A Bayesian Model of Diachronic Meaning Change

Lea Frermann and Mirella Lapata

Institute for Language, Cognition, and Computation
School of Informatics
The University of Edinburgh

lea@frermann.de
www.frermann.de

ACL, August 09, 2016
Language is a dynamic system, constantly shaped by users and their environment.
Language is a dynamic system, constantly shaped by users and their environment.
The Dynamic Nature of Meaning

Language is a dynamic system, constantly shaped by users and their environment.

"nice"
- stupid
- lazy
- precise
- shy
- pleasant

"mouse"

Time: 1300-2000
The Dynamic Nature of Meaning

Language is a dynamic system, constantly shaped by users and their environment.

Meaning changes smoothly (in written language, across societies)
Motivation

Can we understand, model, and predict change?

• aid historical sociolinguistic research
• improve historical text mining and information retrieval

Can we build task-agnostic models?

• learn time-specific meaning representations which
• are interpretable and
• are useful across tasks
SCAN: A Dynamic Model of Sense Change
Model Assumptions

- target word (e.g., mouse)

- target word-specific corpus

<table>
<thead>
<tr>
<th>year</th>
<th>text snippet</th>
<th>text snippet</th>
</tr>
</thead>
<tbody>
<tr>
<td>1749</td>
<td>fortitude time woman shrieks</td>
<td>mouse rat capable poisoning husband</td>
</tr>
<tr>
<td>1915</td>
<td>rabbit lived hole small grey</td>
<td>mouse made nest pocket coat</td>
</tr>
<tr>
<td>1993</td>
<td>moved fire messages click computer</td>
<td>mouse communications appear electronic bulletin</td>
</tr>
<tr>
<td>2009</td>
<td>scooted chair clicking button wireless</td>
<td>mouse hibernate computer stealthy exit</td>
</tr>
</tbody>
</table>

- number of word senses ($K$)

- granularity of temporal intervals ($\Delta T$)
  (e.g., a year, decade, or century)
A **Bayesian** and **knowledge-lean** model of meaning change of individual words (e.g., “mouse”)
A Bayesian and knowledge-lean model of meaning change of individual words (e.g., “mouse”)
A Bayesian and knowledge-lean model of meaning change of individual words (e.g., “mouse”)

\[
p(k|t) \quad \text{and} \quad p(w|k,t)
\]
A Bayesian and knowledge-lean model of meaning change of individual words (e.g., “mouse”)

$p(k|t)$

$p(w|k,t)$
1. Extent of meaning change
   Generate temporal sense flexibility parameter $\kappa \phi \sim \Gamma(a, b)$

2. Time-specific representations
   Generate sense distributions $\phi_t$
   Generate sense-word distributions $\psi_{t,k}$

3. Document generation given time $t$
   Generate sense $z \sim \text{Mult}(\phi_t)$
   Generate context words $w_i \sim \text{Mult}(\psi_{t,k}, k = z)$
1. Extent of meaning change

\[ \kappa^\phi \sim \text{Gamma}(a, b) \]
Model Description: Generative Story

1. Extent of meaning change
   Generate temporal sense flexibility parameter
   \( \kappa^{\phi} \sim \text{Gamma}(a, b) \)

2. Time-specific representations
   Generate sense distributions \( \phi^{t} \)
   Generate sense-word distributions \( \psi_{k,t} \)
Model Description: Generative Story

1. Extent of meaning change
   - Generate temporal sense flexibility parameter
     \[ \kappa^{\phi} \sim \text{Gamma}(a, b) \]

2. Time-specific representations
   - Generate sense distributions \( \phi^t \)
   - Generate sense-word distributions \( \psi^{k,t} \)

3. Document generation given time \( t \)
Model Description: Generative Story

1. Extent of meaning change
   - Generate temporal sense flexibility parameter
     \[ \kappa^\phi \sim \text{Gamma}(a, b) \]

2. Time-specific representations
   - Generate sense distributions \( \phi^t \)
   - Generate sense-word distributions \( \psi^{k,t} \)

3. Document generation given time \( t \)
   - Generate sense
     \[ z \sim \text{Mult}(\phi^t) \]
Model Description: Generative Story

1. Extent of meaning change
   - Generate temporal sense flexibility parameter
     \( \kappa \phi \sim \text{Gamma}(a, b) \)

