Finding diachronic sense changes by Unsupervised methods

Arun Jayapal

Supervised by: Dr. Martin Emms

Trinity College Dublin

April 26, 2017

Semantic Neologism

semantic neologism: when an *old* word or multi-word acquires a *new* usage/meaning

Semantic Neologism

semantic neologism: when an *old* word or multi-word acquires a *new* usage/meaning

example *tweet*

Semantic Neologism

semantic neologism: when an *old* word or multi-word acquires a *new* usage/meaning

example *tweet*

old sense: refers to bird song, as in (from 1995)

> When a bird **tweets**, it's telling you what it is and where it is.

Semantic Neologism

semantic neologism: when an *old* word or multi-word acquires a *new* usage/meaning

example *tweet*

old sense: refers to bird song, as in (from 1995)

> When a bird **tweets**, it's telling you what it is and where it is.

recent sense: to post a status on twitter, as in (from 2009)

> An Embedded **Tweet** brings the best content created on Twitter into your article or website.

Example

gay being happy vs. homosexual

Example

gay being happy vs. homosexual

can make problems for SMT when its training data pre-dates the neologism's emergence

A sample mis-translation into Tamil via Google Translate for :

S1. With Clara, however, his brow cleared, and he was gay again (from 'sons and lovers' by D.H. Lawrence 1931:)

T1. கிளாரா ஆயினும் அவரது புருவம் அகற்றப்படும் மற்றும் அவர் மீண்டும் ஓரினச்சேர்க்கையாளர்

L1. Kiḷārā, āyiṉum, avaratu puruvam akaṟṟappaṭum, maṟṟum avar mīṇṭum ōriṉaccērkkaiyāḷar

The word *gay* is mis-translated as *ōriṉaccērkkaiyāḷar* (L1), which has the *homosexual* sense

Example

gay being happy vs. homosexual

can make problems for SMT when its training data pre-dates the neologism's emergence

A sample mis-translation into Tamil via Google Translate for :

S1. With Clara, however, his brow cleared, and he was gay again (from 'sons and lovers' by D.H. Lawrence 1931:)

T1. கிளாரா ஆயினும் அவரது புருவம் அகற்றப்படும் மற்றும் அவர் மீண்டும் ஓரினச்சேர்க்கையாளர்

L1. Kiḷārā, āyiṉum, avaratu puruvam akaṟṟappaṭum, maṟṟum avar mīṇṭum ōriṉaccērkkaiyāḷar

The word *gay* is mis-translated as *ōriṉaccērkkaiyāḷar* (L1), which has the *homosexual* sense

The question is:

Can semantic neologisms be detected from untagged text?

Representation and Notation

To talk about an occurrence of an ambiguous word will use:

$\mathbf{W}$:	words to left and right of a target
W_i:	i-th word in $\mathbf{W}$
Y:	year of occurrence
S:	sense of target occurrence of targets

Representation and Notation

To talk about an occurrence of an ambiguous word will use:

$\mathbf{W}$:	words to left and right of a target
W_i:	i-th word in $\mathbf{W}$
Y:	year of occurrence
S:	sense of target occurrence of targets

Eg. samples of **bricked**:

2001: *... In 1611 she was* **bricked** *into one of the rooms ...*
2011: *I've tried to flash a custom ROM and now I think I've* **bricked** *my phone*

become instances:

$$
\begin{array}{lll}
Y = 2001, & S = 1, & \mathbf{W} = \langle L, In, 1611, she, was, into, one, of, the, rooms \rangle \\
Y = 2011, & S = 2, & \mathbf{W} = \langle and, now, I, think, I've, my, phone, R, R, R \rangle
\end{array}
$$

Outline

Time dependent Sense Model

Without loss of generality, using the chain rule, we have

$$p(Y, S, \mathbf{W}) = p(Y) \times p(S|Y) \times p(\mathbf{W}|S, Y)$$

The $p(S|Y)$ term directly expresses the idea that the prevalence of a sense can vary with the year

Time dependent Sense Model

Without loss of generality, using the chain rule, we have

$$p(Y, S, \mathbf{W}) = p(Y) \times p(S|Y) \times p(\mathbf{W}|S, Y)$$

The $p(S|Y)$ term directly expresses the idea that the prevalence of a sense can vary with the year

