Other Similarity Measures for Unsupervised Semantic Disambiguation

Roberto Basili,

Department of Computer Science,
University Of Rome, Tor Vergata, Italy
e-mail: basili@info.uniroma2.it

Estratto dalla presentazione ad: LREC 2004, Lisbon, May 27, 2004
Summary

• Motivations
  • Sources of Information for Semantic Disambiguation
  • Semantic similarity in NL Learning
• A similarity measure based on Wordnet
• Application 1: *Unsupervised BNC Tagging through Wordnet*
• Conclusions
Motivations

• The notion of predominant senses require the estimation of probability distributions suitable for the target corpora
• Data sparseness limits the impact of contextual information in NL learning
  • Triples of (w1, w2, R) can be very rare in the target corpus and their frequency counts are unreliable
• Clusters of (similar) words provide more robust information
• Corpus-based clustering has been early used for NLP tasks (e.g. PP-disambiguation as in Pereira 1990)
• Semantically justified classes (e.g. WN high-level synsets) are more attractive
Motivations (2)

- Technologies for unsupervised labelling of target words are attractive as manual tagging is
  - Subjective (inter-annotator agreement is very low for complex cases, around 70%)
  - Expensive ... and sometimes
  - Unavailable in a timely fashion
- Attack data sparseness requires *clustering* and *partition* of source data to enable accurate estimation
Motivations (3)

• Generalization through the WN hierarchy forces a specific form of *clustering*
  
  • *Specific hypernyms* are able to separate different word senses
  
  • *General hypernyms* help to limit the number of target word classes (clusters)
  
• The suitable *trade-off* is usually *local* to single contexts (cfr. underspecification, e.g. (Buitelaar, 98))

• The global optimum should be inferred through a corpus
Motivations (cont’d)

• Wordnet contribution to semantic tagging
  • A source of evidence for a large set of lexical items (even words unseen in training data)
  • A linguistically principled way to generalize single observations
  • (hierarchical) constraints over word usage statistics (in line with other authors)

• Exploiting the similarity of words usage to induce their semantic similarity:
The idea

• To develop a semantic similarity measure that depends on word clusters

• Clusters may express different criteria of similarity local to specific phenomena.

• Examples are:
  • Same syntactic relations (direct objects of the same verb, e.g. *drink*)
  • Same topical informations (*concrete* nouns in “education”)
  • Same topological properties in a taxonomy (labels of concepts in a sub-hierarchy)
Conceptual Density

Basic terminology

• Target noun cluster $W$  
  (e.g. \{beer, water, stock\} as direct objects of a given verb, \(r=VDirobj\))

• (Branching Factor) Average number \(\mu\) of children in the tree rooted at synset \(s\), i.e. the average number of children of any node subsumed by \(s\)

• (Marks) Set of marks \(M\), i.e. the subset of nouns in \(W\) that are subsumed within the WN hierarchy rooted in \(s\). \(N = |M|\)

• (Area) \(area(s)\), total number of nodes in the hierarchy rooted at \(s\)
CD as metric of similarity

dotted edges are edges in the CD graph

entity

1

2

liquid

5

beverage

Empedocles

9

urine

11

body of water

12

water supply

w1ine beer water
CD as metric of similarity

diagram with nodes labeled as follows:
- entity
- liquid
- beverage
- color
- Empedocles
- urine
- body of water
- water supply

Terms:
- wine
- beer
- water
CD as metric of similarity

entity

liquid

beverage

Empedocles

urine

body of water

water supply

wine beer water
CD as metric of similarity

\[ cd^{(r)}(s) = \frac{\sum_{i=0}^{h} \mu^i}{area(s)} \]

\[ h = \begin{cases} \left\lfloor \frac{\log_{\mu}N}{N} \right\rfloor & \text{iff } \mu \neq 1 \\ N & \text{otherwise} \end{cases} \]
Conceptual Density (cont’d)
Conceptual Density (cont’d)
Conceptual Density (cont’d)

\[ cd^{(r)}(s) = \frac{\sum_{i=0}^{h} \mu^i}{area(s)} \]

\[ h = \begin{cases} \log_{\mu} N & \text{iff } \mu \neq 1 \\ N & \text{otherwise} \end{cases} \]
A Web Interface

http://terra.info.uniroma2.it:8080/Estimator/cd.htm

Conceptual Density Estimator

Please insert a set of words. Use a line for each word.

