Word Sense Disambiguation as an Integer Linear Programming Problem

Vicky Panagiotopoulou[1], *Iraklis Varlamis*[2], *Ion Androutsopoulos*[1], *and George Tsatsaronis*[3]

[1] Department of Informatics, Athens University of Economics and Business

[2] Department of Informatics and Telematics, Harokopio University of Athens

[3] Biotechnology Center (BIOTEC), Technische Universitat Dresden

SETN 2012

7th Hellenic Conference on Artificial Intelligence,

May 28-31, 2012, Lamia

Contents

- Problem statement
- Existing solutions
- Our solution
- Implementation
- Results
- Conclusions

Word Sense Disambiguation - WSD

- Assign to every **word** of a document the most appropriate meaning (**sense**) among those offered by a **lexicon** or a **thesaurus** *(inventory of senses)*
  - Some examples:
    - The two friends jumped off the **bank** and into the water.
      - bank = sloping land - especially the slope beside a body of water.
    - They passed by the **bank** to make a deposit.
      - bank = a financial institution that accepts deposits and channels the money into lending activities.
    - They used the **bank** when the army entered the city.
      - bank = a supply or stock held in reserve for future use (especially in emergencies).
  - What is the correct meaning of "**bank**" in each sentence?

How hard is the WSD task?

	Senseval 2					*Senseval 3*				
	N	**V**	**Adj.**	**Adv.**	**All**	**N**	**V**	**Adj.**	**Adv.**	**All**
Mono.	260	33	80	91	464	193	39	72	13	317
Poly.	813	502	352	172	1839	699	686	276	1	1662
Av. Poly.	4.21	9.9	3.94	3.23	5.37	5.07	11.49	4.13	1.07	7.23
Av. Poly. (P. only)	5.24	10.48	4.61	4.41	6.48	6.19	12.08	4.95	2.0	8.41

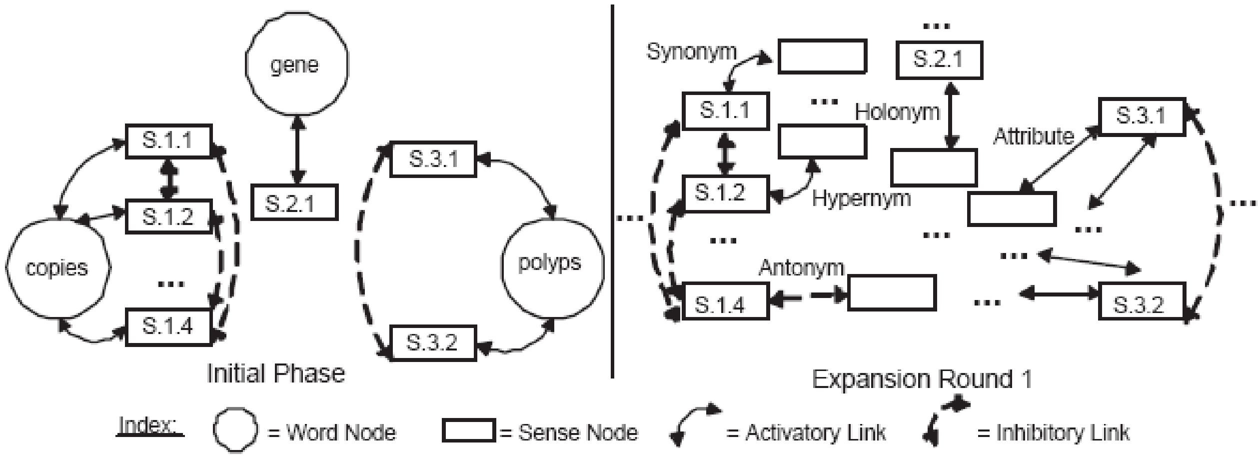
- Upper Bound: Human performace; 95%-99% coarse-grained senses, 65-70% with fine-grained senses [Haliday and Hasan, 1976].
- Lower Bound: Unsupervised Baseline: 13-20%, Supervised Baseline: 61-64%
- Inter-annotator agreement: 67% - 80% [Snyder and Palmer, 2004]

WSD alternatives

- Several options in applying WSD:
  - Unsupervised
    - High coverage, lower accuracy than supervised, no need for manually annotated data set
  - Supervised
    - Lower coverage than unsupervised, higher accuracy, “*knowledge acquisition bottleneck*”

