Sense Disambiguation of Wikipedia Terms Based on Hidden Markov Model

Denis Turdakov
CMC MSU, ISP RAS

RCDL 2009

Outline

- Introduction
  - Word Sense Disambiguation
  - Hidden Markov Model and WSD
  - Wikipedia
- WSD and Wikipedia: State of the art
- Algorithm
  - Parameters estimation
  - Evaluation
- Conclusion

Word Sense Disambiguation

- Most common sense
- Lesk's algorithm (1986): "PINE CONE"
  - PINE

    1.kinds of evergreen tree with needle-sheped leaves

    2.waste away through sorrow or illness

  - CONE

    1.solid body which narrows to a point

    2.something of this shpe whether solid or hollow

    3.fruit of certain evergreen tree

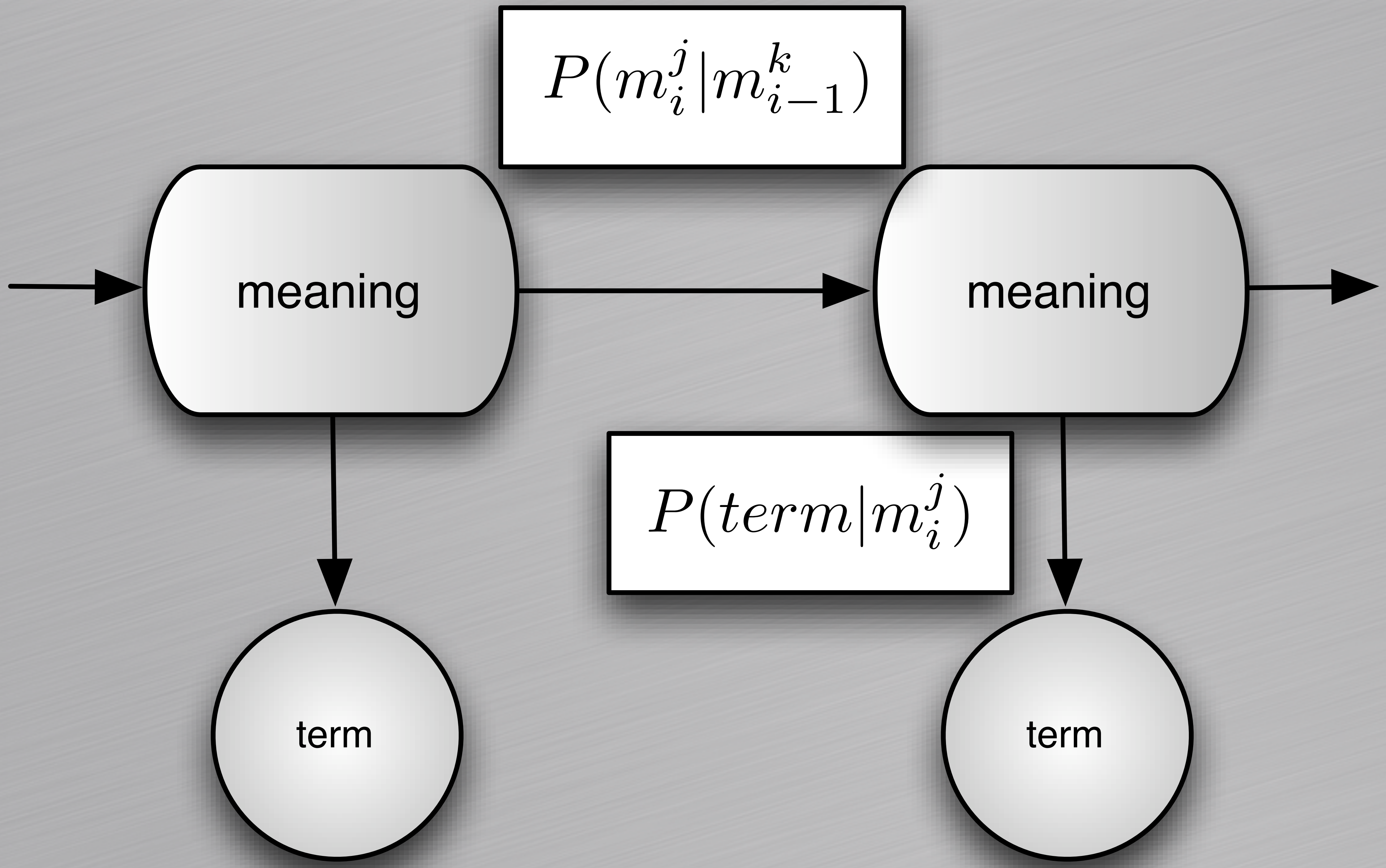
  - PINE#1 ∩ CONE#3 = 2

WSD problems

- What is meaning?
- What is context?
- How to evaluate and compare algorithms?

WSD Book (Springer 2006)

