Word Sense Disambiguation for Ontological Document Classification

Speaker: Georgiana Ifrim
Supervisors: Prof. Gerhard Weikum
Ph.D. Martin Theobald

Outline

- Word Sense Disambiguation
- Motivation
- Our approach
- Summary
- Future work
- References

Words and Semantics

- “He who knows not and knows not he knows not,
  He is a fool - Shun him.
- He who knows not and knows he knows not,
  He is simple - Teach him.
- He who knows and knows not he knows,
  He is asleep - Awaken him.
- He who knows and knows that he knows,
  He is wise - follow him."

Arabic proverb

Word Sense Disambiguation

- Many words have several meanings or senses
- Disambiguation: Determine the sense of an ambiguous word invoked in a particular context
- “He cashed a check at the bank”
- “They pulled the canoe up on the bank”

Word Sense Disambiguation

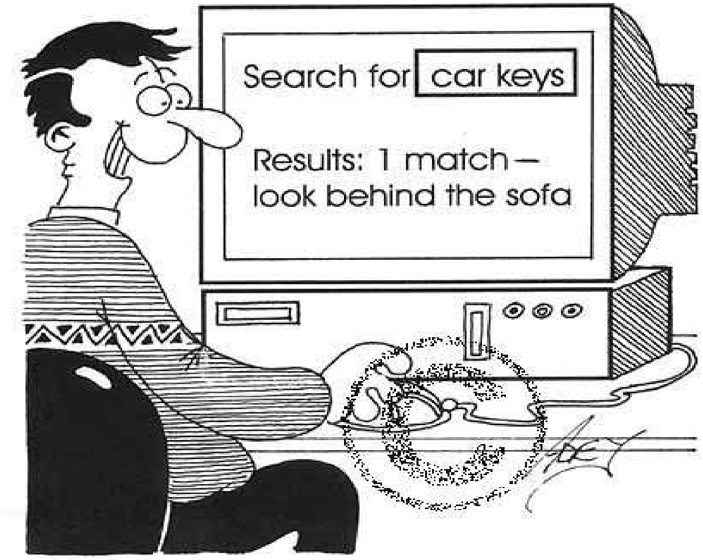
- 2-step process:
  - Determine the set of applicable senses of a word for a particular context
    - E.g: Dictionaries, thesauri, translation dictionaries
  - Determine which sense is most appropriate
    - Based on context or external knowledge sources

Word Sense Disambiguation

- Problems:
  - Difficult to define a WSD standard
    - What is the right separation of word senses?
    - Different dictionaries, different granularity of meanings
  - Clear and hierachical organization of word senses
    - Successful try: WordNet