Example: Empedocle’s Elements
- air
- fire
- earth
- quintessence
- water
- element

Example: Colours
- red
- green
- blue
- white
- black
- gray
A greedy algorithm for disambiguation
A greedy algorithm for disambiguation

A synset $s$ is a useful synset with respect to a cluster $W$ iff $s$ is an hypernym of at least two different words in $W$
A greedy algorithm for disambiguation

A synset $s$ is a useful synset with respect to a cluster $W$ iff
$s$ is an hypernym of at least two different words in $W$

procedure greedy( $W$, Wordnet)
Let the output set of synsets $O$ be the empty set, i.e. $O = \emptyset$.
Let $S$ be the set of all useful synsets $s \in Wordnet$ with respect to $W$.
Rank elements $s \in S$ according to decreasing values of $cd(W)(s)$.
while $W \neq \emptyset$ and $S \neq \emptyset$
begin
Let $s \in S$ be the highest ranked element
Let $C \subset W$ be the set of nouns whose senses are hyponyms of $s$
$W = W - C$
$S = S - \{s\}$
if $C \neq \emptyset$ then
$O = O \cup \{s\}$
end
Using Conceptual Density

Target Noun set T (e.g. subjects of verb to march)

- horse (6 senses in WN1.7),
- ant (1 sense in WN1.7)
- troop (4 senses in WN1.7)
- division (12 senses in WN1.7)
- elephant (2 senses in WN1.7)

FIND the smaller set of synsets \{s\} that covers T and maximizes

\[ CD = \sum_s cd^{(r)}(s) \]

(1) organization organisation:
- horse, troops, divisions

(2) placental placental_mammal...:
- horse, elephant

(3) animal animate_being...:
- horse, elephant, ant

(4) army_unit:
- troop, division
ants, horses, troops, divisions, elephants

**Word:** horse  
**Number of senses:** 6  
**Aptoroi probability of each sense:** 0.1667

<table>
<thead>
<tr>
<th>Sense</th>
<th>Description</th>
<th>Probability (CD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 - 2036647</td>
<td>(solid-footed herbivorous quadruped domesticated since prehistoric times)</td>
<td>0.4365</td>
</tr>
<tr>
<td>2 - 3063881</td>
<td>(a narcotic that is considered a hard drug; a highly addictive morphine derivative)</td>
<td>0.0038</td>
</tr>
<tr>
<td>3 - 3082229</td>
<td>(a padded gymnastic apparatus on legs)</td>
<td>0.0097</td>
</tr>
<tr>
<td>4 - 3157070</td>
<td>(a chessman in the shape of a horse's head; can move two squares horizontally and one vertically (or in the same file)</td>
<td>0.0097</td>
</tr>
<tr>
<td>5 - 3610345</td>
<td>(a framework for holding wood that is being sawed)</td>
<td>0.0097</td>
</tr>
<tr>
<td>6 - 6029004</td>
<td>(troops trained to fight on horseback; &quot;500 horse led the attack&quot;)</td>
<td>0.5307</td>
</tr>
</tbody>
</table>