Graph-based Unsupervised WSD

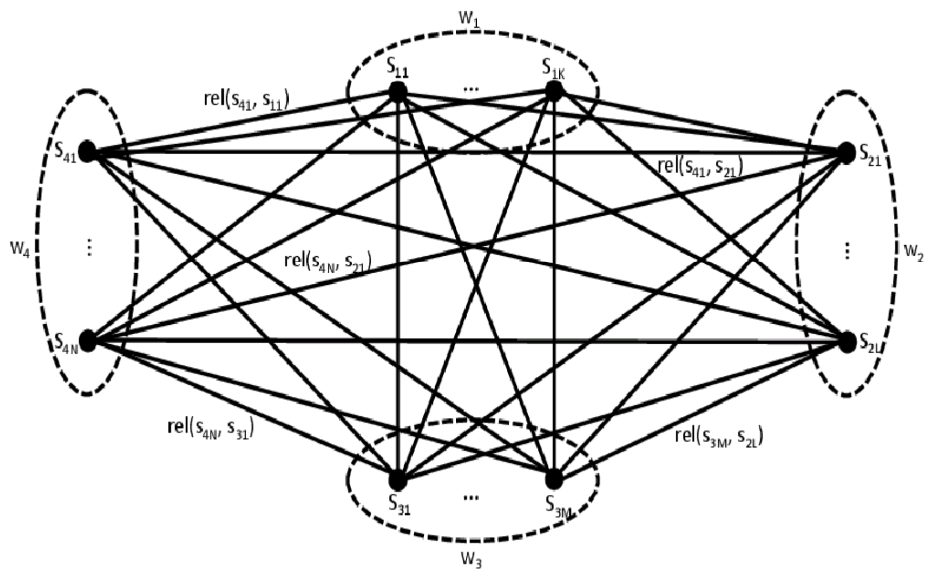
- Map all words' senses to nodes of a graph
- Expand the graph by adding related senses until a connected graph is constructed



- Rank graph nodes (senses) using graph based metrics (or node activation techniques)
- Each word is mapped to its most highly ranked (or most active) sense

Our suggestion

- Model WSD as an Integer Linear Programming (ILP) Problem
- Select exactly one possible sense of each word in the input sentence, so as to maximize the total pairwise relatedness between the selected senses
- Create a graph that contains only the candidate senses for each word
- The edges denote relatedness between senses

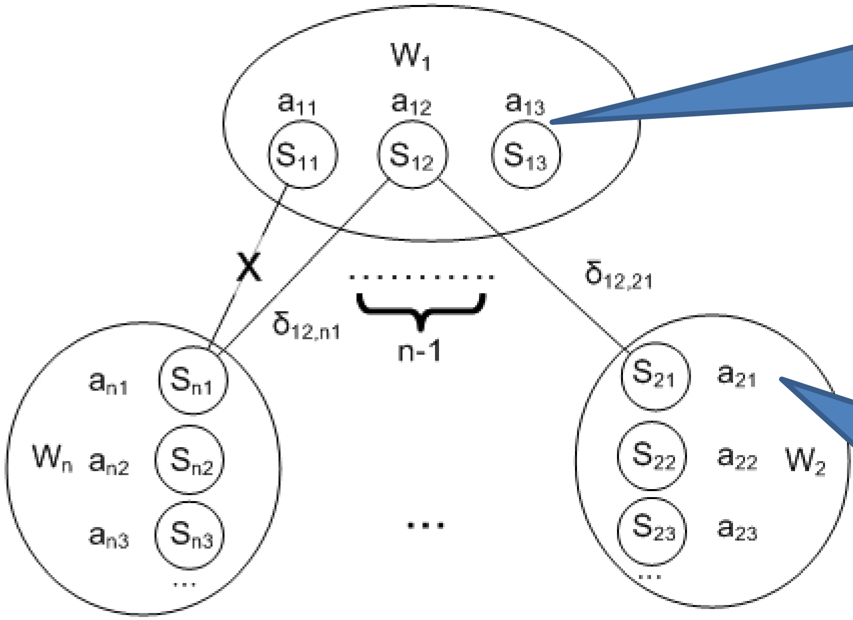


WSD as ILP vs Graph-based WSD

- Complete but smaller graphs that contain only words' senses
- Weighted edges are created using any pairwise sense relatedness measure (semantical or statistical)
- ILP is NP-hard, however for small graphs and using efficient solvers the method is faster than graph-based WSD