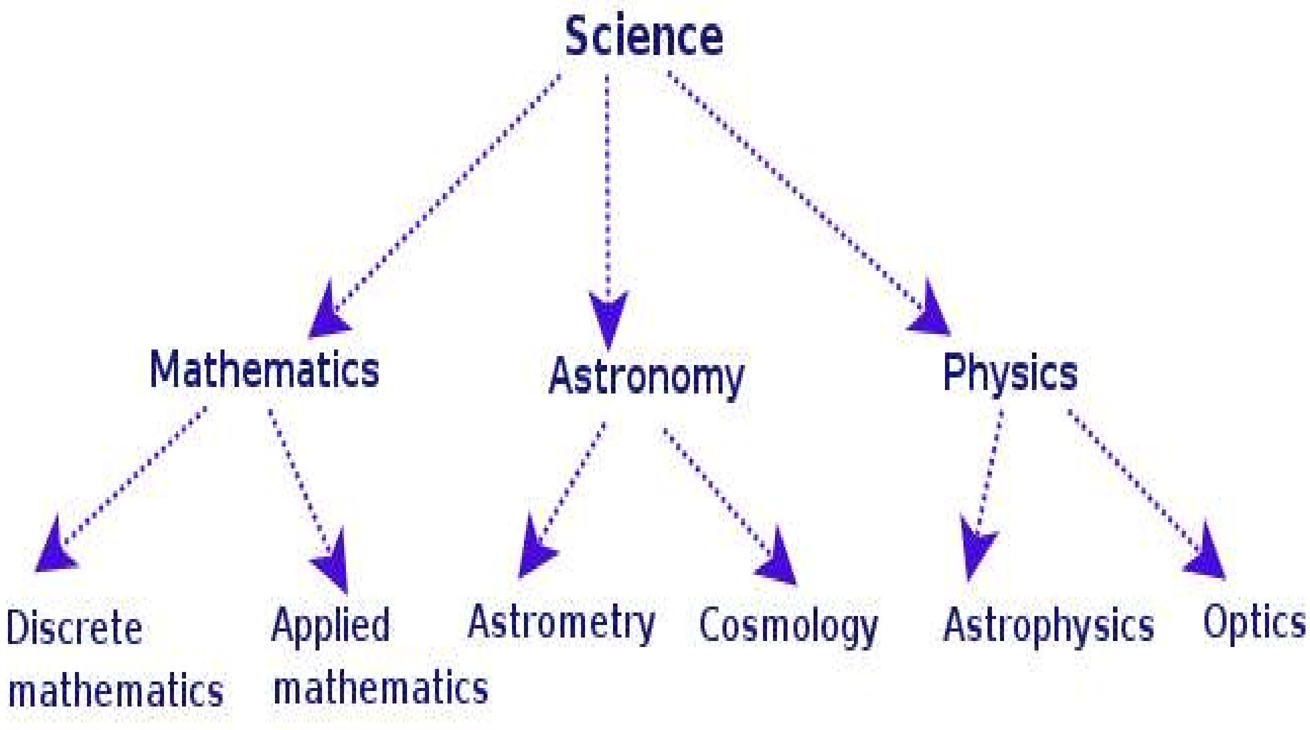
Word Sense Disambiguation

- Use of WSD:
  - NLP
    - Machine translation: English --> German
      - bank (ground bordering a lake or river) = Ufer
        bank (financial institution) = Bank
  - IR
    - Search engines
      - Query expansion
      - Query disambiguation
    - Automatic document classification



Word Sense Disambiguation

- Resources for WSD and classification:
  - Taxonomy: Tree of topics
    - Wikipedia

