Learning Word Sense Distributions, Detecting Unattested Senses and Identifying Novel Senses Using Topic Models

Jey Han Lau[1] Paul Cook[2] Diana McCarthy[3]
Spandana Gella[2] Timothy Baldwin[2]

[1]Dept of Philosophy,
King's College London

[2]Dept of Computing and Information Systems,
The University of Melbourne

[3]University of Cambridge

June, 2014

Table of Contents

Table of Contents

Automatic Sense Adaptation

- Sense priors are known to vary considerably from corpus to corpus:
  - predominant/first sense preferences can be very different
  - certain senses may not be attested at all in a given corpus
  - there may be novel senses not documented in sense inventory
- Knowing the sense priors for a given corpus boosts WSD accuracy substantially
- **Aim:** given a sense inventory and an untagged corpus, automatically learn:
  1. the predominant sense for a given word
  2. the sense distribution for a given word
  3. what senses in the sense inventory aren't attested in the corpus
  4. what usages in the corpus aren't captured in the sense inventory

Example

- Target word = $cheat_V$;
- Domain = New York Times articles;
- Sense inventory = Macmillan; senses of $cheat_V$:

1. to behave dishonestly, or to not obey rules, for example in order to win a game or do well in an examination
2. to treat someone dishonestly
3. to have sex with someone who is not your husband, wife, or partner

Example

- Target word = $cheat_V$;
- Domain = New York Times articles;
- Sense inventory = Macmillan; senses of $cheat_V$:

Predominant sense

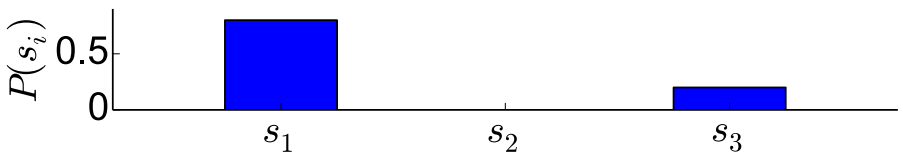
1. to behave dishonestly, or to not obey rules, for example in order to win a game or do well in an examination

Example

- Target word = $cheat_V$;
- Domain = New York Times articles;
- Sense inventory = Macmillan; senses of $cheat_V$:

Sense distribution

1. to behave dishonestly, or to not obey rules, for example in order to win a game or do well in an examination
2. to treat someone dishonestly
3. to have sex with someone who is not your husband, wife, or partner



Example

- Target word = $cheat_V$;
- Domain = New York Times articles;
- Sense inventory = Macmillan; senses of $cheat_V$:

Unattested senses

2. to treat someone dishonestly

Example

- Target word = $cheat_V$;
- Domain = New York Times articles;
- Sense inventory = Macmillan; senses of $cheat_V$:

Novel senses

5. avoid (something undesirable) by luck or skill, e.g. *cheated death*

Table of Contents

Introduction

- Our methodology builds on the **Word Sense Induction (WSI)** system we developed previously [Lau et al., 2012].
- WSI is the task of inducing the different **senses** or **meanings** of a target word.
- WSI is an unsupervised task: an unannotated text corpus is used for learning the senses.
- The core of the WSI system is driven by a Hierarchical Dirichlet Process (HDP), a non-parametric topic model [Teh et al., 2006].

HDP-WSI

- **Input:** collection of usages/sentences of a target word.
- **Output:**
  - HDP topics ($\leftrightarrow$ senses), each represented as a multinomial distribution over words;
  - Topic assignment in usages, each usage represented as a multinomial distribution over topics.
- Advantage of HDP: non-parametric method, meaning we do not need to pre-specify the number of senses.

Senses Induced for *cheat*

Sense	Top-N Terms
1	cheat think want ... love feel tell guy find
2	cheat student cheating test game school teacher exam study
3	husband wife cheat tiger on ... woman relationship
4	cheat woman relationship cheating partner reason man spouse
5	cheat game play player cheating poker card cheated money
6	cheat exchange china chinese foreign china team
7	tina bette kirk walk accuse mon pok symkyn nick star
8	fat jones ashley pen body taste weight expectation parent able
9	euro goal luck fair france irish single 2000 point complain

Induced WSI Topics vs. Inventory Senses

- We assign one topic to each usage by choosing its highest probability topic.
- This produces a distribution of topics over usages.
- In other words, it gives the **predominant topic**.
- The topic, however, does not have any direct relationship with the senses defined by sense inventories.
- We therefore require some way to align the topics with the senses.