- Connected big graphs that contain extra (interconnecting) senses
- The edges are lexical relations from a thesaurus, and are usually unweighed
- Semantic network construction is slow
- Spreading of activation or node ranking run for each new graph

Towards an ILP formulation



s_{1j}: possible senses of w_1.
a_{1j}: shows whether s_{1j} is selected (a_{1j} =1) or not (a_{1j} = 0)

s_{2j}: possible senses of w_2.
a_{2j}: shows whether s_{2j} is selected (a_{2j} =1) or not (a_{2j} = 0)

max $\sum_{i,j,i',j',i<i'} rel(s_{ij}, s_{i'j'}) \cdot a_{ij} \cdot a_{i'j'}$

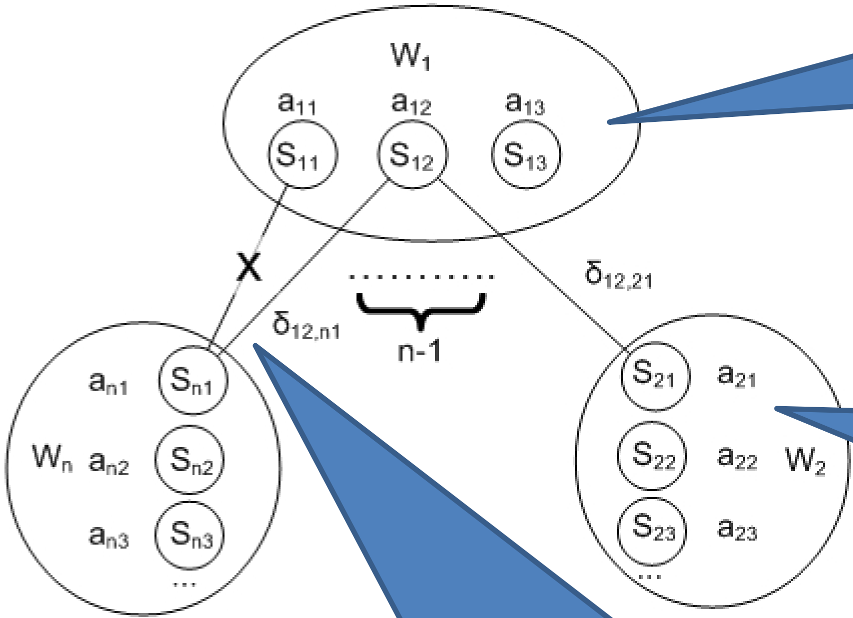
s.t. $a_{ij} \in \{0,1\}, \quad \forall i,j$

and $\sum_j a_{ij} = 1, \quad \forall i$

Maximize the total pairwise relatedness between selected senses, using only one sense per word

Results in a quadratic objective function

Our ILP model for WSD



s_{1j}: possible senses of w_1.
a_{1j}: shows if s_{1j} is selected (a_{1j} =1) or not (a_{1j} = 0)

s_{2j}: possible senses of w_2.
a_{2j}: shows if s_{2j} is selected (a_{2j} =1) or not (a_{2j} = 0)

$\delta_{ij,i'j'}$: Shows whether the ***edge*** is ***active*** (1) or not (0).
The edge is active *iff* both connected senses are active ($a_{ij} = a_{i'j'}$ =1).

