# Learning Context Selection Models for Knowledge-Based WSD


אוניברסיטת חיפה University of Haifa

Evgenia Wasserman Pritsker, Einat Minkov and William W. Cohen


## Background

***Word sense disambiguation (WSD):*** Determine which sense, out of known senses, is invoked in a given context, e.g., **STAR**:

***Supervised WSD:*** highly effective, but requires large amounts of sense-tagged examples. Learns contextual cues.

***Knowledge-based WSD:*** use lexico-semantic resources -- find the sense that agrees most with the given context. No annotated data required! Typically, considers all words surrounding the target word within a pre-defined window size.

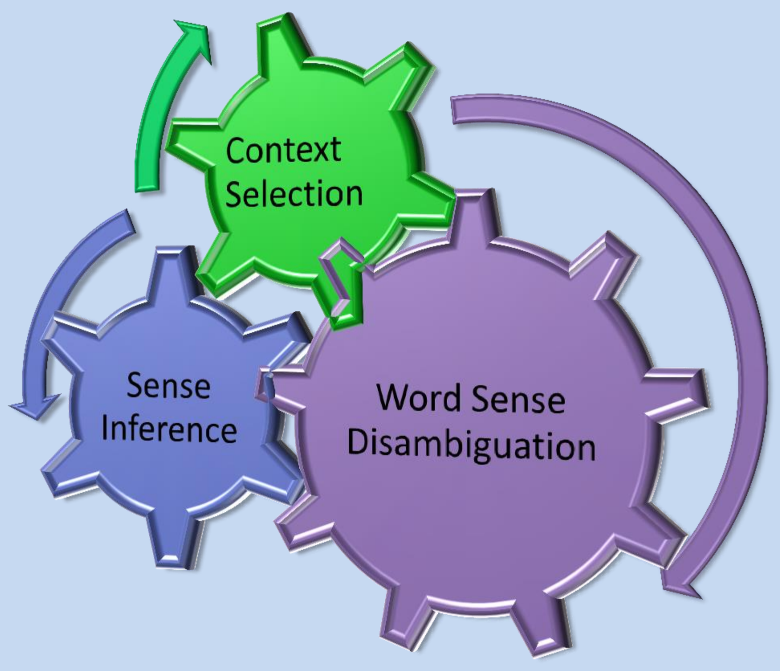
***Our goal: improve WSD performance by identifying <u>informative</u> contextual cues.***

*Any real **star** - which would never be perfectly **spherical** - could therefore only **collapse** to form a **naked singularity**.*

## Main contributions

- A learning framework for context selection.
- Statistically significant and consistent gains on benchmark WSD datasets.
- Performance comparable with, or exceeding, state-of-the-art.
- Found that lexico-statistical information (PMI, IDF) provides the strongest evidence (vs. traditional distance).

## Framework




- Learning framework that evaluates the reliability of each context word.

$$\arg\max_{s \in S(w)} \sum_{c_j \in Ctx} weight(c_j)\, Sim(s, c_j)$$

***S(w)*** – candidate senses
***Sim(.)*** - a similarity function
***Ctx*** – context represented as a bag-of-words

- Assign weights according to relevancy of the context word.
- In this work, boolean weight: unreliable context words ignored: *weight*($c_j$)*=0* ; otherwise: 1

## KB WSD Methods

- ***Lesk*** (1986) – *Sim(s,c)* = word overlap between the dictionary glosses of the context word and the candidate sense gloss.
- ***Gloss Vectors (GV)*** (Patwardhan Pedersen, `06) – extended Lesk (glosses extended with related synsets glosses and co-occurring words derived from row text).
- ***Personalized PageRank (PPR)*** (Agirre and Soroa, `09) – model WordNet as a graph; *Sim(s,c)* = graph-based similarity (random walks) between the nodes representing the senses of the context word and the target sense.

## Learning

- *Supervised*: dataset of context-target word pairs.
- Noisy labels: indicate whether sense prediction given the context word is correct.
- Unbalanced datasets (most examples are negative). Good results obtained using Naïve Bayes.

Explicit lexical information not encoded into the features. →
The learned models general rather than word-specific. →
The learned models fit within unsupervised KB WSD settings.