Design Philosophy

- Methodology should be portable and applicable to any sense inventories.
- As such, our methodology assumes access to conventional sense gloss or definition only (i.e. no reliance on ontological/structural knowledge).

Computing Similarity Between Topic and Sense

Formally, similarity between sense s_i and topic t_j:

$$\text{sim}(s_i, t_j) = 1 - JS(S||T)$$

T: multinomial distribution over words for topic t_j;

S: multinomial distribution over words for sense s_i (converted from words in gloss and example based on MLE);

JS: Jensen Shannon divergence.

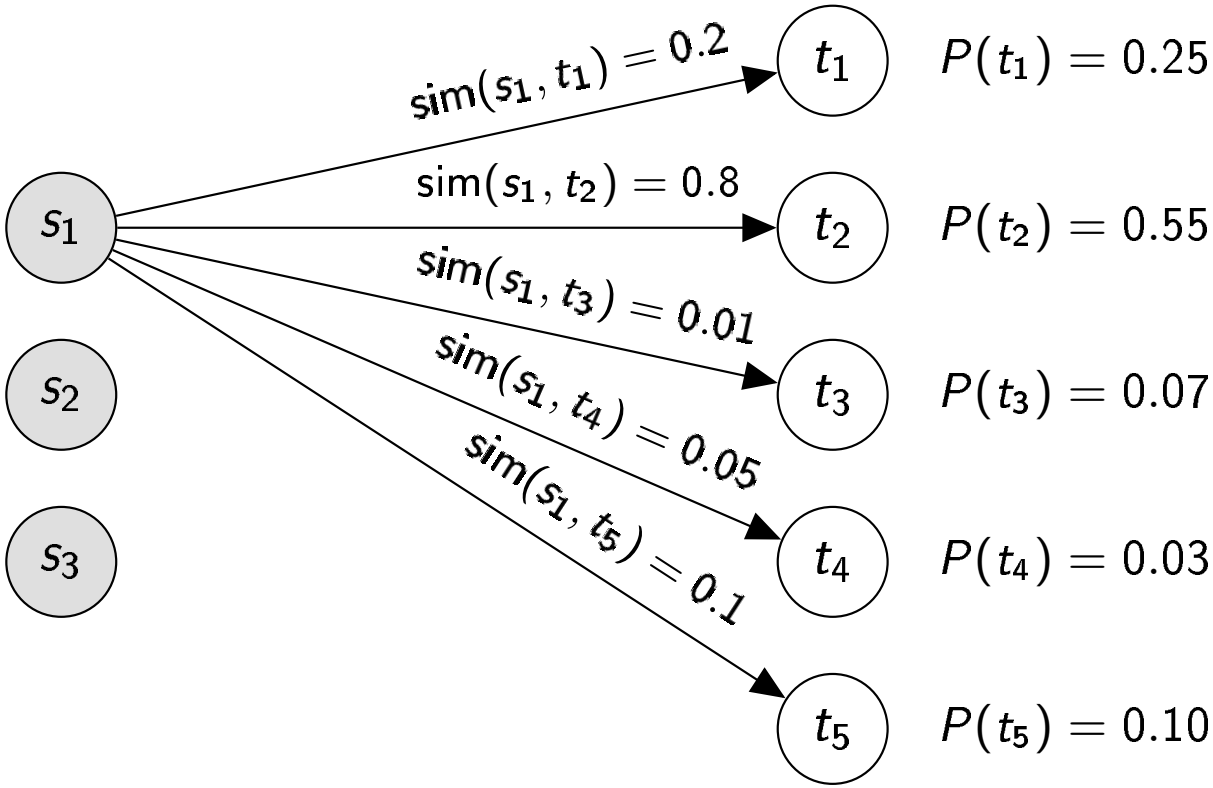
Finding Predominant Sense

To learn the predominant sense, we compute **prevalence score**, and take the sense with the highest prevalence score as the predominant sense.

