A Bayesian Model of Diachronic Meaning Change

Lea Frermann and Mirella Lapata

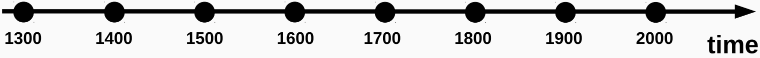
Institute for Language, Cognition, and Computation
School of Informatics
The University of Edinburgh

lea@frermann.de
www.frermann.de

ACL, August 09, 2016

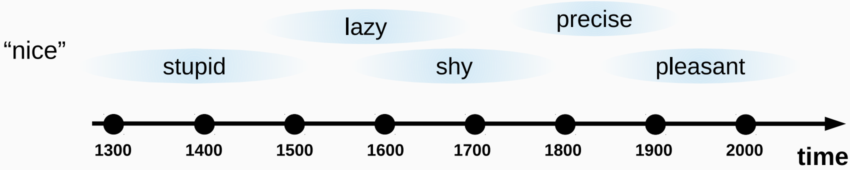
The Dynamic Nature of Meaning

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



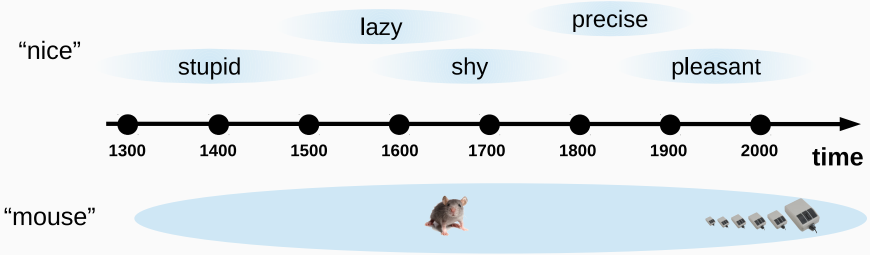
The Dynamic Nature of Meaning

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



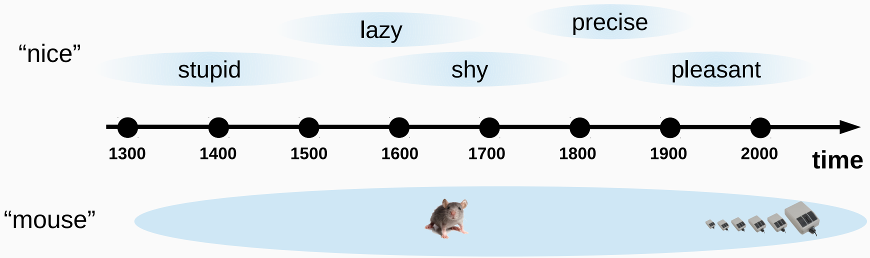
The Dynamic Nature of Meaning

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



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

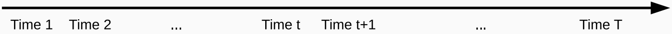
- target word (e.g., *mouse*)
- target word-specific corpus

year	text snippet		
1749	fortitude time woman shrieks	*mouse*	rat capable poisoning husband
1915	rabbit lived hole small grey	*mouse*	made nest pocket coat
1993	moved fire messages click computer	*mouse*	communications appear electronic bulletin
2009	scooted chair clicking button wireless	*mouse*	hibernate computer stealthy exit
	...		

- number of word senses (K)
- granularity of temporal intervals (ΔT)
(e.g., a year, decade, or century)

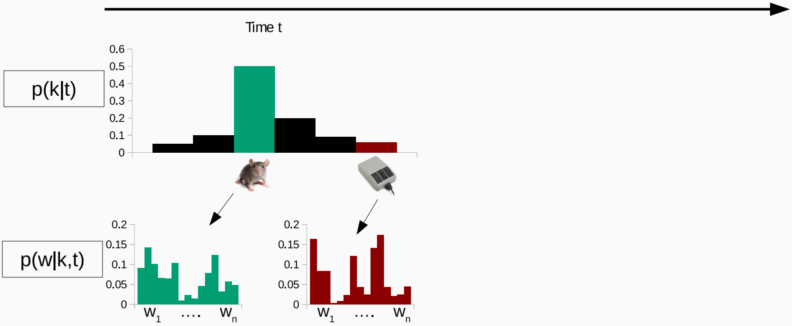
Model Overview

A **Bayesian** **and** **knowledge-lean** **model of meaning change of individual words** (e.g., "mouse")