If we now assume that $p(\mathbf{W}|S, Y) = p(\mathbf{W}|S)$ ie. **$\mathbf{W}$ is conditionally independent of Y given S** we get first line below

Definition (Dynamic Sense Model)

$$\begin{aligned} p(Y, S, \mathbf{W}) &= p(Y) \times p(S|Y) \times p(\mathbf{W}|S) && (1) \\ &= && (2) \end{aligned}$$

Time dependent Sense Model

Without loss of generality, using the chain rule, we have

$$p(Y, S, \mathbf{W}) = p(Y) \times p(S|Y) \times p(\mathbf{W}|S, Y)$$

The $p(S|Y)$ term directly expresses the idea that the prevalence of a sense can vary with the year

If we now assume that $p(\mathbf{W}|S, Y) = p(\mathbf{W}|S)$ ie. $\mathbf{W}$ is conditionally independent of Y given S we get first line below

Definition (Dynamic Sense Model)

$$\begin{aligned} p(Y, S, \mathbf{W}) &= p(Y) \times p(S|Y) \times p(\mathbf{W}|S) & (1) \\ &= p(Y) \times p(S|Y) \times \prod_i p(W_i|S) & (2) \end{aligned}$$

Second line above by treating $\mathbf{W}$ as 'bag of words'

Outline

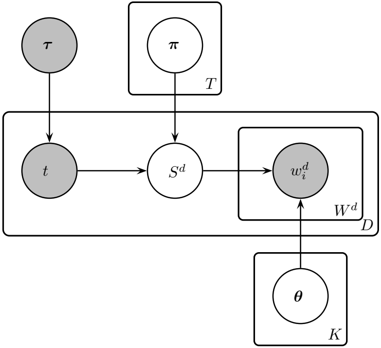
If we further assume that $p(S|Y) = p(S)$ we get:

Definition (Static Sense Model)

$$p(Y, S, \mathbf{W}) = p(Y) \times p(S) \times p(\mathbf{W}|S)$$

Outline

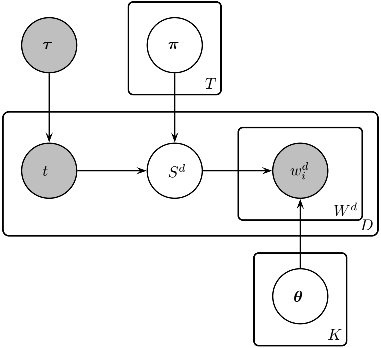
Generative story - EM



$\boldsymbol{\pi}_t$: parameter for senses (sense probs at t)
S^d : sense label for document d
w_i^d : word i in document d
W^d : words in document d
D : number of documents
K : total number of senses (classes)
$\boldsymbol{\theta}_k$: parameter for words (word probs at sense k)
$\boldsymbol{w}^{1:D}$: set of all documents
$\boldsymbol{\tau}$: probability for all time periods
τ_t : probability for a particular time t
V : all words in the vocabulary
T : total unique time instances

Figure: Sense emergence by EM - Graphical model

Generative story - EM



$\boldsymbol{\pi}_t$: parameter for senses (sense probs at t)
S^d : sense label for document d
w_i^d : word i in document d
W^d : words in document d
D : number of documents
K : total number of senses (classes)
$\boldsymbol{\theta}_k$: parameter for words (word probs at sense k)
$\boldsymbol{w}^{1:D}$: set of all documents
$\boldsymbol{\tau}$: probability for all time periods
τ_t : probability for a particular time t
V : all words in the vocabulary
T : total unique time instances

Figure: Sense emergence by EM - Graphical model

(1) Choose sense label S^d for a document from $p(S^d|Y_t;\boldsymbol{\pi})$
(2) For each word position w_i^d in the document
    Choose a word w from $p(\mathbf{W}|S^d;\boldsymbol{\theta})$

EM-training

let $\boldsymbol{\pi}_t$ be a vector of sense parameters at time t ie., $P(\mathbf{S}|Y=t)$
let $\boldsymbol{\theta}_k$ be a vector of word parameters with sense k ie., $P(\mathbf{W}|S=k)$ and
let $\boldsymbol{\tau}$ be a vector of year probabilities ie., $p(\mathbf{Y})$

