Statistical Semantic Change Detection via Usage Similarities

Taichi Aida[1], Daichi Mochihashi[2], Hiroya Takamura[3], Toshinobu Ogiso[4], Mamoru Komachi[1]

[1]Hitotsubashi University
[2]The Institute of Statistical Mathematics
[3]National Institute of Advanced Industrial Science and Technology
[4]National Institute for Japanese Language and Linguistics

Lexical Semantic Change

The meaning of words naturally evolve across different periods/domains.

- *plane* in 19th:
  - "If a *plane*() be parallel to the horizontal…"
  - "The sun is in the same *plane*() as the picture…"
- *plane* in 20th:
  - "The President's *plane*() landed at Goose Bay…"
  - "The *plane*() kept climbing and climbing…"

Lexical Semantic Change Detection

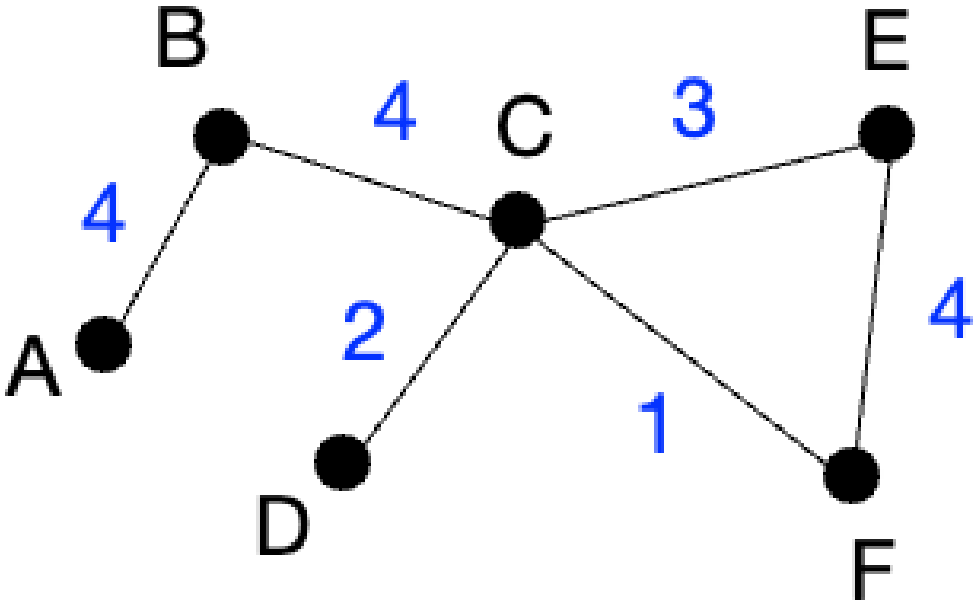
Detecting semantically changed words across different periods/domains.

- *plane* in 19th:
  - "If a *plane*() be parallel to the horizontal…"
  - "The sun is in the same *plane*() as the picture…"
- *plane* in 20th:
  - "The President's *plane*() landed at Goose Bay…"
  - "The *plane*() kept climbing and climbing…"