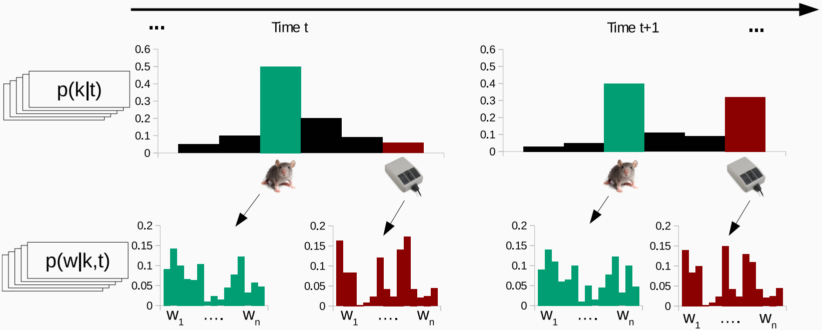
Model Overview

A Bayesian and knowledge-lean model of meaning change of individual words (e.g., “mouse”)

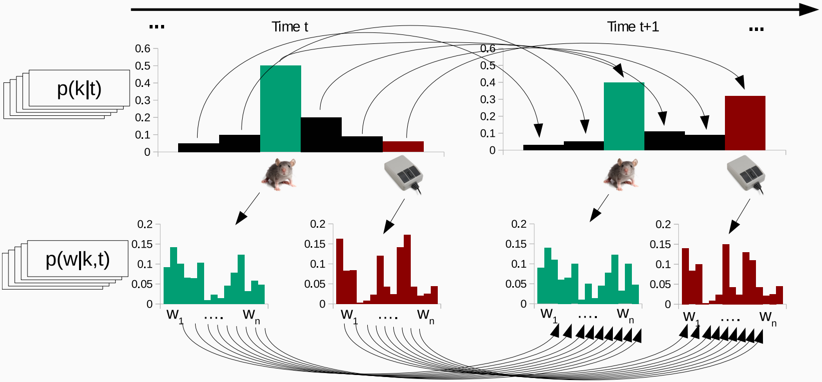


Model Overview

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")

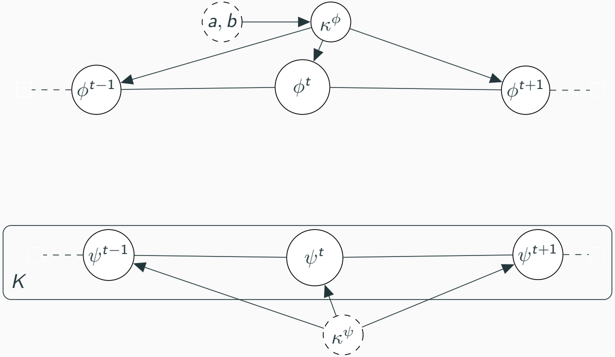


Model Description: Generative Story

Model Description: Generative Story

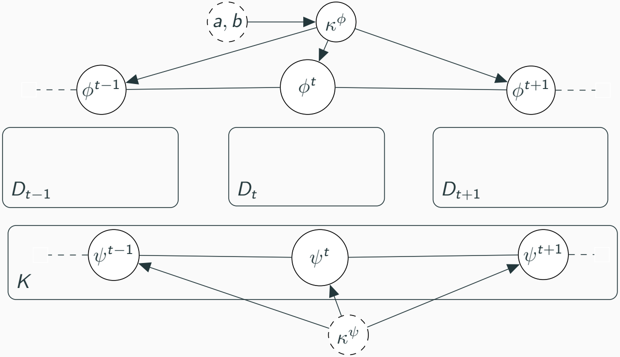
1. Extent of meaning change	Generate temporal sense flexibility parameter $\kappa^\phi \sim Gamma(a, b)$