Now $P(Y, S, \mathbf{W})$ can be expressed as

$$p(d; \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\tau}) = p(Y_t; \boldsymbol{\tau}) \times p(S_d|Y_t; \boldsymbol{\pi}_t) \times \prod_i p(w_i^d|S_d; \boldsymbol{\theta}_{1:K})$$

$$p(d; \boldsymbol{\Theta}) = \tau_t \times \pi_{t,k} \times \prod_{i=1}^{W^d} \theta_{k,i}$$

EM-training

let $\boldsymbol{\pi}_t$ be a vector of sense parameters at time t ie., $P(\mathbf{S}|Y = t)$
let $\boldsymbol{\theta}_k$ be a vector of word parameters with sense k ie., $P(\mathbf{W}|S = k)$ and
let $\boldsymbol{\tau}$ be a vector of year probabilities ie., $p(\mathbf{Y})$

Now $P(Y, S, \mathbf{W})$ can be expressed as

$$p(d; \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\tau}) = p(Y_t; \boldsymbol{\tau}) \times p(S_d|Y_t; \boldsymbol{\pi}_t) \times \prod_i p(w_i^d|S_d; \boldsymbol{\theta}_{1:K})$$

$$p(d; \boldsymbol{\Theta}) = \tau_t \times \pi_{t,k} \times \prod_{i=1}^{W^d} \theta_{k,i}$$

Data has no sense annotation. So use EM to make converging sequence of estimates

The training algorithm will consist of iterations to get the new estimates for $\Theta_n(\tau, \pi, \boldsymbol{\theta})$ and map them to $\Theta_{n+1}(\tau, \pi, \boldsymbol{\theta})$ by an **E-step** followed by **M-step**.

EM-updates

(E-step) Compute the completions for each of the training instances $(Y^d, \mathbf{W}^d)$ of the dataset $\mathbf{D}$ with a sense $(Y^d, S, \mathbf{W}^d)$, by computing the conditional probability $P(S|Y = t, \mathbf{W}^d)$ under the current estimate $\Theta_n(\boldsymbol{\tau}, \boldsymbol{\pi}, \boldsymbol{\theta})$.

EM-updates

(E-step) Compute the completions for each of the training instances $(Y^d, \mathbf{W}^d)$ of the dataset $\mathbf{D}$ with a sense $(Y^d, S, \mathbf{W}^d)$, by computing the conditional probability $P(S|Y = t, \mathbf{W}^d)$ under the current estimate $\Theta_n(\boldsymbol{\tau}, \boldsymbol{\pi}, \boldsymbol{\theta})$.

(M-step) Use maximum likelihood estimate to derive the new estimates for $\Theta_{n+1}(\tau_t, \boldsymbol{\pi}_t, \boldsymbol{\theta}_k)$.

EM-updates

(E-step) Compute the completions for each of the training instances $(Y^d, \mathbf{W}^d)$ of the dataset **D** with a sense $(Y^d, S, \mathbf{W}^d)$, by computing the conditional probability $P(S|Y = t, \mathbf{W}^d)$ under the current estimate $\Theta_n(\boldsymbol{\tau}, \boldsymbol{\pi}, \boldsymbol{\theta})$.

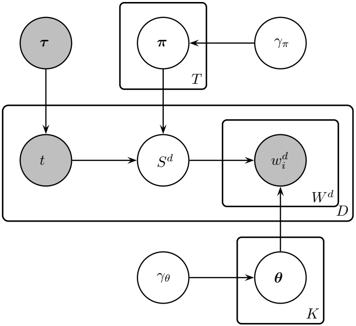
(M-step) Use maximum likelihood estimate to derive the new estimates for $\Theta_{n+1}(\tau_t, \boldsymbol{\pi}_t, \boldsymbol{\theta}_k)$.

