\
0 \ldots 0 & 1
\end{bmatrix}
= \begin{bmatrix}
W_t \\
1 \ldots 1
\end{bmatrix}
\cdot
\begin{bmatrix}
W_s \\
1 \ldots 1
\end{bmatrix}^+
\]

\[
A^+ = A^T (A A^T)^{-1}
\]
Our Problem fits..

• Build datasets corresponding to different epochs
• Train Word2Vec over each epoch individually.
• Find transformation matrices for all pairs/ convert all into one frame
• Vectors can then be compared between time epochs, once we have them in the same frame.
Challenges

- Filip Ginter et al. [5] –

<table>
<thead>
<tr>
<th>Task</th>
<th>Finnish</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>base</td>
<td>n-gram</td>
</tr>
<tr>
<td>Wordsim</td>
<td>22.95</td>
<td>19.28</td>
</tr>
<tr>
<td>SRL</td>
<td>63.81</td>
<td>66.29</td>
</tr>
</tbody>
</table>

- Dataset Collection
  - No directly available dataset with diachronic tags
  - Dataset corresponding to different languages was to be generated

- Handle existing polysemy within a time epoch.
Progress so far..

• Collected out the British parliament debate dataset, distributed over the span of 1890 to 2007
  ◦ Had to scrape out over 1500 links and extract xml text ~8gb
  ◦ Process the xml files extracted for specific parts

• Collected Hindi stories, fiction, and novels from hindisamay.com.
  • Around 250 fiction, 500 stories spread over 250 years

• Transformation matrix code done. WordVec training code done. Training and all pair comparison needs to be done.
Further Thoughts..

- Train and test over *Peoples daily* (simplified Chinese).

- Handle polysemy within one epoch and appropriately train word2vec separately for different words within an time epoch; propose birth, death merge or split of a sense.

- Use of Distributed thesaurus to train word2vec over Google 5-grams
References