The prevalence score for a sense s_i is the sum of the product of similarity scores and topic proportions:

$$\begin{aligned}\text{prevalence}(s_i) &= \sum_j^T (\text{sim}(s_i, t_j) \times P(t_j)) \\ &= \sum_j^T \left(\text{sim}(s_i, t_j) \times \frac{f(t_j)}{\sum_k^T f(t_k)} \right)\end{aligned}$$

Prevalence Score Example



$$\text{prevalence}(s_1) = (0.2 \times 0.25) + (0.8 \times 0.55) + (0.01 \times 0.07) + (0.05 \times 0.03) + (0.1 \times 0.1)$$

Table of Contents

State-of-the-Art

- McCarthy et al. [2004] proposed a method that uses the association of the target word with its nearest neighbours in an automatically acquired thesarus.
- Association is computed using WordNet similarity.
- Predominant sense is the highest ranked sense based on similarity scores.
- Similarity measures exploits WordNet hierarchy.

WordNet Dataset

- The authors developed a gold standard dataset for evaluating their methodology.
- Three domains were experimented: BNC, Reuters Sports and Reuters Finance.
- Usages of 40 target words were sense-annotated, using WordNet as the sense inventory.

Evaluation

- Acc: Word Sense Disambiguation (WSD) accuracy using predominant sense.
- FS_{corpus}/Acc_{ub}: Upper bound WSD accuracy using gold-standard predominant sense.
- ERR: Error rate reduction (Acc/Acc_{ub}).
- JS-Div: JS divergence between computed sense distribution and gold-standard sense distribution.

Results

Dataset	FS_{corpus} Acc_{ub}	MKWC Acc	MKWC ERR	HDP-WSI Acc	HDP-WSI ERR
BNC	0.524	**0.407**	(0.777)	0.376	(0.718)
FINANCE	0.801	0.499	(0.623)	**0.555**	(0.693)
SPORTS	0.774	**0.437**	(0.565)	0.422	(0.545)

Table: Predominant sense results (WSD Acc)

Dataset	MKWC	HDP-WSI
BNC	0.226	**0.214**
FINANCE	0.426	**0.375**
SPORTS	0.420	**0.363**

Table: Sense distribution results (JS-Div)

Findings

- Results fairly even: each outperforms the other at a level of statistical significance over one dataset.
- HDP-WSI is better at inducing overall sense distribution.
- MKWC uses full-text parsing in calculating distributional similarity thesaurus and WordNet graph structure in computing association.
- HDP-WSI uses no parsing (input is raw text), and only synset definitions of WordNet.

Table of Contents

The Macmillan dataset

- Gella et al. [2014] developed another sense-annotated dataset using the Macmillan dictionary as the sense inventory.
- 2 domains: ukWaC and Twitter; 20 target words.
- The Macmillan senses are coarser than WordNet senses (average polysemy in dataset = 5.6 vs. 12.3, resp.);
- We apply our methodology to the dataset to learn the predominant sense of the 20 target words.

Results

Dataset	FS_{corpus}	FS_{dict}		HDP-WSI	
	Acc_{ub}	Acc	ERR	Acc	ERR
ukWaC	0.574	0.387	(0.674)	**0.514**	(0.895)
Twitter	0.468	0.297	(0.635)	**0.335**	(0.716)

Table: Predominant sense results (WSD Acc)

Dataset	FS_{corpus}	FS_{dict}	HDP-WSI
ukWaC	0.210	0.393	**0.156**
Twitter	0.259	0.472	**0.171**

Table: Sense distribution results (JS-Div)

FS_{dict} = WSD Accuracy using the first-listed sense in Macmillan.

Extensions

- Our methodology does not just learn the predominant sense — it learns the overall sense distribution.

- Extensions:
  1. **Identification of unattested senses**: to find senses that are not used in the corpus;
  2. **Identification of novel senses**: to find novel senses that are not recorded in the sense inventory but seen in the corpus.

Identification of Novel Senses

Synthetic Data

- Task: Find usages/sentences of a novel sense that is not recorded by the sense inventory but seen in the data.
- Novel senses are synthesised by artificially removing an inventory sense.
- Three types of senses are removed: low, medium and high frequency senses.
- Only one sense is removed for each target word.