Compute the conditional probabilities $\gamma[d][k_i] = p(S = k_i|Y = t^d, \mathbf{W} = \mathbf{W}^d)$ for each sense of $\mathbf{K} = \{k_1, k_2, \ldots, k_n\}$, in dataset D. The updates are:

$$p(S = k_i|Y = t; \Theta_{n+1}) = \frac{\sum_d (\text{if } Y^d = t \text{ then } \gamma[d][k_i] \text{ else } 0)}{\sum_d (\text{if } Y^d = t \text{ then } 1 \text{ else } 0)}$$

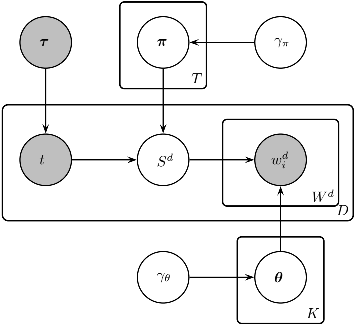
$$p(w|S = k_i; \Theta_{n+1}) = \frac{\sum_d (\gamma[d][k_i] \times freq(w \in \mathbf{W}^d))}{\sum_d (length(\mathbf{W}^d))}$$

Outline



γ_π : hyperparam of dirichlet distribution to sample for π
γ_θ : hyperparam of dirichlet distribution to sample for θ
$\boldsymbol{\pi}_t$: parameter for senses (sense probs at t)
S^d : sense label for document d
w_i^d : word i in document d
W^d : words in document d
D : number of documents
K : total number of senses (classes)
$\boldsymbol{\theta}_k$: parameter for words (word probs at sense k)
$\boldsymbol{w}^{1:D}$: set of all documents
$\boldsymbol{\tau}$: probability for all time periods
τ_t : probability for a particular time t
V : all words in the vocabulary
T : total unique time instances

Figure: Sense emergence by Gibbs sampling - Graphical model



γ_π : hyperparam of dirichlet distribution to sample for π
γ_θ : hyperparam of dirichlet distribution to sample for θ
$\boldsymbol{\pi}_t$: parameter for senses (sense probs at t)
S^d : sense label for document d
w_i^d : word i in document d
W^d : words in document d
D : number of documents
K : total number of senses (classes)
$\boldsymbol{\theta}_k$: parameter for words (word probs at sense k)
$\boldsymbol{w}^{1:D}$: set of all documents
$\boldsymbol{\tau}$: probability for all time periods
τ_t : probability for a particular time t
V : all words in the vocabulary
T : total unique time instances

Figure: Sense emergence by Gibbs sampling - Graphical model

(1) Choose $\boldsymbol{\pi} \sim \text{Dir}(\gamma_\pi)$
(2) Choose $\boldsymbol{\theta} \sim \text{Dir}(\gamma_\theta)$
(3) Choose sense label S^d from $p(S^d|t;\boldsymbol{\pi})$
(4) For each position w_i^d in the document d
Choose a word w from $p(w|S^d;\boldsymbol{\theta})$

Gibbs Sampling

Idea: Get a number of samples from the posterior distribution.

Gibbs Sampling

Idea: Get a number of samples from the posterior distribution.

Difference from EM:

1. Compute $\gamma[d][k_i] = p(S = k_i | Y = t^d, \mathbf{W} = \mathbf{W}^d)$ (as in EM) and sample for sense label from $\gamma[d][k_i]$

Gibbs Sampling

Idea: Get a number of samples from the posterior distribution.

Difference from EM:

1. Compute $\gamma[d][k_i] = p(S = k_i | Y = t^d, \mathbf{W} = \mathbf{W}^d)$ (as in EM) and sample for sense label from $\gamma[d][k_i]$
2. Sample for posteriors π and θ from Dirichlet distribution based on the document-sense counts and word-sense counts respectively.

Outline

Dataset

The data needed timestamp for a number of years

1. Google 5-gram dataset
2. Google timeline search

Dataset

Google 5-grams: Data set released by Google giving per-year counts of 5-grams from their digitized books holdings.

Dataset

Google 5-grams: Data set released by Google giving per-year counts of 5-grams from their digitized books holdings.

Google 5-gram datasets for the positive targets mouse, gay, strike, bit, compile, paste, surf, boot, rock, stoned, hip, export, mirror, domain, high and negative targets ostensible, cinema, present, promotion, theatre, play, spirit were considered for experiments using EM and Gibbs.

Dataset

Google 5-grams: Data set released by Google giving per-year counts of 5-grams from their digitized books holdings.

Google 5-gram datasets for the positive targets mouse, gay, strike, bit, compile, paste, surf, boot, rock, stoned, hip, export, mirror, domain, high and negative targets ostensible, cinema, present, promotion, theatre, play, spirit were considered for experiments using EM and Gibbs.