Our ILP model for WSD

$\max \sum_{i,j,i',j',i<i'} rel(s_{ij}, s_{i'j'}) \cdot \delta_{ij,i'j'}$

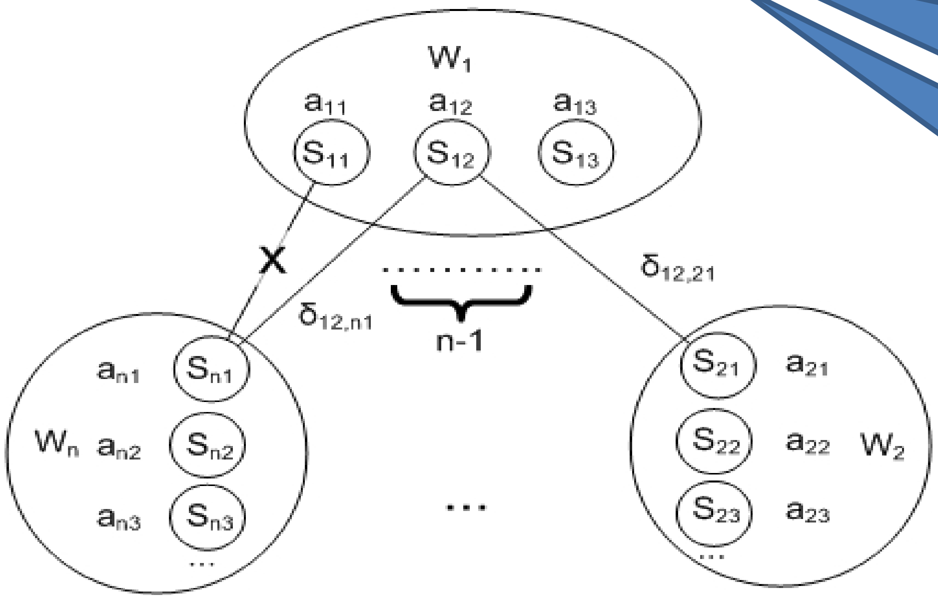
s.t. $a_{ij} \in \{0,1\}, \quad \forall i,j$

and $\sum_j a_{ij} = 1, \quad \forall i$

and $\delta_{ij,i'j'} \in \{0,1\}, \quad \forall i,j,i',j'$

and $\delta_{ij,i'j'} = \delta_{i'j',ij}, \quad \forall i,j,i',j'$

and $\sum_{j'} \delta_{ij,i'j'} = a_{ij}, \forall i,j,i'$



Maximize the total pairwise relatedness between senses, taking into account senses connected via active edges