Key problem: **preparing annotated words**

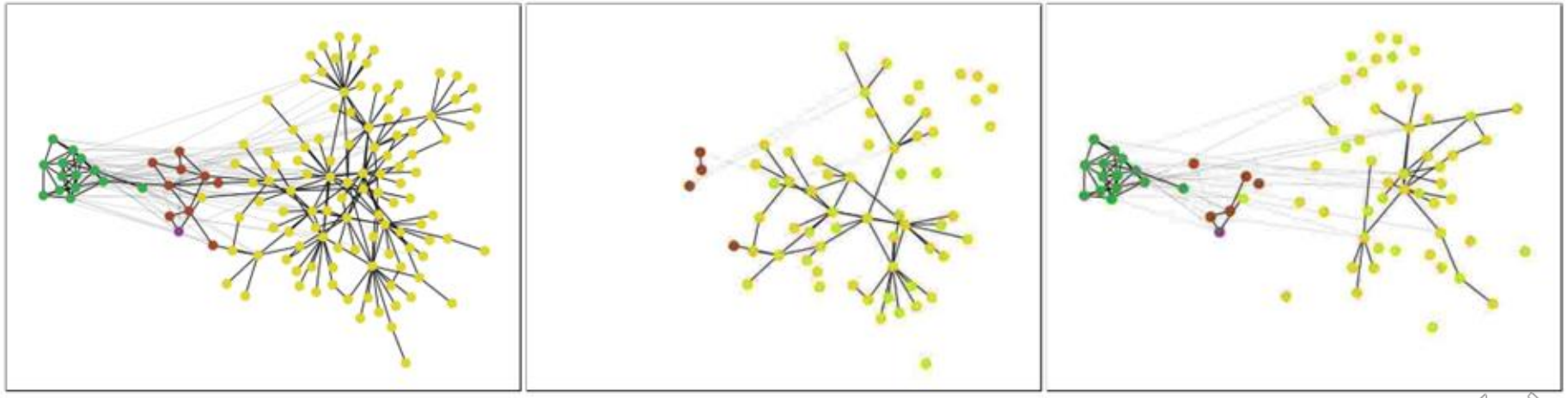
DURel Annotation Pipeline [Schlechtweg+2018]

- **Annotating word usage** (A-F) via usage similarity (1: Unrelated - 4: Identical)
- It can be viewed as an annotation graph = **word usage graph**



Word Usage Graph [Schlechtweg+2021]

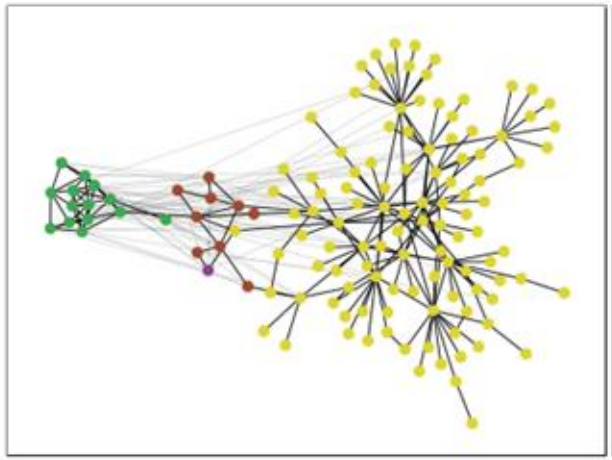
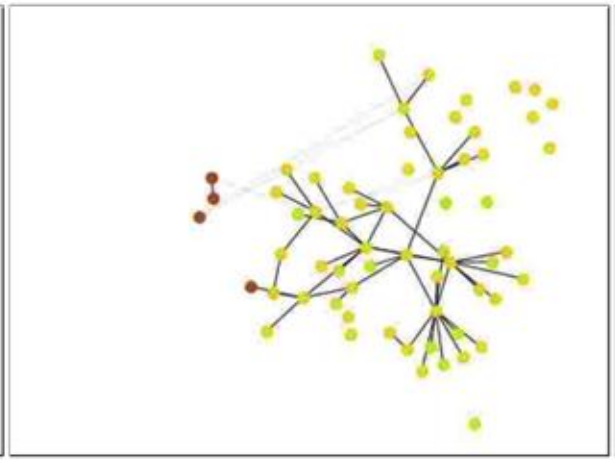
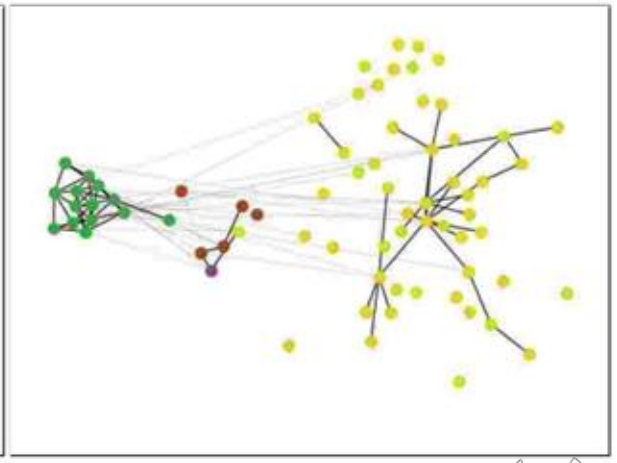
- These graphs are clustered to identify senses
- **The degree of sense gain/loss** between different periods/domains
  = **The degree of semantic change**