**Most probable sense:** 6029004  
**Probability:** 0.5307

---

**Word:** division  
**Number of senses:** 12  
**Aptoroi probability of each sense:** 0.0833

<table>
<thead>
<tr>
<th>Sense</th>
<th>Description</th>
<th>Probability (CD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 - 292500</td>
<td>(the act or process of dividing)</td>
<td>0</td>
</tr>
<tr>
<td>2 - 301937</td>
<td>(the act of dividing or partitioning; separation by the creation of a boundary that divides or keeps)</td>
<td>0</td>
</tr>
<tr>
<td>3 - 659171</td>
<td>(an arithmetic operation that is the inverse of multiplication; the quotient of two numbers is computed)</td>
<td>0</td>
</tr>
<tr>
<td>4 - 5022434</td>
<td>(one of the portions into which something is regarded as divided and which together constitute a whole)</td>
<td>0.0123</td>
</tr>
<tr>
<td>5 - 5989110</td>
<td>(discord that splits a group)</td>
<td>0</td>
</tr>
<tr>
<td>6 - 6796169</td>
<td>(an army unit large enough to sustain combat; &quot;two infantry divisions were held in reserve&quot;)</td>
<td>0.4457</td>
</tr>
<tr>
<td>7 - 6801431</td>
<td>(a group of ships of similar type)</td>
<td>0.2054</td>
</tr>
<tr>
<td>8 - 6801518</td>
<td>(a unit of the United States air force usually comprising two or more wings)</td>
<td>0.2054</td>
</tr>
<tr>
<td>9 - 6802831</td>
<td>(an administrative unit in government or business)</td>
<td>0.0743</td>
</tr>
<tr>
<td>10 - 6803008</td>
<td>(botany) taxonomic unit of plants corresponding to a phylum)</td>
<td>0.0037</td>
</tr>
<tr>
<td>11 - 6803446</td>
<td>(biology) a group of organisms forming a subdivision of a larger category)</td>
<td>0.0037</td>
</tr>
</tbody>
</table>
| 12 - 6818817 | (a league ranked by quality; "he played baseball in class D for two years", "Princeton is in the NCA"

**Most probable sense:** 6796169  
**Probability:** 0.4457
Appl1: Semantic Tagging

- \( \Omega = \arg \max_\Gamma p(\Gamma | tw, r_1 \ldots r_n) \)

\[ \Rightarrow \quad \Omega = \arg \max_\Gamma \prod_i p(\Gamma | tw, r_i) \]

- Pairs \((tw, r_i)\) are syntactic trigrams in the target context, e.g.
  - (dish, wash, DO), "John washed the \textit{dishes} …"
  - (dish, cook, DO), "Mary cooked \textit{wonderful dishes} for dinner …"
  - (company, declare, S), "The \textit{company} declares the joint venture with …"
The model parameters

- $\Omega = \arg \max_\Gamma p(\Gamma | tw, r_1 \ldots r_n)$

  $\Rightarrow \Omega = \arg \max_\Gamma \prod_j p(\Gamma | tw, r_j)$

- CD can be used to estimate the following parameters:
  - $p(\Gamma | tw, r_i)$ \textit{contextual probabilities}
  - $p(\Gamma | tw)$ \textit{(back-off) lexical probabilities}
  - $p(\Gamma | r_i)$ \textit{(back-off) syntactic probabilities}
Apply a simple Bayesian model to incoming contexts given by

\[ <tw \ r_1, \ ..., \ r_k> \]

by selecting

\[ \Omega = \arg\max_{\Gamma} \ p(\Gamma \mid <tw \ r_1, \ ..., \ r_k>) = \]

\[ \approx \arg\max_{\Gamma} \ \hat{p}(\Gamma \mid tw\ r_1) \cdot \hat{p}(\Gamma \mid tw\ r_2) \cdot \ldots \cdot \hat{p}(\Gamma \mid tw\ r_k)) \]

(OBS: if \ counts(<tw \ r_j>)<N \ then

\[ \hat{p}(\Gamma \mid tw\ r_j) = \text{(back-off)} = \alpha\hat{p}(\Gamma \mid tw) + \beta\hat{p}(\Gamma \mid r_j)) \]
## Backing-off (Katz, 1987)

| $P_{katz}(z|x,y)$ | $P^*(z|x,y)$, if $C(x,y,z) > 0$ |
|-------------------|----------------------------------|
|                   | $\alpha(x,y)P_{katz}(z|y)$, else if $C(x,y) > 0$ |
|                   | $P^*(z)$, otherwise.              |

| $P_{katz}(z|y)$   | $P^*(z|y)$, if $C(y,z) > 0$ |
|-------------------|----------------------------------|
|                   | $\alpha(y)P^*(z)$, otherwise.   |

- The CD-based back-off exploits semantic classes to derive:
  - $p(\Gamma | tw)$ probability that $tw$ is a kind of $\Gamma$
  - $p(\Gamma | r_j)$ probability that $r_j$ lexically prefers some $\Gamma$
  - $p(\Gamma)$ global probability of class $\Gamma$ in the corpus
Unsupervised Tagging through Wordnet