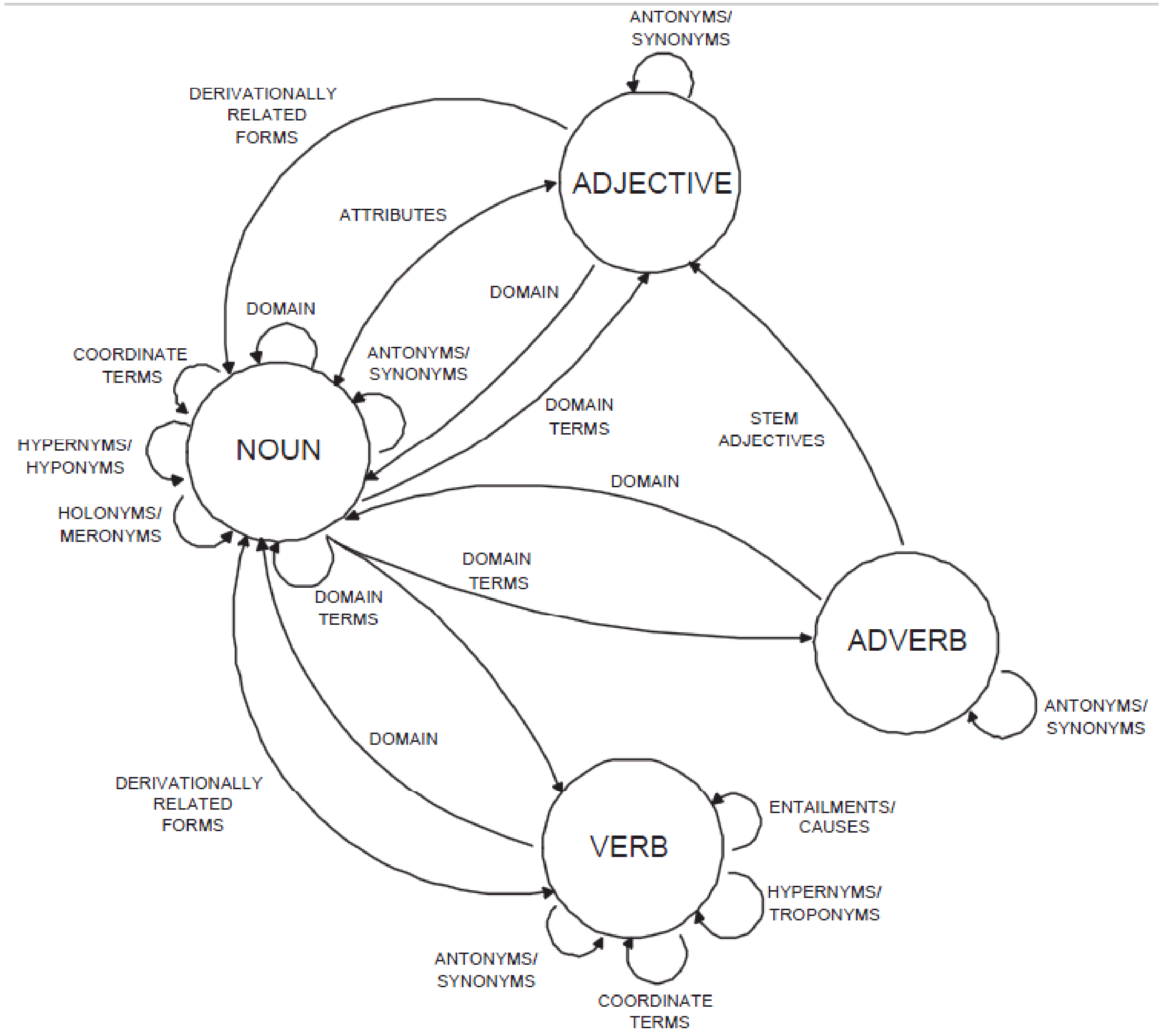
Edges are undirected

If s_{ij} is selected ($a_{ij} = 1$), then there is exactly one active edge from **s_{ij}** to the senses of all other words **$w_{i'}$**.

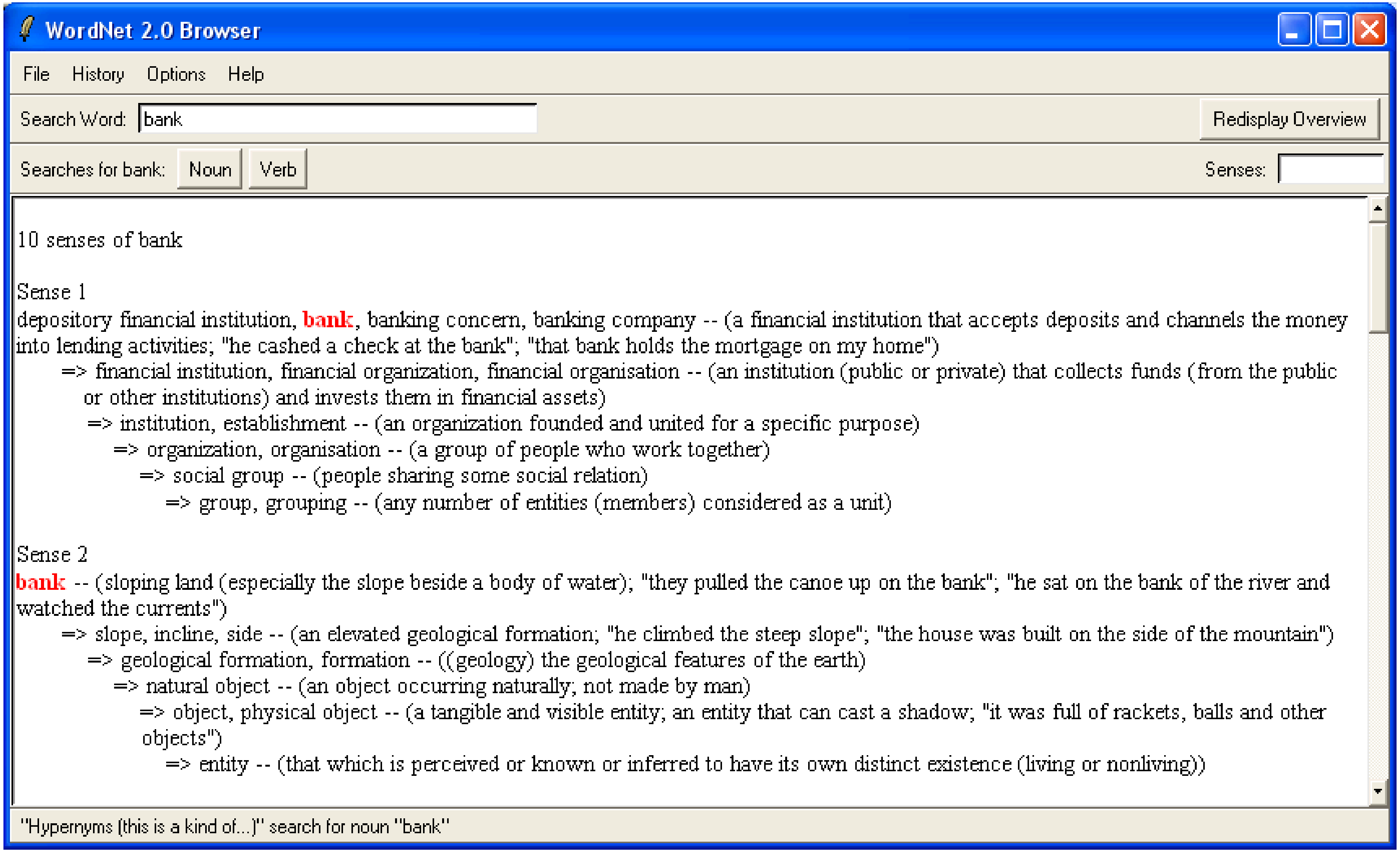
If s_{ij} is not selected ($a_{ij} = 0$), then there is no active edge from **s_{ij}** to senses of other words **$w_{i'}$**.