Word Usage Graph [Schlechtweg+2021]

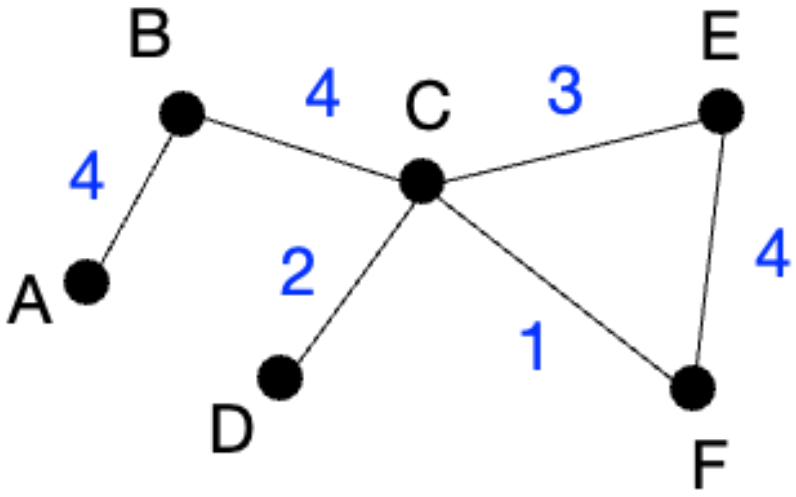
- [illegible]es
- [illegible]

How can we model semantic change from word usage graphs directly?

Proposal

- Represent the word usage graph as an adjacency matrix (A-C: Doc1, D-F: Doc2)
- **Model the usage similarity matrix directly**



$$
\begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \end{array}
\begin{array}{c}
\begin{array}{cccccc} A & B & C & D & E & F \end{array} \\
\begin{pmatrix}
 & 4 & & & & \\
4 & & 4 & & & \\
 & 4 & & 2 & 3 & 1 \\
 & & 2 & & & \\
 & & 3 & & & 4 \\
 & & 1 & & 4 &
\end{pmatrix}
\end{array}
$$

Proposal

- If the meaning of the word is **stable**
- All similarity scores **X** are generated from **the same distribution**

$$
\begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \end{array}
\begin{array}{c}
\begin{array}{cccccc} A & B & C & D & E & F \end{array} \\
\begin{pmatrix}
4 & 2 & 3 & 4 & 1 & 2 \\
2 & 4 & 3 & 2 & 4 & 4 \\
3 & 3 & 4 & 2 & 3 & 1 \\
4 & 2 & 2 & 4 & 3 & 2 \\
1 & 4 & 3 & 3 & 4 & 1 \\
2 & 4 & 1 & 2 & 1 & 4
\end{pmatrix}
\end{array}
= \mathbf{X}
$$

Proposal

- If the meaning of the word is **changed**
- All similarity scores **X1~X4** are generated from **the distinct distributions**

$$
\begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \end{array}
\begin{array}{c}
\begin{array}{cccccc} A & B & C & D & E & F \end{array} \\
\left(\begin{array}{cccc|cc}
4 & 4 & 4 & 1 & 1 & 1 \\
4 & 4 & 4 & 1 & 1 & 1 \\
4 & 4 & 4 & 1 & 1 & 1 \\
\hline
1 & 1 & 1 & 4 & 4 & 4 \\
1 & 1 & 1 & 4 & 4 & 4 \\
1 & 1 & 1 & 4 & 4 & 4
\end{array}\right)
\end{array}
=
\begin{array}{c}
\begin{array}{cccccc} A & B & C & D & E & F \end{array} \\
\left(\begin{array}{cc|cc}
\multicolumn{2}{c|}{\multirow{3}{*}{$\mathbf{X}_1$}} & \multicolumn{2}{c}{\multirow{3}{*}{$\mathbf{X}_2$}} \\
\multicolumn{2}{c|}{} & \multicolumn{2}{c}{} \\
\multicolumn{2}{c|}{} & \multicolumn{2}{c}{} \\
\hline
\multicolumn{2}{c|}{\multirow{3}{*}{$\mathbf{X}_3$}} & \multicolumn{2}{c}{\multirow{3}{*}{$\mathbf{X}_4$}} \\
\multicolumn{2}{c|}{} & \multicolumn{2}{c}{} \\
\multicolumn{2}{c|}{} & \multicolumn{2}{c}{}
\end{array}\right)
\end{array}
\begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \end{array}
$$