HMM and WSD



$P(m_i^j|m_{i-1}^k)$

meaning

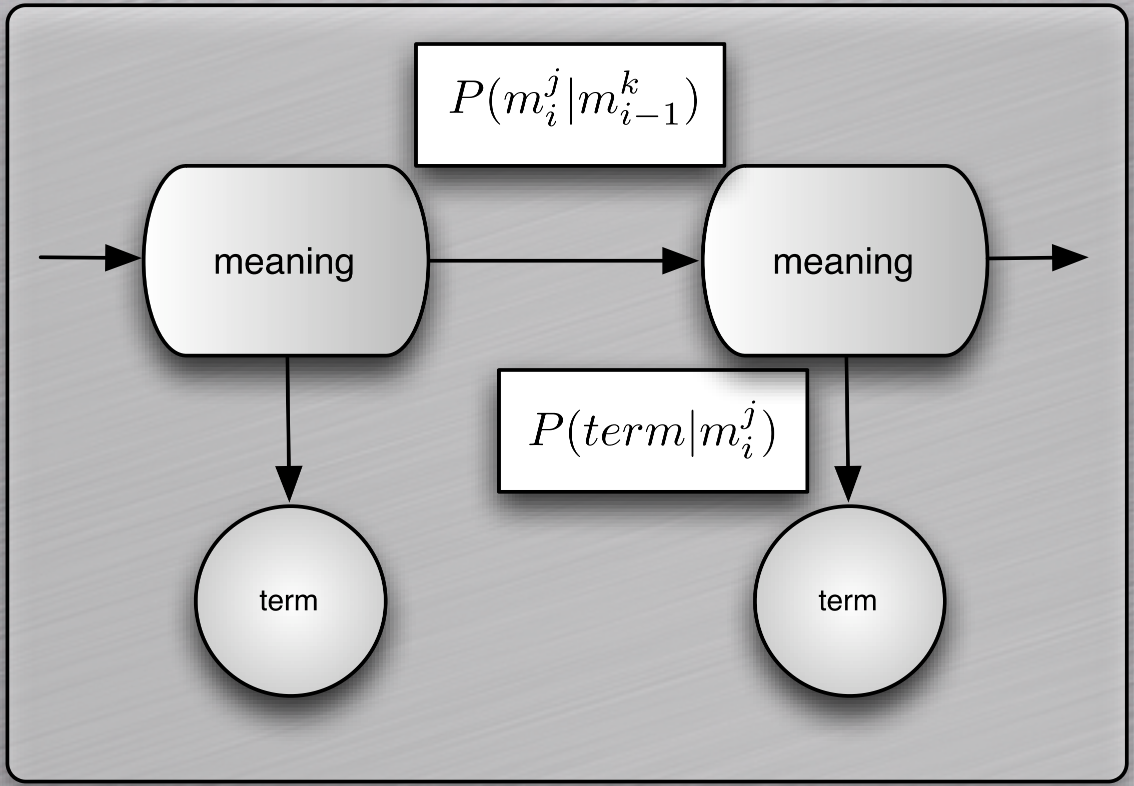
meaning

$P(term|m_i^j)$

term

term