2. Time-specific representations
   - Generate sense distributions \( \phi^t \)
   - Generate sense-word distributions \( \psi^{k,t} \)

3. Document generation given time \( t \)
   - Generate sense \( z \sim \text{Mult}(\phi^t) \)
   - Generate context words \( w_i \sim \text{Mult}(\psi^{t,k=z}) \)
First-order random walk model
intrinsic Gaussian Markov Random Field (Rue, 2005; Mimno, 2009)

draw local changes from a normal distribution
mean temporally neighboring parameters
variance meaning flexibility parameter $\kappa_{\phi}$
Blocked Gibbs sampling

Details in the paper...
Related Work
## Word meaning change


<table>
<thead>
<tr>
<th>Word</th>
<th>Neighboring Words in 1900</th>
<th>Neighboring Words in 2009</th>
</tr>
</thead>
<tbody>
<tr>
<td>gay</td>
<td>cheerful, pleasant, brilliant</td>
<td>lesbian, bisexual, lesbians</td>
</tr>
</tbody>
</table>

- ✗ word-level meaning
- ✗ two time intervals
- ✗ representations are independent
- ✓ knowledge-lean
Related work

Word meaning change


<table>
<thead>
<tr>
<th>Word</th>
<th>Neighboring Words in 1900</th>
<th>Neighboring Words in 2009</th>
</tr>
</thead>
<tbody>
<tr>
<td>gay</td>
<td>cheerful pleasant brilliant</td>
<td>lesbian bisexual lesbians</td>
</tr>
</tbody>
</table>

- ✓ word-level meaning
- ✓ two time intervals
- ✓ representations are independent
- ✓ knowledge-lean

Graph-based tracking of word sense change

Mitra (2014, 2015)

- ✓ sense-level meaning
- ✓ multiple time intervals
- ✓ representations are independent
- ✓ knowledge-heavy
Evaluation
Evaluation: Overview

- no gold standard test set or benchmark corpora
- small-scale evaluation with hand-picked test examples

DATE: Diachronic Text Corpus (years 1710 – 2010)

1. COHA Corpus (Davies, 2010)
2. SemEval DTE Task Training Data (Popescu, 2015)
3. parts of the CLMET3.0 corpus (Diller, 2011)
Evaluation: Overview

- no gold standard test set or benchmark corpora
- small-scale evaluation with hand-picked test examples

We evaluate on various previously proposed tasks and metrics

1. qualitative evaluation
2. perceived word novelty (Gulordava, 2011)
3. temporal text classification SemEval DTE (Popescu, 2015)
4. usefulness of temporal dynamics
5. novel word sense detection (Mitra, 2014)
1. Qualitative Evaluation

- **love** power **life** time woman heart **god** tell little day
- **mind** power time **life** friend woman nature **love** world reason
- power country **government** nation war increase world **political** people europe
- power time **company** water force line **electric** plant day run
1. Qualitative Evaluation
2. Human-perceived Word Meaning Change (Gulordava (2011))

Task: Rank 100 target words by meaning change.

How much did \{ \textit{baseball}, \textit{network} \ldots \} change between the 1960s and the 1990s?

4-point scale 0: no change ... 3: significant change
2. Human-perceived Word Meaning Change (Gulordava (2011))

**Task:** Rank 100 target words by meaning change.

How much did \( \begin{align*} \text{baseball} \\
\text{network} \\
\ldots \end{align*} \) change between the 1960s and the 1990s?

4-point scale 0: no change ... 3: significant change

**Gulordava (2011)’s system**

- Compute word vectors from time-specific corpora (shared space):  
  \( w^{1960}, w^{1990} \)
- Compute \( \text{cosine}(w^{1960}, w^{1990}) \)
- Rank words by cosine: greater angle \( \rightarrow \) greater meaning change
2. Human-perceived Word Meaning Change (Gulordava (2011))

**Task:** Rank 100 target words by meaning change.

How much did \( \begin{align*}
    \text{baseball} \\
    \text{network} \\
    \text{...}
\end{align*} \) change between the 1960s and the 1990s?

4-point scale 0: no change ... 3: significant change

![Bar chart showing Spearman's \( \rho \) values for different methods.]