Diagram:
- Longman
- WN2Longman
- WN
- Annotation
- Corpus
- Training
- Test
- WN2LD
- Sensum
- CD Estimation
- Word classes & probs
- BO SemTagger

Processes:
- BO SemTagger
- WN2LD
- Sensum
- CD Estimation
Experimental Set-UP

- Estimate over Wordnet and the corpus and then map them into Longman Dictionary categories the following quantities:
  \[ p(\Gamma \mid hw \ r), \quad p(\Gamma \mid r), \quad p(\Gamma \mid hw) \]

- \( r \) ranges among the main grammatical dependencies:
  - SubjV (e.g. tw drinks …)
  - DirObjV (e.g. drink an tw)
  - V_P_hw (e.g. drinking with an tw)
  - N_P_hw (e.g. a drink with tw)
  - hw_P_N (e.g. a tw for a drink)
  - Adj_hw (e.g. drinkable tw)
The target tag set

Abstract

Concrete

Animate

Inanimate

Plant

Animal

Human

Liquid

Gas

Solid

Movable

Non-movable

MaleAnim

FemaleAnim.

MaleAnim.

FemaleAnim.
The target tag set

Abstract
The target tag set

- Concrete
  - Animate
    - Male
    - Female
    - Plant
    - Animal
    - Human
    - Liquid
    - Gas
    - Solid
  - Inanimate
- Abstract
  - Physical Qualities
  - Organic
- Movable
  - MaleAnim
  - FemaleAnim
- Non-movable
  - MaleAnim.
  - FemaleAnim.
The target tag set
Evaluation Set-Up

• Target Corpus: BNC

• Controlled material: BNC texts manually annotated or all the heads of simple and complex noun phrases

• Total number of annotated instances in BNC:
  • 214,446 human annotated noun phrases
  • 29,071 unique vocabulary items (Unlemmatized)
  • 59 % of head nouns tagged as Abstract
Experimental Set-Up

- Held-out (size: \(\approx 99,000\))
  - only words ambiguous in the dictionary
  - Baseline: most frequent tag (\(\approx 54\%\) accuracy)
- Blind (size: \(\approx 13,097\))
  - random selections
  - Baseline: most frequent tag (\(\approx 80\%\) accuracy)
- Wordnet baseline
  - Algorithm:
    I. Pick the 1st Wordnet sense for tw
    II. Map it to Longman
## Evaluation

<table>
<thead>
<tr>
<th>Held-Out</th>
<th>Accuracy (WITHIN LD)</th>
<th>First 2 choices</th>
<th>First 3 choices</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised (i.e. ME)</td>
<td>~83.70%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pick 1\textsuperscript{st} WN sense</td>
<td>72.40%</td>
<td>89.21%</td>
<td>92.02%</td>
</tr>
<tr>
<td>Unsupervised Tagger</td>
<td>75.45%</td>
<td>91.28%</td>
<td>96.68%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Blind</th>
<th>Accuracy (WITHIN LD)</th>
<th>First 2 choices</th>
<th>First 3 choices</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pick 1\textsuperscript{st} WN sense</td>
<td>68.74%</td>
<td>92.28%</td>
<td>95.10%</td>
</tr>
<tr>
<td>Unsupervised Tagger</td>
<td>81.05%</td>
<td>95.17%</td>
<td>98.43%</td>
</tr>
</tbody>
</table>
Conclusions

• A robust semantic similarity measure over Wordnet.
  • Less prone to sparse data and overfitting
  • Generalize to meaningful noun classes as explanations of corpus phenomena

• Successfully applied to:
  • Semantic Tagging (syntactic clusters) (about 81% accurate)
  • Harmonisation of semantic resources (dictionaries vs. lexicons)

• It supports development of:
  • corpus-dependent semantic dictionaries
  • “lexicalized” contextual cues (as probabilities)

• Already used for ontology alignment problems (see WebInt03)

• URL: http://terra.info.uniroma2.it:8080/Estimator/cd.htm