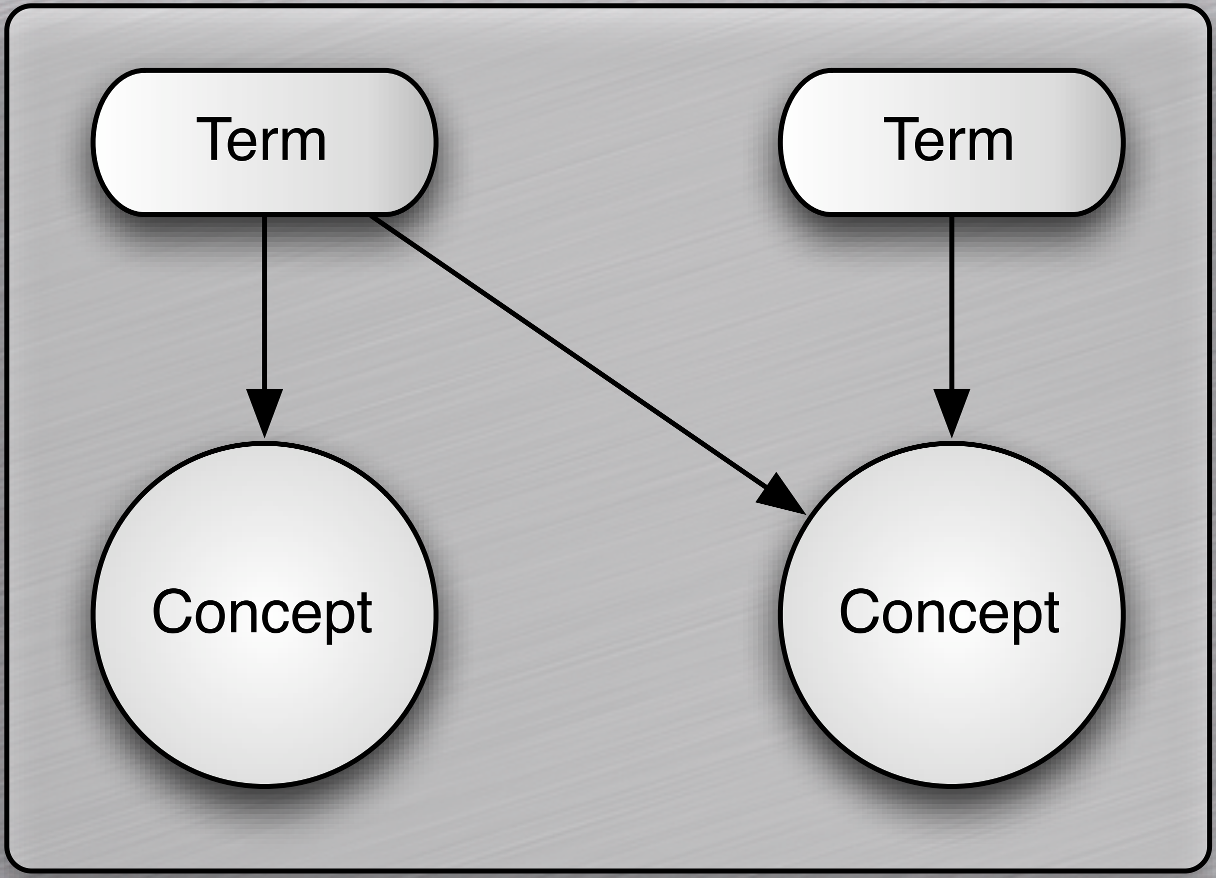
HMM and WSD



$$\hat{\mu} = \arg\max_{\mu} \left(\prod_{i=1}^{n} P(m_i|m_{i-k:i-1}) \cdot P(t_i|m_i) \right)$$