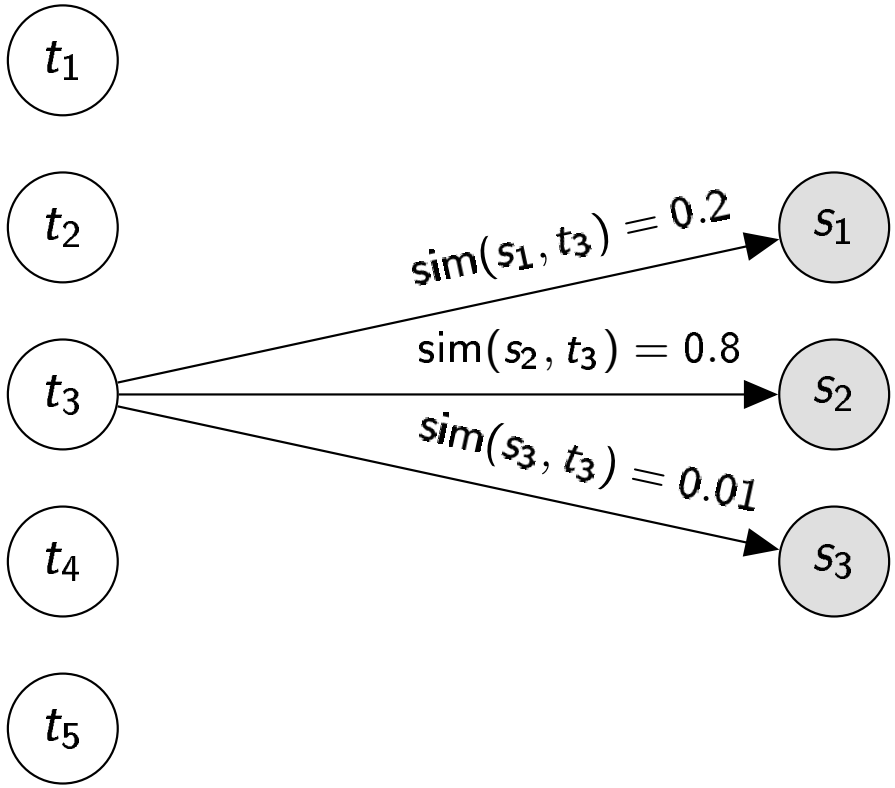
Experiment Setup

- Treat the task as a binary classification task: classify whether a sentence/usage contains a novel sense.
- Feature: topic-to-sense affinity score.
- Tune the threshold of this feature for separating the two classes with 10-fold cross validation.

$$\text{ts-affinity}(t_j) = \frac{\sum_i^S \text{sim}(s_i, t_j)}{\sum_l^T \sum_k^S \text{sim}(s_k, t_l)}$$

Intuition: a usage with a novel sense should have a topic that has low association with pre-existing senses.

Example



$$\text{ts-affinity}(t_3) = \frac{0.2 + 0.8 + 0.01}{\sum_{j=1}^{5} \text{ts-affinity}(t_j)}$$

Novel Sense Experiment: Results

No. Lemmas with a Removed Sense	Relative Freq of Removed Sense	P	R	F
20	0.0–0.2	0.35	0.42	0.36
9	0.2–0.4	0.50	0.66	0.52
6	0.4–0.6	0.73	0.90	0.80

- Usages with high frequency novel senses are more easily identifiable.
- Unsurprising as high frequency senses have a higher probability of generating related topics.

Conclusion

- We proposed a topic modelling-based method for estimating word sense distribution based on HDP.
- We evaluated the method to learn predominant senses and induce word sense distributions.
- The method is found to be comparable with a state-of-the-art system.
- We demonstrated the applicability of our method by proposing two new tasks that identify: (1) unattested senses; and (2) novel senses.

The End

Infel yor...

Questions?

S. Gella, P. Cook, and T. Baldwin. One sense per tweeter ... and other lexical semantic tales of Twitter. pages 215–220, Gothenburg, Sweden, 2014.

J.H. Lau, P. Cook, D. McCarthy, D. Newman, and T. Baldwin. Word sense induction for novel sense detection. pages 591–601, Avignon, France, 2012.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. Finding predominant senses in untagged text. In *Proceedings of the 42nd Annual Meeting of the Association of Computational Linguistics*, pages 280–287, Barcelona, Spain, 2004.

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101:1566–1581, 2006.