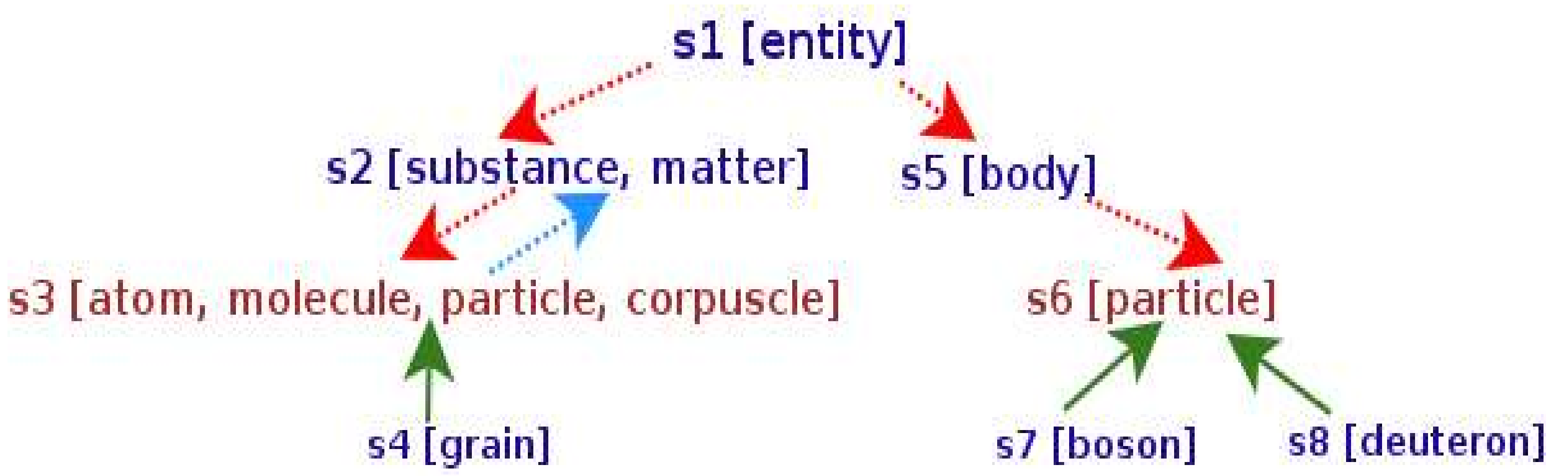


Word Sense Disambiguation

- Resources:
  - Ontology: DAG of concepts
  - WordNet
    - Large graph of concepts (semantic network)
      - Nodes: Set of words representing a concept (synset)
      - Edges: Hierarchical relations among concepts
        - Hypernym (generalization), Hyponym (specialization)
          e.g. tree hypernym of oak (IS-A)
        - Holonym (whole of), Meronym (part of)
          e.g. branch meronym of tree (PART-OF)
    - Contains ca. 150.000 nodes: nouns, verbs, adjectives, adverbs

Word Sense Disambiguation

- WordNet
  - Senses of particle
    - Hypernym
    - Hyponym
    - Meronym



Word Sense Disambiguation

- Resources:
  - Natural Language corpora
    - Wikipedia
    - BNC (British National Corpus)
    - SemCor
      - Sense-tagged corpus of 200.000 words
      - Subset of BNC
      - Each word type is tagged with its PoS and its sense-id in WordNet

Motivation

- Use WSD for automatic document classification
  - Capture semantics of documents by the concepts their words map to, in an ontology
  - Elimination of synonymy
    - Multiple terms with the same meaning are mapped to a single concept
  - Elimination of polysemy
    - The same term can be mapped to different concepts according to its true meaning in a given context
  - Reduction of training set size
    - Approximate matches can be found for formerly unknown concepts

Motivation

- Room for improving
  - Better selection of the feature space
    - Existing criteria: Counting of terms w.r.t. a given topic (MI criterion)
    - No stress on selecting the semantically significant terms that give the most benefit by disambiguation
  - New approaches for mapping words onto word senses
    - Use linguistics tools to extract more richly annotated word context
    - Feature sets mapped onto most compact ontological sub-domain
    - Enhance ontological topology by edges across PoS
  - Use WSD into a generative model

Our approach

- Given
  - A taxonomy tree of topics (Wikipedia)
    - Each topic has a label and a set of training documents
  - An ontology DAG of concepts (WordNet, customized)
    - Each concept has a set of synonyms, a short textual description and is linked by hierarchical relations
  - A set of lexical features observed in documents
  - A set of training documents with known topic labels and observed features, but unknown concepts
- Goal
  - For a given document, predict its topic label

Our approach

- 3 Stages:
  1. Naïve mapping
     - Map single features to single concepts using similarity of contexts measures (bag-of-words, no structure)
     - Select the most semantically representative concepts to feed to a classifier (MI on concepts)

Naïve mapping

- Naïve mapping example:
  - Nature or Computers?
    - mouse => WordNet => 2 senses:
      1. {mouse, rodent, gnawer, gnawing animal}
      2. {mouse, computer mouse, electronic device}
    - Compare term context con(mouse) with synset context con(sense) using some similarity measure
    - Term context: sentence in the document
    - Synset context: hypernyms, hyponyms + WordNet descriptions
    - Select the sense with the highest similarity

Naïve mapping

- Use:
  - Obtain sense-tagged resources
  - Estimate statistics about concepts:
    - Frequency (specificity)
    - Co-occurrence probabilities (quantified relations)
    - New edges in the ontology across PoS (verb-noun edges)
  - Extract better features (MI on concepts)