***Inference*** - rank available context words using the learned model; aggregate the predictions of the best ones.

## Context Features

- **Distance**: direct word distance
- **Syntactic**: target-context dependency relation path, path length, POS tag of context word
- **Word properties**: target-context PMI score; context word IDF score; context word number of senses

## Datasets

- Lexical sample due to Koeling *et al* ('05). Annotated examples for 41 nouns – 300 sentences each, extracted from domain-specific (sports/finance) and general (BNC) texts. 7 senses on average. Derived 121K target-context word pairs.
- All-words Senseval2&3.

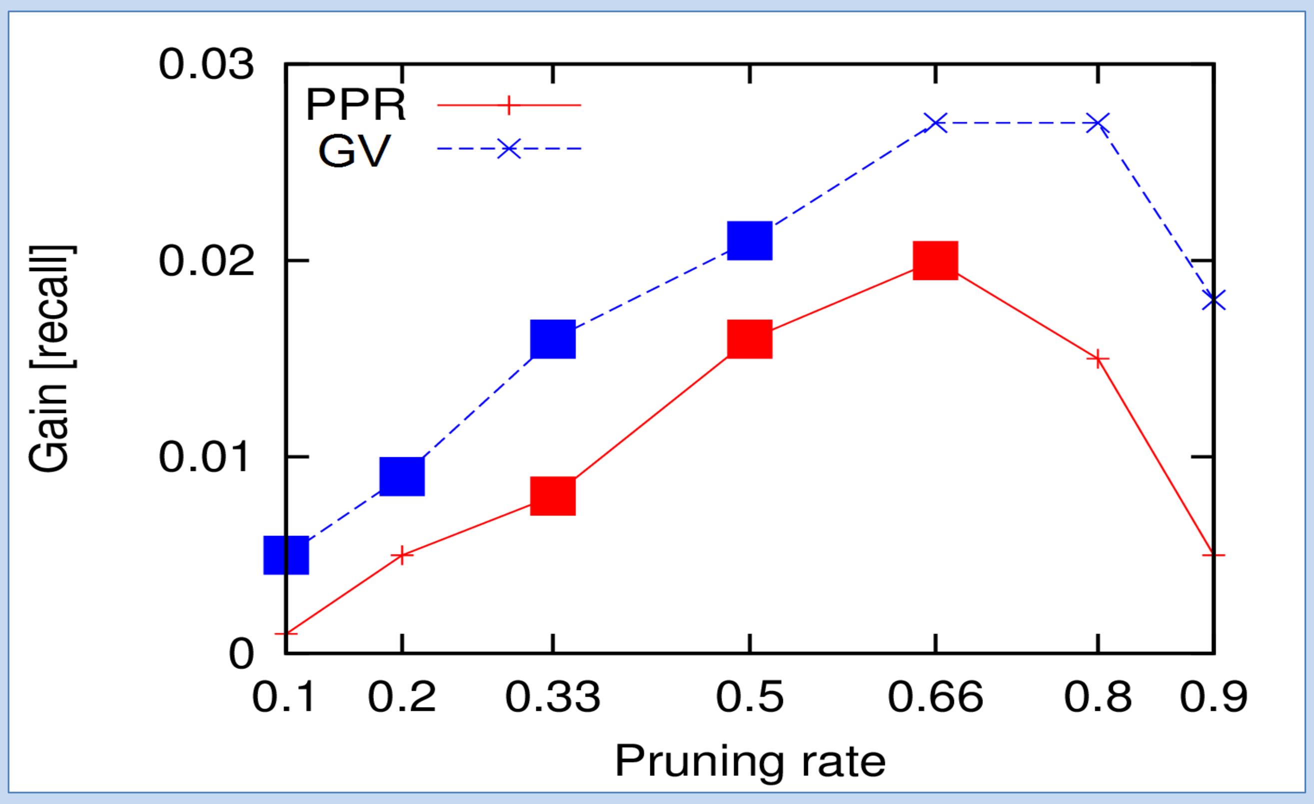
## Results

| Method | All | BNC | Sports | Finance |
|---|---|---|---|---|
| PPR | 0.49 | 0.49 | 0.44 | 0.55 |
| **Context Selection** | **0.51*** | **0.50** | **0.46** | **0.57** |
| Gloss Vectors | 0.39 | 0.38 | 0.36 | 0.42 |
| **Context Selection** | **0.41*** | **0.40** | **0.38** | **0.45** |

* Statistically significant results 50% top ranked words used

Koeling *et al* ('05) lexical sample dataset [recall]

## Robustness




Recall gains using different pruning rate.


**The research was funded by BSF.**