Proposal

- Our goal: estimate $p(\theta|\mathbf{X})$ from similarity matrix $\mathbf{X}$ ($\theta \in \{0, 1\}$: semantic change indicator)

$$p(\theta|\mathbf{X}) \propto p(\mathbf{X}|\theta)p(\theta) \propto \begin{cases} p(\mathbf{X}|\theta=0) \\ p(\mathbf{X}|\theta=1) \end{cases}$$

Bayes' theorem

- To predict semantic change, we can compare

$$\begin{cases} p(\mathbf{X}) = \int p(\mathbf{X}|\mathbf{p})p(\mathbf{p})d\mathbf{p} = \dfrac{\Gamma(\sum_k \alpha_k)}{\Gamma(L+\sum_k \alpha_k)} \prod_{k=1}^{K} \dfrac{\Gamma(\alpha_k+n_k)}{\Gamma(\alpha_k)} \\ p(\mathbf{X}|\theta=0) = p(\mathbf{X}), \quad p(\mathbf{X}|\theta=1) = \prod_{n=1}^{4} p(\mathbf{X}_n) \end{cases}$$

Pólya distribution

Results

- Our method **outperforms** previous SotA approaches using word embeddings

Method	EN	DE	LA	SV
SGNS [11]	73.0	54.2	45.0	61.3
SGNS [12]	62.2	75.0	**70.0**	67.7
BERT [13]	70.3	75.0	55.0	74.2
Pólya (**ours**)	**76.1**	**80.0**	N/A	**88.6**

Results

Data	Language	Grouping 1	Grouping 2	Accuracy (%) MostFreq	Pólya
DWUG EN	EN	1810–1860	1960–2010	54.3	**76.1**
DWUG EN Resampled	EN	1810–1860	1960–2010	60.0	**80.0**
DWUG DE	DE	1800–1899	1946–1990	60.0	**80.0**
DWUG DE Resampled	DE	1800–1899	1946–1990	60.0	**73.3**
DiscoWUG	DE	1800–1899	1946–1990	51.0	**72.0**
RefWUG	DE	1750–1800	1850–1900	**54.5**	45.5
DURel	DE	1750–1800	1850–1900	**63.6**	**63.6**
SURel	DE	general	domain specific	63.6	**68.2**
RuSemShift 1	RU	1682–1916	1918–1990	**77.5**	**77.5**
RuSemShift 2	RU	1918–1990	1991–2016	**62.3**	**62.3**
RuShiftEval 1	RU	1700–1916	1918–1990	**74.8**	**74.8**
RuShiftEval 2	RU	1918–1990	1992–2016	**70.3**	**70.3**
RuShiftEval 3	RU	1700–1916	1992–2016	**68.5**	**68.5**
DWUG ES	ES	1810–1906	1994–2020	55.5	**78.0**
DiaWUG	ES	Spanish variant 1	Spanish variant 2	65.6	**81.3**
DWUG SV	SV	1790–1830	1895–1903	68.1	**88.6**
DWUG SV Resampled	SV	1790–1830	1895–1903	60.0	**73.3**
ChiWUG	ZH	1954–1978	1979–2003	**57.5**	52.5
DWUG IT	IT	1948–1970	1990–2014	**69.2**	N/A
DWUG LA	LA	−200–0	0–2000	**55.5**	N/A
NorDiaChange 1	NO	1929–1965	1970–2013	67.5	**75.0**
NorDiaChange 2	NO	1980–1990	2012–2019	**77.5**	70.0

Discussion

- Only 5→2% of the total entries work!