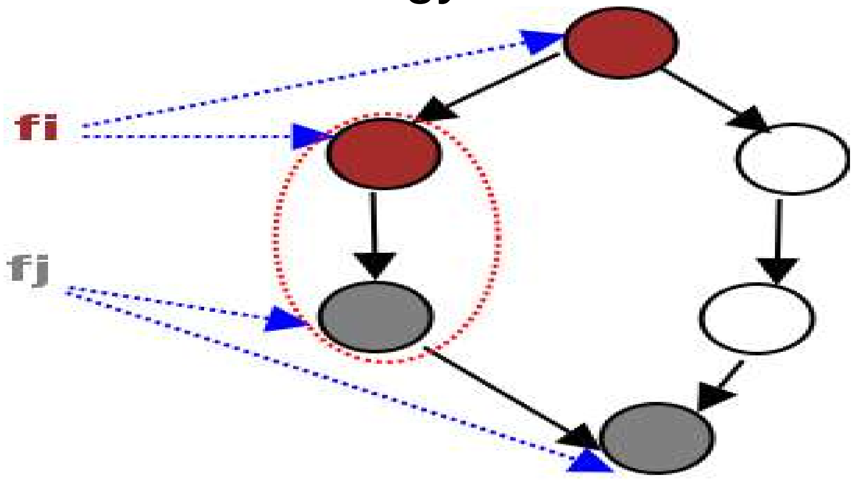
Naïve mapping

- Problems:
  - Context in the ontology very sensitive to noise
  - No structure of the ontology taken into account (bag of words approach, no structure)

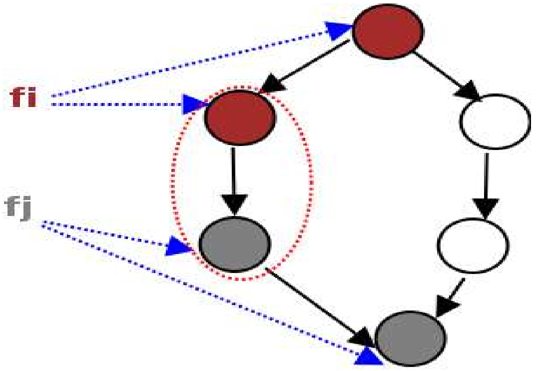
Our approach

2. Compact mapping

- Map sets of features to sets of concepts
- Consider structure of the ontology
- Select the most compact ontological subdomain to represent that set of terms
- Intuition: Concepts close in meaning are close in the DAG structure of the ontology



Compact mapping



- Try with pairs: verb-noun (same sentence)
  - v --> $\{s_v^1, \ldots, s_v^{l1}\}$
  - n --> $\{s_n^1, \ldots, s_n^{l2}\}$
  - Choose subset $\{s_v^i, s_n^j\}$ most compact: shortest path
  - Use statistics about concepts estimated in stage 1
  - Try with triplets: object (l1 senses)-verb (l2 senses)-subject (l3 senses): weighted MST

$$compactness(s_1^i, s_2^j, s_3^k) = \frac{1}{weight(MST(s_1^i, s_2^j, s_3^k))}$$

- l1 x l2 x l3 possible triplets
  - Wordnet worst case: 30x30x30 = 27,000 possible MSTs