**Task:** predict the time frame of origin of a given text snippet

*President de Gaulle favors an independent European nuclear striking force [...] (1962)*

Prediction granularity

<table>
<thead>
<tr>
<th>Granularity</th>
<th>Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>medium</td>
<td>6-year intervals</td>
</tr>
<tr>
<td>coarse</td>
<td>12-year intervals</td>
</tr>
</tbody>
</table>

**SCAN** temporal word representations

- 883 nouns and verbs from the DTE development dataset
- $\Delta T = 5$ years
- $K = 8$ senses

→ predict time of a test snippet using **SCAN** representations

accuracy: precision measure discounted by distance from true time
A dynamic Bayesian model of diachronic meaning change

• sense-level meaning change
• arbitrary time spans and intervals
• knowledge lean
• explicit model of smooth temporal dynamics
Conclusions

A dynamic Bayesian model of diachronic meaning change

- sense-level meaning change
- arbitrary time spans and intervals
- knowledge lean
- explicit model of smooth temporal dynamics

Future Work

- learn the number of word senses (non-parametric)
- model short term opinion change from twitter data
Thank you!

lea@frermann.de
www.frermann.de
Blocked Gibbs sampling with three components

Block 1 Document sense assignments $\{z\}^D$

Block 2 Time-specific sense prevalence parameters $\{\phi\}^T$
Time- and sense-specific word parameters $\{\psi\}^{T \times K}$

Block 3 Degree of temporal sense flexibility $\kappa^\phi$
Learning

Block 2 Word- / sense parameters \( \{\phi\}^T \) and \( \{\psi\}^{T \times K} \)

- Logistic Normal is not conjugate to Multinomial \( \rightarrow \) ugly math!
- auxiliary variable method (Mimno et al, 2008)
- resample each \( \phi^t_k \) (and \( \psi^{t,k}_w \)) from a weighted, bounded area

\[
\begin{align*}
\text{current labels (likhd)} & \quad \text{iGMRF (prior)} \\
\begin{cases}
z^{d,t} = k \\
z^{d,t} \neq k
\end{cases}
\end{align*}
\]
1. The **COHA Corpus** *(Davies, 2010)*
   - large collection of text from various genres
   - years 1810 – 2009
   - 142,587,656 words

2. The **SemEval DTE Task Training Data** *(Popescu, 2015)*
   - news text snippets
   - years 1700 – 2010
   - 124,771 words

3. **Parts of the CLMET3.0 corpus** *(Diller, 2011)*
   - texts of various genres from open online archives
   - use years 1710–1810
   - 4,531,505 words
2. Capturing Perceived Word Novelty (Gulordava, 2011)

**Task:** Given a word, predict its novelty in a focus time (1990s) compared to a reference time (1960s).

- orange → 0
- crisis → 2
- net → 3
- sleep → 0
- virus → 2
- program → 3

··· ··· ···
2. Capturing Perceived Word Novelty (Gulordava, 2011)

**Task:** Given a word, predict its novelty in a focus time (1990s) compared to a reference time (1960s).

**A gold test set of 100 target words**

- how much did w’s meaning change between the 1960s and 1990s?
- ratings on a 4-point scale
  [0=no change, ..., 3=change significantly]

<table>
<thead>
<tr>
<th>orange</th>
<th>crisis</th>
<th>net</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>sleep</td>
<td>virus</td>
<td>program</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

...     ...     ...     ...
2. Capturing Perceived Word Novelty (Gulordava, 2011)

Gulordava et al’s system

- vector space model
- data: the Google Books bigram corpus
- compute a novelty score based on similarity of word vectors
  low similarity $\rightarrow$ significant change

SCAN

- data: DATE subcorpus covering 1960 – 1999; $\Delta T = 10$, $K = 8$
- we measure word novelty using the relevance score (Cook, 2014)
  - compute sense novelty based on time-specific keyword probabilities
    (Kilgarriff, 2000)
  - word novelty = max sense novelty
# 2. Capturing Perceived Word Novelty (Gulordava, 2011)

## Performance

<table>
<thead>
<tr>
<th>system</th>
<th>corpus</th>
<th>Spearman’s $\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gulordava (2011)</td>
<td>Google</td>
<td>0.386</td>
</tr>
<tr>
<td>SCAN</td>
<td>DATE</td>
<td>0.377</td>
</tr>
<tr>
<td>SCAN- NOT</td>
<td>DATE</td>
<td>0.255</td>
</tr>
<tr>
<td>frequency baseline</td>
<td>DATE</td>
<td>0.325</td>
</tr>
</tbody>
</table>