Model Description: Generative Story



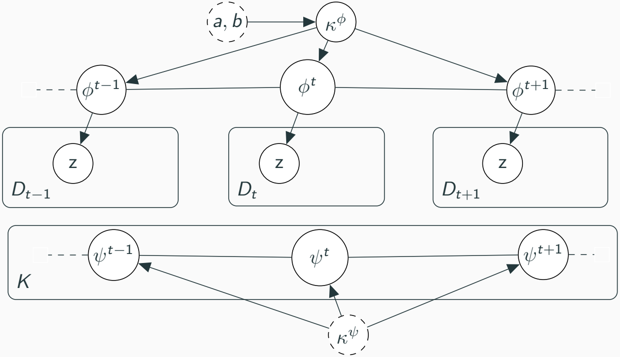
1. Extent of meaning change	Generate temporal sense flexibility parameter $\kappa^{\phi} \sim Gamma(a, b)$
2. Time-specific representations	Generate sense distributions ϕ^{t} Generate sense-word distributions $\psi^{k,t}$

Model Description: Generative Story



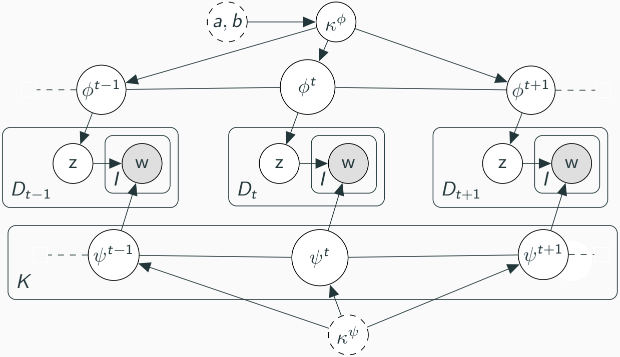
1. Extent of meaning change	Generate temporal sense flexibility parameter $\kappa^\phi \sim Gamma(a, b)$
2. Time-specific representations	Generate sense distributions ϕ^t Generate sense-word distributions $\psi^{k,t}$
3. Document generation given time t	

Model Description: Generative Story



1. Extent of meaning change	Generate temporal sense flexibility parameter $\kappa^\phi \sim \mathit{Gamma}(a, b)$
2. Time-specific representations	Generate sense distributions ϕ^t Generate sense-word distributions $\psi^{k,t}$
3. Document generation given time t	Generate sense $z \sim \mathit{Mult}(\phi^t)$

Model Description: Generative Story

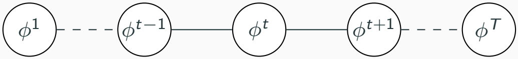


1. Extent of meaning change	Generate temporal sense flexibility parameter $\kappa^{\phi} \sim Gamma(a, b)$
2. Time-specific representations	Generate sense distributions ϕ^{t} Generate sense-word distributions $\psi^{k,t}$
3. Document generation given time t	Generate sense $z \sim Mult(\phi^{t})$ Generate context words $w_i \sim Mult(\psi^{t,k=z})$

SCAN: The Prior

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 κ^ϕ

Blocked Gibbs sampling

Details in the paper...

Related Work

Related work

Word meaning change

Gulordava (2011), Popescu (2013), Kim (2014) , Kulkarni (2015)

Word	Neighboring Words in	
	1900	2009
gay	*cheerful* *pleasant* *brilliant*	*lesbian* *bisexual* *lesbians*

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

Related work

Word meaning change

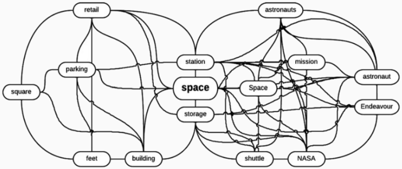
Gulordava (2011), Popescu (2013), Kim (2014) , Kulkarni (2015)

Word	Neighboring Words in	
	1900	2009
gay	*cheerful* *pleasant* *brilliant*	*lesbian* *bisexual* *lesbians*