Resources: WordNet [Miller *et al.*]

- Each **sense** is a set of **synonym words** (***synset***)
  - has a gloss and a POS (noun, verb, adjective, adverb)
  - is connected to other senses



WordNet [Miller et al.]



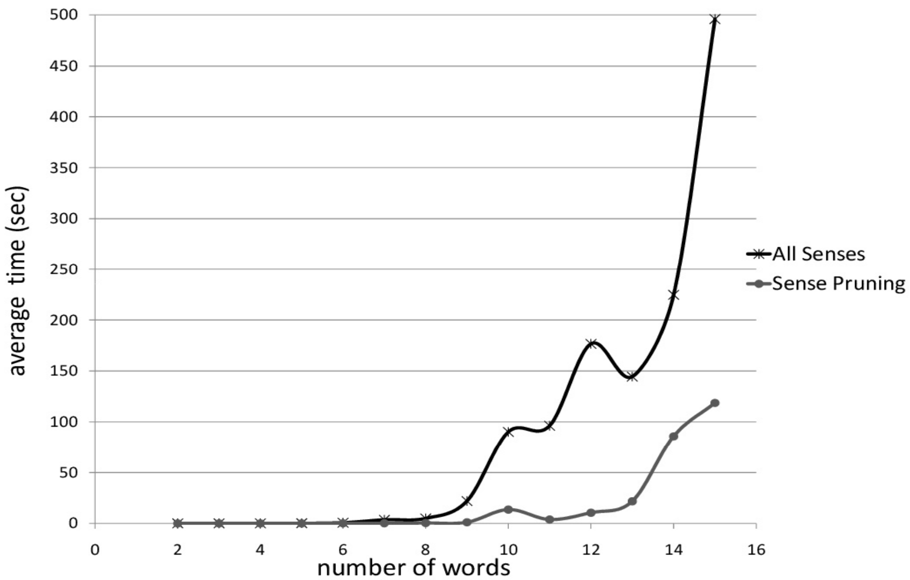
Implementation

- Relatedness measures
  - Semantic Relatedness (SR) is a knowledge-based measure that uses WordNet $SR(s_1, s_2) = \max_{P=\langle s_1,\ldots,s_2\rangle} \{SCM(P) \cdot SPE(P)\}$
    - Semantic compactness (SCM): the semantic path from s1 to s2 is short and contains highly related senses
    - Semantic Path Elaboration (SPE): the senses in the path are very specific
  - Pointwise mutual information (PMI) is a statistical similarity measure
  - We use the WordNet glosses of each sense s_i, and a non-sense tagged corpus (953 million tokens)

$$PMI(s_1, s_2) = \frac{\sum_{w_1 \in g(s_1),\, w_2 \in g(s_2)} PMI(w_1, w_2)}{|g(s_1)| \cdot |g(s_2)|}$$

Implementation

- ILP Solver
  - lp_solve: A **branch-and-bound** implementation that uses **Simplex for** LP subproblems. Available at http://lpsolve.sourceforge.net/
- Sense pruning: The sense s_{ij} of a word w_i is removed from the graph if the gloss of s_i and the sentence of w_i do not overlap
  - The resulting graph is smaller, faster execution with comparable WSD performance



Method	Avg. time (secs) per sentence
SAN	101.38
PR	91.92
ILP-SR-FULL	82.81
ILP-SR-PRUN	23.45
ILP-PMI-FULL	81.46
ILP-PMI-PRUN	17.40