**SCAN predictions:** Most novel words w/ most novel sense (1960s vs 1990s)

<table>
<thead>
<tr>
<th>environmental users</th>
<th>supra note law protection id agency impact policy factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>virtual disk</td>
<td>computer window information software system wireless web</td>
</tr>
<tr>
<td></td>
<td>reality virtual computer center experience week community</td>
</tr>
<tr>
<td></td>
<td>hard disk drive program computer file store ram business</td>
</tr>
</tbody>
</table>

**Task:** predict the time frame of origin of a given text snippet

**subtask 1 – explicit cues**

*President de Gaulle favors an independent European nuclear striking force [...] (1962)*

**Prediction granularity**

<table>
<thead>
<tr>
<th>Type</th>
<th>Duration</th>
<th>Range</th>
</tr>
</thead>
</table>
Task: predict the time frame of origin of a given text snippet

subtask 2 – implicit (language) cues

*The local wheat market was not quite so strong to-day as yesterday.* (1891)

Prediction granularity

<table>
<thead>
<tr>
<th>Granularity</th>
<th>Time Frame</th>
</tr>
</thead>
<tbody>
<tr>
<td>fine 6-year</td>
<td>{1699–1705, 1706–1712, ..., 1888–1894, ..., 2007–2013}</td>
</tr>
<tr>
<td>coarse 20-year</td>
<td>{1692–1712, 1713–1733, ..., 1881–1901, ..., 2007–2027}</td>
</tr>
</tbody>
</table>
Scan

learn temporal word representations

- for all nouns and for all verbs that occur at least twice in the DTE development dataset (883 words)
- $\Delta T = 5\ \text{years}, \ K = 8$

**SCAN**

learn temporal word representations
- for all nouns and for all verbs that occur at least twice in the DTE development dataset (883 words)
- $\Delta T = 5$ years, $K = 8$

Predicting time of a test news snippet
1. Detect mentions of target words $\{c\}$; for each target
   1.1 construct document with $c$ and $\pm 5$ surrounding words $w$
   1.2 compute distribution over time slices:
      $$p^{(c)}(t|w) \propto p^{(c)}(w|t) \times p^{(c)}(t)$$
2. combine target-wise predictions into final distribution
3. predict time $t$ with highest probability

SCAN

learn temporal word representations

- for all nouns and for all verbs that occur at least twice in the DTE development dataset (883 words)
- $\Delta T = 5$ years, $K = 8$

Supervised Classification – Multiclass SVM

- SVM SCAN
  1. $\arg \max_k p^{(c)}(k|t)$ (most likely sense from SCAN models)
- SVM SCAN+n-gram
  1. $\arg \max_k p^{(c)}(k|t)$ (most likely sense from SCAN models)
  2. character n-grams
### 3. Diachronic Text Evaluation (DTE) (SemEval, 2015)

<table>
<thead>
<tr>
<th>Subtask 1 – factual cues</th>
<th>2 yr</th>
<th>6 yr</th>
<th>12 yr</th>
<th>6 yr</th>
<th>12 yr</th>
<th>20 yr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>.097</td>
<td>.214</td>
<td>.383</td>
<td>.199</td>
<td>.343</td>
<td>.499</td>
</tr>
<tr>
<td>SCAN-NOT</td>
<td>.265</td>
<td>.435</td>
<td>.609</td>
<td>.259</td>
<td>.403</td>
<td>.567</td>
</tr>
<tr>
<td>SCAN</td>
<td>.353</td>
<td>.569</td>
<td>.748</td>
<td>.376</td>
<td>.572</td>
<td>.719</td>
</tr>
<tr>
<td>IXA</td>
<td>.187</td>
<td>.375</td>
<td>.557</td>
<td>.261</td>
<td>.428</td>
<td>.622</td>
</tr>
<tr>
<td>AMBRA</td>
<td>.167</td>
<td>.367</td>
<td>.554</td>
<td>.605</td>
<td>.767</td>
<td>.868</td>
</tr>
<tr>
<td>UCD</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>.759</td>
<td>.846</td>
<td>.910</td>
</tr>
<tr>
<td>SVM SCAN</td>
<td>.192</td>
<td>.417</td>
<td>.545</td>
<td>.573</td>
<td>.667</td>
<td>.790</td>
</tr>
<tr>
<td>SVM SCAN+ngram</td>
<td>.222</td>
<td>.467</td>
<td>.627</td>
<td>.747</td>
<td>.821</td>
<td>.897</td>
</tr>
</tbody>
</table>