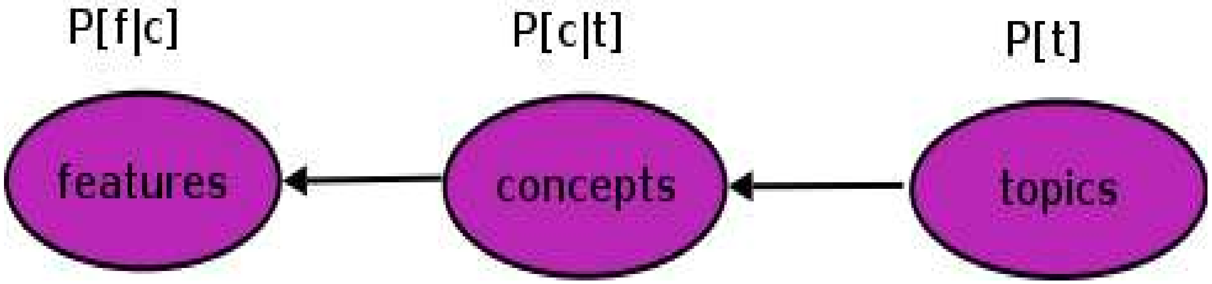
Compact mapping

- Use:
  - Disambiguating words with many equally likely meanings
- Advantages:
  - Avoids the context selection problem in the ontology
  - Investigation of triplets possible giving the best benefit, at low computational cost
- Problems:
  - General case: combinatorial explosion of possible number of MSTs

Our approach

3. Generative model – Bayesian approach

- Topics generate concepts
- Concepts generate features



$$P[d|t] = \Pi_f \sum_c P[f|c] * P[c|t]$$

$$P[t|d] = \frac{P[d|t] * P[t]}{P[d]} = \frac{P[d|t] * P[t]}{\sum_t P[d|t] * P[t]}$$

Generative model

- EM algorithm
  - Select a topic t with probability P[t]
  - Pick a latent variable c with probability P[c|t] (prob that topic t generated concept c)
  - Generate a feature f with probability P[f|c] (prob that word f means concept c)
  - Estimate parameters by maximizing the expected complete data log-likelihood
  - Initialize the parameters by a WSD step

Generative model

EM algorithm

1. Initialize parameters:

$$P[f|c] = sim(context(f), context(c))$$

$$P[c|t] = sim(context(c), context(t))$$

2. **E-step:**

$$P[c|f,t] = \frac{P[f|c] \cdot P[c|t]}{\sum_c P[f|c] \cdot P[c|t]}$$

n(f,t) = frequency of feature f in topic t

3. **M-step:**

$$P[f|c] = \frac{\sum_t n(f,t) P[c|f,t]}{\sum_f \sum_t n(f,t) P[c|f,t]}$$

$$P[c|t] = \frac{\sum_f n(f,t) P[c|f,t]}{\sum_c \sum_f n(f,t) P[c|f,t]}$$

$$P[t] = \frac{\sum_{f,c} n(f,t) P[c|f,t]}{\sum_t \sum_{f,c} n(f,t) P[c|f,t]}$$

4. Iterate until some termination condition.

5. Use estimates of parameters $P[f|c]$, $P[c|t]$, $P[t]$ into the classifier.

Generative model

- Advantages:
  - Semi-supervised approach
  - Uses unlabeled data to overcome the training set size problem
  - Combines WSD and statistical learning
- Problems:
  - Many parameters to estimate

Summary

- 3 modular approaches for ontological document classification
  - Naïve mapping
    - WSD using most similar concept (cosine measure)
    - Use hybrid feature space: terms+ concepts
  - Compact mapping
    - WSD using most compact ontological subdomain
    - Explore pairs: verb-noun, triplets: subject-verb-object
  - Generative model
    - Combines WSD and statistical modelling
    - Learn from unlabeled data

Future Work

- Tackle the details of the theoretical framework design
- Modular implementation of the 3 stages described
- Experiments
- Performance assessment

References

- "Foundations of Statistical Natural Language Processing", C. Manning, H, Schuetze, MIT, 1999
- "WordNet: An Electronic Lexical Database", C. Fellbaum, MIT, 1999
- "Exploiting Structure, Annotation and Ontological Knowledge for Automatic Classification of XML Data", M. Theobald, R. Schenkel, G. Weikum
- "Global organization of the WordNet lexicon", M. Sigman, G. Cecchi, 2002
- "Unsupervised Learning by Probabilistic Latent Semantic Analysis", T. Hofmann, 2001
- http://www.wikipedia.org

Thank you!