- C. Loupy, et. al. (1998): 72.3% (71.5% SemCor)
- A. Molina, et. al. (02, 04): 60.2% (58.0% SensEval-2)

Wikipedia

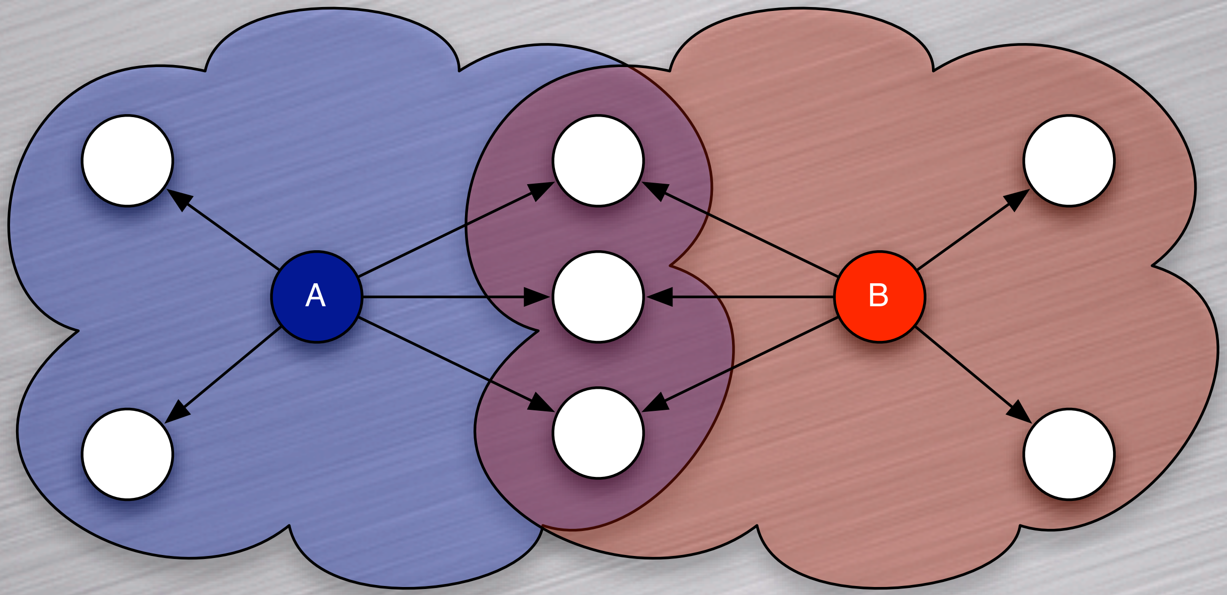


- >3M concepts
- Compound terms
- Disambiguation pages
- Synonyms
- Meanings
- Links: [[Concept | Term]]

WSD and Wikipedia

- R. Bunescu, M. Pasca (2007)
- S. Cucerzan (2007)
- R. Mihalcea, A. Csomai. Wikify! (2007)
- D. Turdakov, P. Velikhov (2008)
- O. Medelyan. Topic indexing with Wikipedia (2008)
- D. Milne, I. Witten. Learning to link with Wikipedia (2008)

Semantic Similarity



$$Sim(A,B) = \frac{2 \times \mid n(A) \cap n(B) \mid}{\mid n(A) \mid + \mid n(B) \mid}$$

$$n(B_1 B_2 \ldots B_m) = \bigcup_{i-1}^{m} n(B_i)$$

Estimation of parameters

- Transition model:

$$P(m_i|m_{i-k:i-1}) = \alpha \cdot [sim(m_i; m_{i-k:i-1}) + \beta \cdot P(m_i)]$$