- ✗ 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

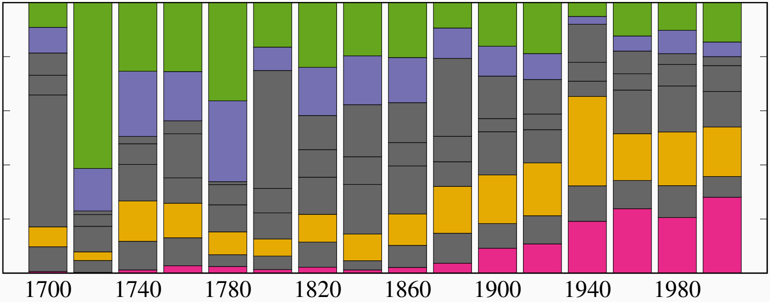
✗ small-scale evaluation with hand-picked test examples

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

power



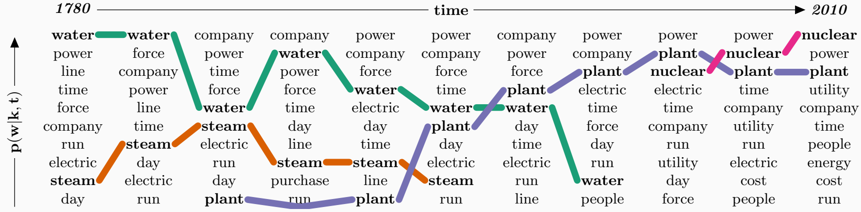
love power **life** time woman heart **god** tell little day

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 $\begin{cases} \textit{baseball} \\ \textit{network} \\ \dots \end{cases}$ 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 $\begin{cases} \textit{baseball} \\ \textit{network} \\ \ldots \end{cases}$ change between the 1960s and the 1990s?

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

Gulordava (2011)'s system

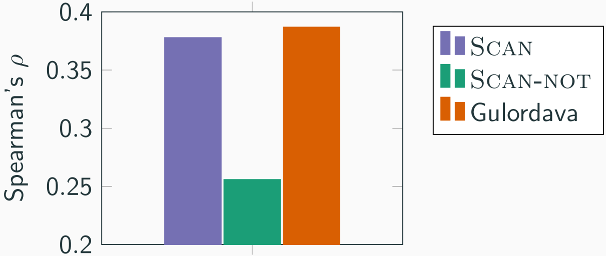
- Compute word vectors from time-specific corpora (shared space): w^{1960}, w^{1990}
- Compute $cosine(w^{1960}, w^{1990})$
- Rank words by cosine: greater angle $\rightarrow$ greater meaning change

2. Human-perceived Word Meaning Change (Gulordava (2011))

Task: Rank 100 target words by meaning change.

How much did { *baseball* / *network* / ... } change between the 1960s and the 1990s?

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



3. Diachronic Text Evaluation (DTE) (SemEval, 2015)

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

fine	2-year intervals	{1700–1702, ..., 1961–1963, ..., 2012–2014}
medium	6-year intervals	{1699–1706, ..., 1959–1965, ..., 2008–2014}
coarse	12-year intervals	{1696–1708, ..., 1956–1968, ..., 2008–2020}

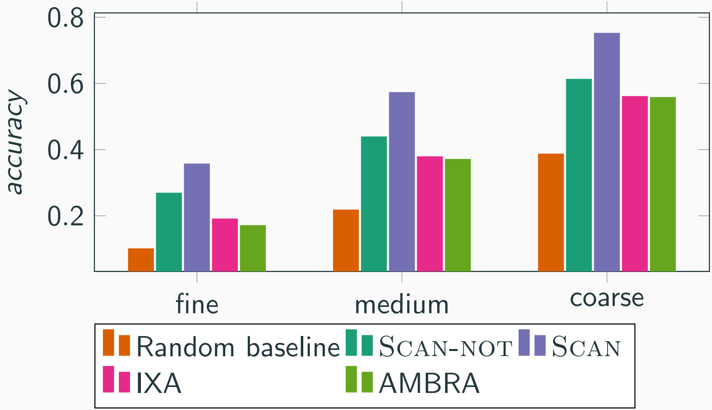
3. Diachronic Text Evaluation (DTE) (SemEval, 2015)