-P problem

Experimental results

	Noun				Verb				Adjective				All			
Method	**C**	**P**	**R**	**F**	**C**	**P**	**R**	**F**	**C**	**P**	**R**	**F**	**C**	**P**	**R**	**F**
SAN	72.2	27.8	20.0	23.3	71.1	19.6	13.9	16.3	72.4	39.6	28.7	33.3	71.9	27.9	20.0	23.3
PR	72.2	45.5	32.8	38.1	71.1	30.0	21.3	24.9	72.4	38.8	28.1	32.6	71.9	39.4	28.4	33.0
ILP-SR-FULL	99.6	38.6	38.4	38.5	99.6	25.0	24.9	24.9	92.8	37.4	34.7	36.0	98.1	34.2	33.5	**33.8**
ILP-SR-PRUN	99.6	38.6	38.4	38.5	99.6	24.6	24.5	24.5	92.8	37.7	35.0	36.3	98.1	34.1	34.4	**33.8**
ILP-PMI-FULL	99.6	27.9	27.7	27.8	98.9	23.4	23.2	23.3	100.0	37.9	37.9	37.9	99.5	28.6	28.4	28.5
ILP-PMI-PRUN	99.6	28.6	28.5	28.6	98.9	24.7	24.5	24.6	100.0	43.5	43.5	43.5	99.5	30.5	30.4	30.5

Table 2: Coverage (C), precision (P), recall (R), and F_1-measure (F) of WSD methods on the Senseval 2 dataset, polysemous words only, excluding adverbs, without using the first-sense heuristic. The results are percentages.

	Noun				Verb				Adjective				All			
Method	**C**	**P**	**R**	**F**	**C**	**P**	**R**	**F**	**C**	**P**	**R**	**F**	**C**	**P**	**R**	**F**
SAN	97.9	30.6	29.9	30.2	94.2	28.8	27.1	27.9	94.9	37.8	35.9	36.8	95.8	31.0	29.7	30.4
PR	97.9	38.3	37.5	37.9	94.2	39.6	37.3	38.4	94.9	40.5	38.4	39.4	95.8	39.2	37.6	**38.4**
ILP-SR-FULL	99.9	32.3	32.2	32.3	98.0	25.8	25.3	25.6	97.0	38.3	37.1	37.7	98.6	30.6	30.2	30.4
ILP-SR-PRUN	99.9	32.0	31.9	32.0	98.0	25.8	25.3	25.6	97.0	38.7	37.5	38.1	98.6	30.5	30.1	30.3
ILP-PMI-FULL	96.7	30.2	29.2	29.7	94.1	18.1	17.1	17.6	96.9	39.4	38.2	38.8	95.7	26.9	25.8	26.3
ILP-PMI-PRUN	96.7	27.3	26.4	26.8	94.1	19.3	18.2	18.7	96.9	39.0	37.8	38.4	95.7	26.1	24.9	25.5

Table 3: Coverage (C), precision (P), recall (R), and F_1-measure (F) of WSD methods on the Senseval 3 dataset, polysemous words only, excluding adverbs, without using the first-sense heuristic. The results are percentages.

Conclusions

- Better results with SR (WordNet), instead of PMI (co-occurence).
  - We are currently evaluating the performance of PMI using sense tagged corpora
  - Sense pruning improves time performance without significantly affecting WSD performance
- In Senseval2 we outperform graph based unsupervised WSD methods (SAN and PageRank)
- In Senseval3 we performed comparable to SAN, but worst than PageRank.
  - Higher polysemy than Senseval2.
  - SAN and PageRank create bigger graphs than our method.
- Almost 100% coverage
  - We may probably compare all methods in 100% coverage, by forcing other methods to give an answer in all cases (without using the First Sense heuristic)

Next steps

- **PMI** in sense-tagged corpora
  - But then we will be supervised
- **Test with other similarity measures** or combinations.
  - **χ^2**, **likelihood ratio**, **LSA**, …
  - Our model works with **any relatedness measure**
- More evaluation datasets: Semeval 2007, 2010
- **LP relaxations**, in order to use Simplex
  - **Faster solution**: real time/scale implementations
  - **WSD in paragraph or section level**

Thank you !

Questions ?

Iraklis Varlamis

varlamis@hua.gr