$$\alpha = 1/2$$

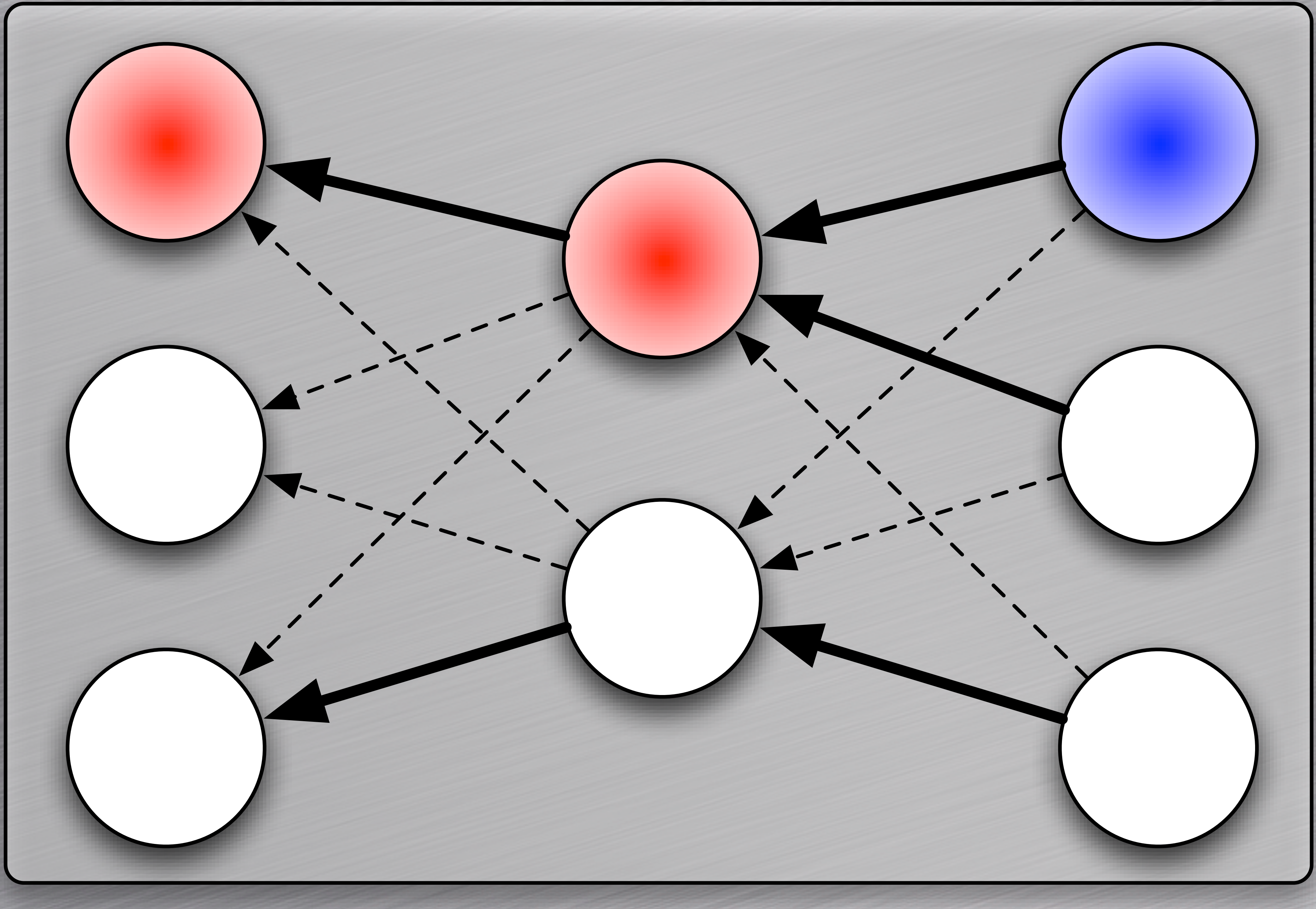
$$\beta = 1$$

$$P(m_i) = \frac{C(m_i)}{\sum_i C(m_i)}$$

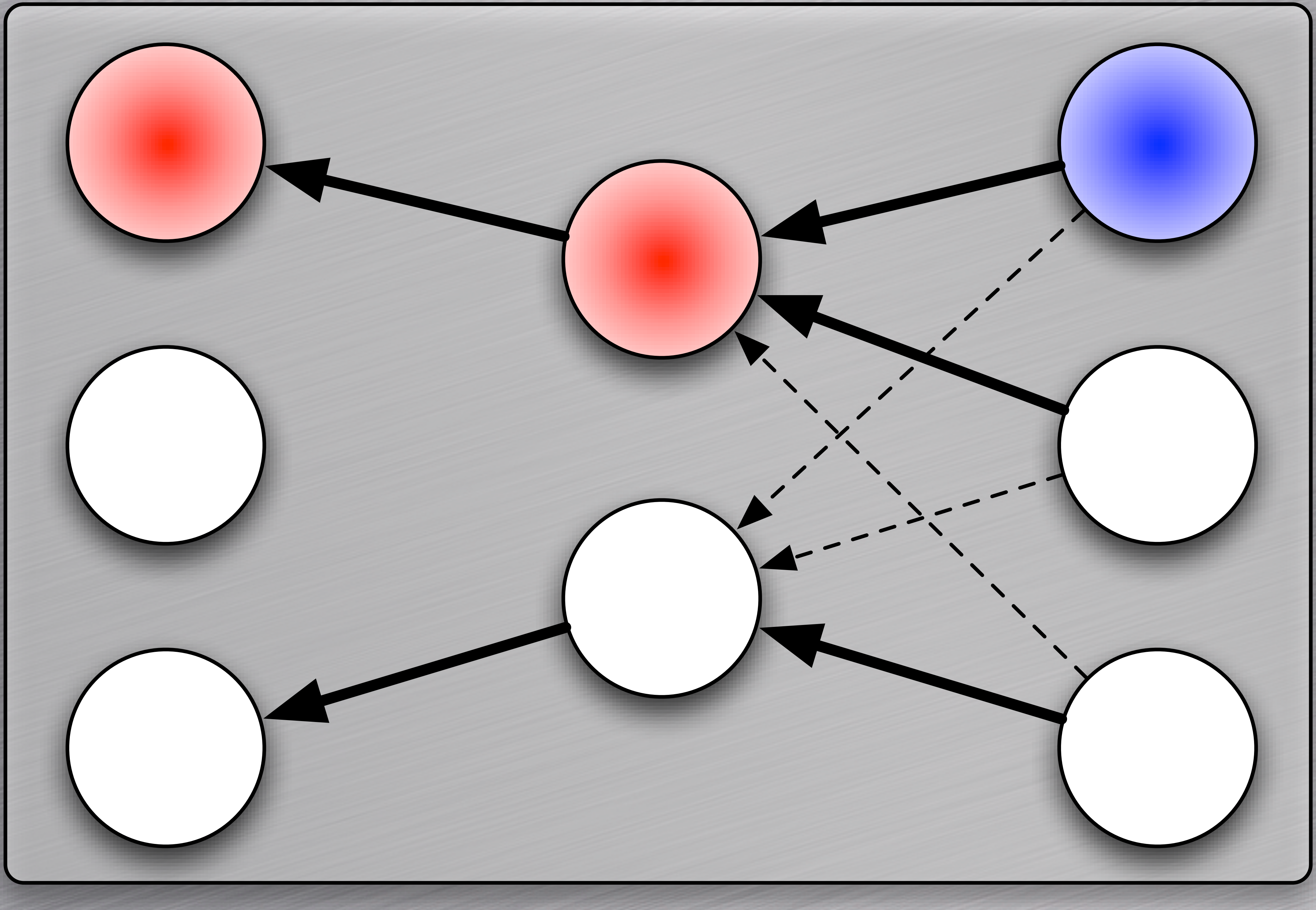
- Observation model:

$$P(t_i^j|m_i) = \frac{C(t_i^j, m_i)}{C(m_i)}$$

Algorithm Viterbi



Heuristics



Evaluation: Tests Collections

	News and scientific articles	Wikipedia articles
Number of documents	131	500
Number of terms	8236	50974
Ambiguous terms	6952	39332
Avg. number of meanings	22,34	35,34