**Scores:** *accuracy* – precision measure discounted by distance from true time
### 3. Diachronic Text Evaluation (DTE) (SemEval, 2015)

<table>
<thead>
<tr>
<th></th>
<th>2 yr</th>
<th>6 yr</th>
<th>12 yr</th>
<th>6 yr</th>
<th>12 yr</th>
<th>20 yr</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baseline</strong></td>
<td>.097</td>
<td>.214</td>
<td>.383</td>
<td>.199</td>
<td>.343</td>
<td>.499</td>
</tr>
<tr>
<td><strong>SCAN-NOT</strong></td>
<td>.265</td>
<td>.435</td>
<td>.609</td>
<td>.259</td>
<td>.403</td>
<td>.567</td>
</tr>
<tr>
<td><strong>SCAN</strong></td>
<td>.353</td>
<td>.569</td>
<td>.748</td>
<td>.376</td>
<td>.572</td>
<td>.719</td>
</tr>
<tr>
<td><strong>IXA</strong></td>
<td>.187</td>
<td>.375</td>
<td>.557</td>
<td>.261</td>
<td>.428</td>
<td>.622</td>
</tr>
<tr>
<td><strong>AMBRA</strong></td>
<td>.167</td>
<td>.367</td>
<td>.554</td>
<td>.605</td>
<td>.767</td>
<td>.868</td>
</tr>
<tr>
<td><strong>UCD</strong></td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>.759</td>
<td>.846</td>
<td>.910</td>
</tr>
<tr>
<td>SVM <strong>SCAN</strong></td>
<td>.192</td>
<td>.417</td>
<td>.545</td>
<td>.573</td>
<td>.667</td>
<td>.790</td>
</tr>
<tr>
<td>SVM <strong>SCAN+ngram</strong></td>
<td>.222</td>
<td>.467</td>
<td>.627</td>
<td>.747</td>
<td>.821</td>
<td>.897</td>
</tr>
</tbody>
</table>

**Scores:**  
*accuracy* – precision measure discounted by distance from true time
### Subtask 1

<table>
<thead>
<tr>
<th>Method</th>
<th>2 yr</th>
<th>6 yr</th>
<th>12 yr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.097</td>
<td>0.214</td>
<td>0.383</td>
</tr>
<tr>
<td>Scan−Not</td>
<td>0.265</td>
<td>0.435</td>
<td>0.609</td>
</tr>
<tr>
<td>Scan</td>
<td>0.353</td>
<td>0.569</td>
<td>0.748</td>
</tr>
<tr>
<td>IXA</td>
<td>0.187</td>
<td>0.375</td>
<td>0.557</td>
</tr>
<tr>
<td>AMBRA</td>
<td>0.167</td>
<td>0.367</td>
<td>0.554</td>
</tr>
<tr>
<td>UCD</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

### Subtask 2

<table>
<thead>
<tr>
<th>Method</th>
<th>6 yr</th>
<th>12 yr</th>
<th>20 yr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.199</td>
<td>0.343</td>
<td>0.499</td>
</tr>
<tr>
<td>Scan−Not</td>
<td>0.259</td>
<td>0.403</td>
<td>0.567</td>
</tr>
<tr>
<td>Scan</td>
<td>0.376</td>
<td>0.572</td>
<td>0.719</td>
</tr>
<tr>
<td>IXA</td>
<td>0.261</td>
<td>0.428</td>
<td>0.622</td>
</tr>
<tr>
<td>AMBRA</td>
<td>0.605</td>
<td>0.767</td>
<td>0.868</td>
</tr>
<tr>
<td>UCD</td>
<td>0.759</td>
<td>0.846</td>
<td>0.910</td>
</tr>
</tbody>
</table>

### SVM Scan

<table>
<thead>
<tr>
<th>Method</th>
<th>6 yr</th>
<th>12 yr</th>
<th>20 yr</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM Scan</td>
<td>0.573</td>
<td>0.667</td>
<td>0.790</td>
</tr>
<tr>
<td>SVM Scan+ngram</td>
<td>0.747</td>
<td>0.821</td>
<td>0.897</td>
</tr>
</tbody>
</table>

### Discussion

→ did we just use more data? (no)
→ our system is not application specific
→ use different systems for different DTE subtasks