# Regular Polysemy: A Distributional Model

## by Boleda, Padó and Utt

Maximilian Fillinger

December 11, 2012

# Outline

# Regular Polysemy and Word Sense Disambiguation

- In Word Sense Disambiguation (WSD), polysemy is treated as word specific.
- However, there are regularities in polysemy.
- WSD can profit from making use of these regularities.

# Long-term Goal

A model of polysemy that

- allows unsupervised learning.
- extracts patterns that are applicable to multiple words.
- is able to generalize to previously unknown words.

# Contributions

- A general formal framework for characterizing polysemy
- A concrete model that instantiates this framework

# Outline

# Meta-Senses

The authors introduce some novel terminology:

- A “meta sense” is a general category or type of senses.
- In this paper, meta senses are CoreLex categories.
- Long-term goal: The model creates its own set of meta senses unsupervised.
- A “meta alternation” is a set of two meta senses.

# The Scoring Function

- The *score*-function takes as input a meta alternation and two word senses and outputs a real number.
- It measures how well the meta alternation corresponds to these two senses.

$$score(\{animal, food\}, \text{lamb}_{animal}, \text{lamb}_{food}) = \text{high}$$

$$score(\{agent, human\}, \text{lamb}_{animal}, \text{lamb}_{food}) = \text{low}$$

$$score(\{part, entity\}, \text{lamb}_{animal}, \text{lamb}_{food}) = ?$$

# The Scoring Function

- The *score*-function takes as input a meta alternation and two word senses and outputs a real number.
- It measures how well the meta alternation corresponds to these two senses.

$$score(\{animal, food\}, \text{lamb}_{animal}, \text{lamb}_{food}) = \text{high}$$

$$score(\{agent, human\}, \text{lamb}_{animal}, \text{lamb}_{food}) = \text{low}$$

$$score(\{part, entity\}, \text{lamb}_{animal}, \text{lamb}_{food}) = ?$$

# The Scoring Function

- The *score*-function takes as input a meta alternation and two word senses and outputs a real number.
- It measures how well the meta alternation corresponds to these two senses.

$$score(\{animal, food\}, \text{lamb}_{animal}, \text{lamb}_{food}) = \text{high}$$

$$score(\{agent, human\}, \text{lamb}_{animal}, \text{lamb}_{food}) = \text{low}$$

$$score(\{part, entity\}, \text{lamb}_{animal}, \text{lamb}_{food}) = ?$$

# Outline

# Representation as Vectors

- Words are represented by context vectors.
- A lemma is represented by the centroid of all its instances.
- A meta sense $m$ is represented by the centroid of all lemmata that *only* have meta sense $m$.
- A meta alternation $\{m_1, m_2\}$ is represented by the centroid of $m_1$ and $m_2$.

# The Scoring Function in CAM

- CAM has no direct representation of word senses.
- The scoring function is defined only for disemous lemmata, i.e., lemmata with exactly two senses.
- For a meta alternation $a$ and $s_1, s_2$ the two word senses of lemma $l$, let

$$score(a, s_1, s_2) = sim(a, l)$$

where *sim* is some similarity measure, e.g. the cosine.

# Outline

# Data

- The co-occurence vectors were extracted from the BNC.
- For each meta alternation that is instantiated often enough
  - 10 targets - disemous lemmata instantiating that alternation
  - and 30 distractors were chosen at random
- The model ranks the targets and distractors according to its scoring function.

# Evaluation Measure

- The success of the model for each alternation is measured with the average precision (AP) function.
- Let $p_1, ..., p_{40}$ be the ranking produced by the model.
- The AP-value is 1 if $p_1, ..., p_{10}$ are all targets and 0 if they are all distractors.

# Definition of AP

- Let $p_1, ..., p_{40}$ be the ranking of the targets and distractors.
- For $j = 1, ..., 10$, let $I(j) = 1$ if $p_j$ is a target.
- The value of AP is

$$\frac{1}{10} \sum_{j=1}^{10} I(j) \cdot \frac{\sum_{k=1}^{j} I(k)}{j}$$

# Outline

# Parameters

Various parameters of CAM have not been fixed yet. The authors have tested it with the following parameters:

- Three types of context vectors:
  - gram: Grammatical paths of length one to three
  - lex: Word co-occurence frequencies
  - gramlex: Combination of grammatical paths and word co-occurence
- With or without log-likelihood transformation
- Centroid computation for meta senses and alternations by micro-averaging or macro-averaging

# Results

- On average, the model achieves scores of at least 3.5.
- It significantly ($p < 0.01$) outperforms both the random baseline (0.313) and the frequency baseline (0.291)
- There is a high degree of variance.

# Summary

- The CAM model for evaluating polysemy captures polysemy/meta alternations to some extent.
- However, it is far from perfect.
- Future goals: An unsupervised version.


G. Boleda, S. Pádo, J. Utt.
Regular Polysemy: A Distributional Model.
*SEM 2012, p. 151-160.