Evaluation: Results

News

Order	HMM	Heuristics
0	53,12	53,12
1	54,00	54,00
2	54,50	54,49
3	**54,76**	54,72

Wikipedia articles

Order	HMM	Heuristics
0	91,34	91,34
1	91,64	91,64
2	92,40	92,37
3	**92,51**	92,41

Conclusion

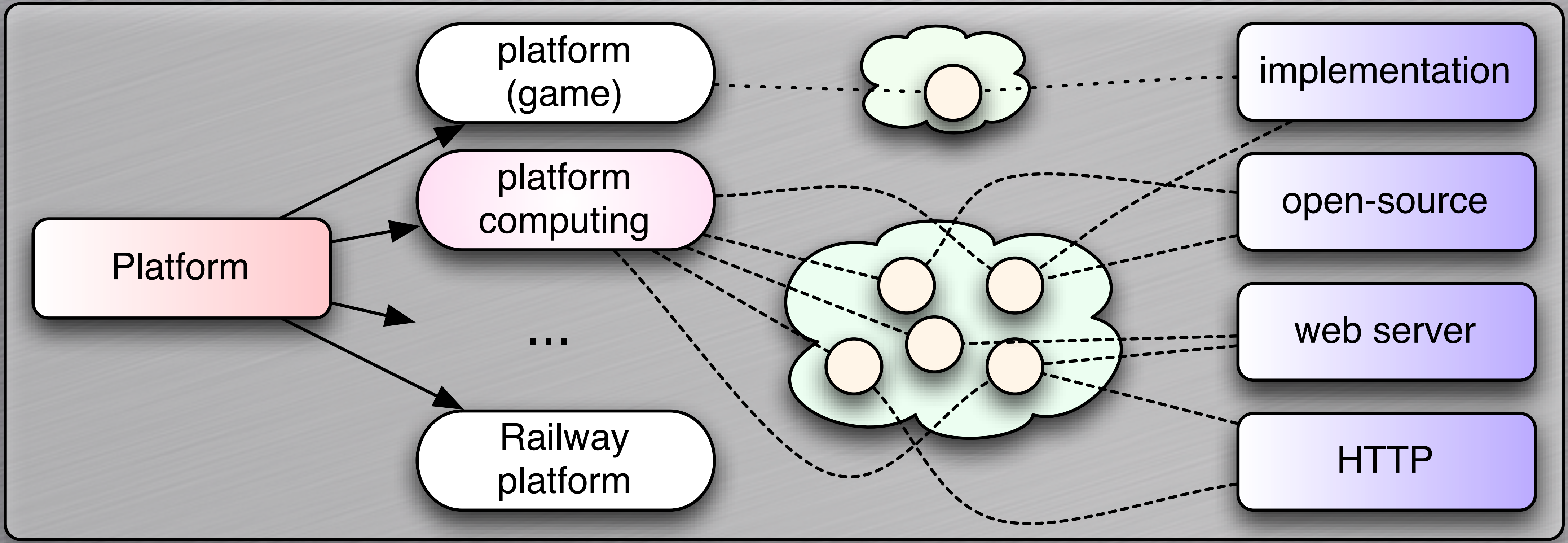
- Semantic similarity helps to estimate parameters of HMM in order to apply it to WSD
- Heuristics produces good results
- HMM is not the best model for WSD of multi theme documents

Thank you

Analog of the Lesk's algorithm

D.Turdakov, P.Velikhov (2008)

Jigsaw is W3C's open-source project that started in May 1996. It is a web server platform that provides a sample HTTP 1.1 implementation and …



Precision (Wikipedia July'08): 59,19%, 91,93%

Precision (Wikipedia March'09): 43,41%, 79,58%