Here is a sample of dataset for *mouse* from 1821:

5-gram	count
diligence and patience the mouse	2
and patience the mouse ate	2
patience the mouse ate in	2
the mouse ate in two	2
mouse ate in two the	2

Google timeline search: Data (text snippets) collected by searching for target using Google time-line search feature.

Google timeline search: Data (text snippets) collected by searching for target using Google time-line search feature.

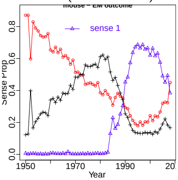
Google timeline search datasets for the following targets were tested: bricked, crawled, smashed it, me time, going forward, biological clock

Outline

Google 5-gram: Evaluation - by *introspection* comparing with tracks plot. For any word w let *track*(w) be the sequence of its *per-year probabilities* of occurrence in the 5-grams for a given target.

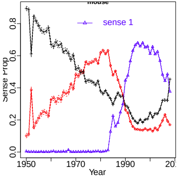
Neologism target: Mouse

mouse (EM outcome)



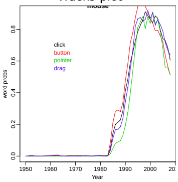
sense 1 words(:corp) *button, pointer, left, right, release, over, move, down, your, drag, you, hold, to, then, on, when, Release, cursor, use, clicking, click, Move, position, press, Click, while, changes, When, moving, user*

Gibbs outcome



sense 1 words(:corp) *button, pointer, left, click, right, you, over, release, your, down, move, to, drag, _START_, is, hold, use, when, then, Release, or, cursor, clicking, on, ,, Move, can, position, press, it*

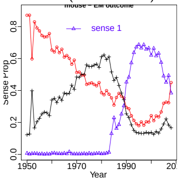
Tracks-plot



Google 5-gram: Evaluation - by *introspection* comparing with tracks plot. For any word w let *track*(w) be the sequence of its *per-year probabilities* of occurrence in the 5-grams for a given target.

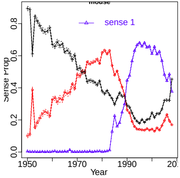
Neologism target: Mouse

mouse (EM outcome)



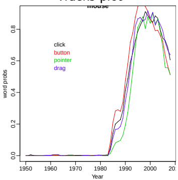
sense 1 words(:corp) *button, pointer, left, right, release, over, move, down, your, drag, you, hold, to, then, on, when, Release, cursor, use, clicking, click, Move, position, press, Click, while, changes, When, moving, user*

Gibbs outcome



sense 1 words(:corp) *button, pointer, left, click, right, you, over, release, your, down, move, to, drag, _START_, is, hold, use, when, then, Release, or, cursor, clicking, on, ,, Move, can, position, press, it*

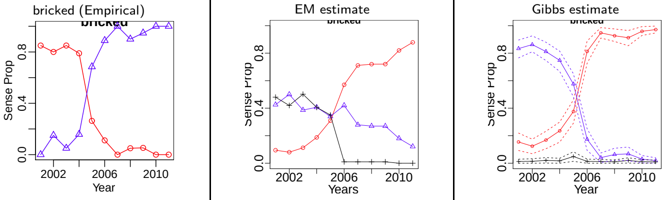
Tracks-plot



Additionally *Oxford English Dictionary* provides first citation date - used to verify

Google timeline search: Evaluation - by *introspection* comparing with empirical estimates.

Google timeline search: Evaluation - by *introspection* comparing with empirical estimates.



gist words(SENSE 0)-, ||, *L, How, ?, my, Forums, My, ·, think, Fix, Linksys, BlackBerry, PS3, iPhone, Apple, Is, Samsung, , iPod, Update, Please, Firmware, Unbrick, &, recover, Community, Android, Help, upgrade*

gist words(SENSE 0)-, *my, L, i, How*, ||, *I, ?, My, iPhone, your, PSP, Forums, psp, fix, < b > . . . < /b >, ", Fix, think, Linksys, phone, firmware, router, how, update, Firmware, do, it, BlackBerry, '*

The emerging sense from unsupervised estimates closely follow the empirical estimates.

Thank you