SCAN **temporal word representations**

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

$\rightarrow$ predict time of a test snippet using SCAN representations

3. Diachronic Text Evaluation (DTE) (SemEval, 2015)



accuracy: precision measure discounted by distance from true time

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

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@fremann.de
www.frermann.de

References I

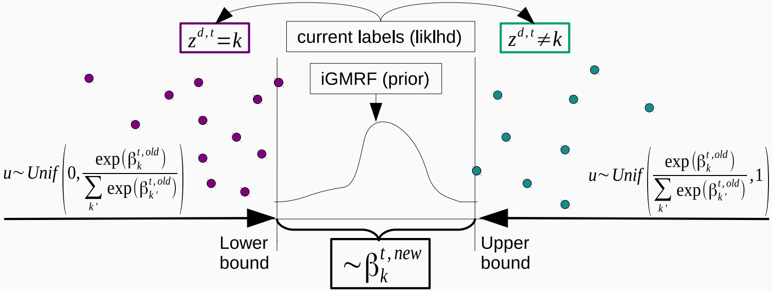
Learning

Blocked Gibbs sampling with three components

Block 1	Document sense assignments	$\{z\}^D$
Block 2	Time-specific sense prevalence parameters Time- and sense-specific word parameters	$\{\phi\}^T$ $\{\psi\}^{T \times K}$
Block 3	Degree of temporal sense flexibility	κ^ϕ

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 ϕ_k^t (and $\psi_w^{t,k}$) from a weighted, bounded area



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).

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]

orange $\rightarrow$ 0	crisis $\rightarrow$ 2	net $\rightarrow$ 3
sleep $\rightarrow$ 0	virus $\rightarrow$ 2	program $\rightarrow$ 3
...	...	...

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 → 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

Performance

system	corpus	Spearman's ρ
Gulordava (2011)	Google	0.386
SCAN	DATE	0.377
SCAN-NOT	DATE	0.255
frequency baseline	DATE	0.325

SCAN predictions: Most novel words w/ most novel sense

(1960s vs 1990s)

environmental	supra note law protection id agency impact policy factor
users	computer window information software system wireless web
virtual	reality virtual computer center experience week community
disk	hard disk drive program computer file store ram business

3. Diachronic Text Evaluation (DTE) (SemEval, 2015)

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

fine	2-year	{1700–1702, 1703–1705, ..., 1961–1963, ..., 2012–2014}
medium	6-year	{1699–1706, 1707–1713, ..., 1959–1965, ..., 2008–2014}
coarse	12-year	{1696–1708, 1709–1721, ..., 1956–1968, ..., 2008–2020}

3. Diachronic Text Evaluation (DTE) (SemEval, 2015)

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

fine	6-year	{1699–1705, 1706–1712, ..., 1888–1894, ..., 2007–2013}
medium	12-year	{1703–1715, 1716–1728, ..., 1885–1897, ..., 2002–2014}
coarse	20-year	{1692–1712, 1713–1733, ..., 1881–1901, ..., 2007–2027}

3. Diachronic Text Evaluation (DTE) (SemEval, 2015)

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 , $K = 8$

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 ± 5 surrounding words $\mathbf{w}$
   1.2 compute distribution over time slices :
   $$p^{(c)}(t|\mathbf{w}) \propto p^{(c)}(\mathbf{w}|t) \times p^{(c)}(t)$$
2. combine target-wise predictions into final distribution
3. predict time t with highest probability

3. Diachronic Text Evaluation (DTE) (SemEval, 2015)

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. $\arg\max_k \quad p^{(c)}(k|t)$ (most likely sense from SCAN models)
- SVM SCAN+n-gram
  1. $\arg\max_k \quad p^{(c)}(k|t)$ (most likely sense from SCAN models)
  2. character n-grams

3. Diachronic Text Evaluation (DTE) (SemEval, 2015)

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

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

3. Diachronic Text Evaluation (DTE) (SemEval, 2015)

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

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

3. Diachronic Text Evaluation (DTE) (SemEval, 2015)

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

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