An Experimental Study on Unsupervised Graph-based Word Sense Disambiguation

George Tsatsaronis
Web: http://www.idi.ntnu.no/~gbt/
e-mail: gbt@idi.ntnu.no

NTNU
Norwegian University of Science and Technology

Iraklis Varlamis
Web: http://www.dit.hua.gr/~varlamis/
e-mail: varlamis@hua.gr

HAROKOPION UNIVERSITY

Kjetil Nørvåg
Web: http://www.idi.ntnu.no/~noervaag/
e-mail: Kjetil.Norvag@idi.ntnu.no

NTNU
Norwegian University of Science and Technology

Presentation Layout

- Introduction and Motivation
- Contributions
- Unsupervised Graph-based Word Sense Disambiguation
  - Semantic Networks Representation
  - Techniques
- Experimental Evaluation
  - Unsupervised Techniques
  - Level of Inter-agreement
  - Comparison with State of the Art
- Conclusions and Future Directions

Introduction: The WSD task

- Assign to every **word** of a document the most appropriate meaning (**sense**) among those offered by a **lexicon** or a **thesaurus**.
  - 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?

- Polysemous and monosemous words in Senseval.

	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]

Motivation

- Several options in applying WSD:
  - Unsupervised
    - High coverage, lower accuracy than supervised, no need for manually annotated data set, low complexity
  - Supervised
    - Lower coverage than unsupervised, higher accuracy, "*knowledge acquisition bottleneck*", higher complexity
- Graph-based Unsupervised WSD
  - Truncated the accuracy gap from supervised
  - Map words and senses to semantic graphs
  - Research Questions:
    - How to construct such graphs, and how to process them?
    - What are the benefits from each processing technique?

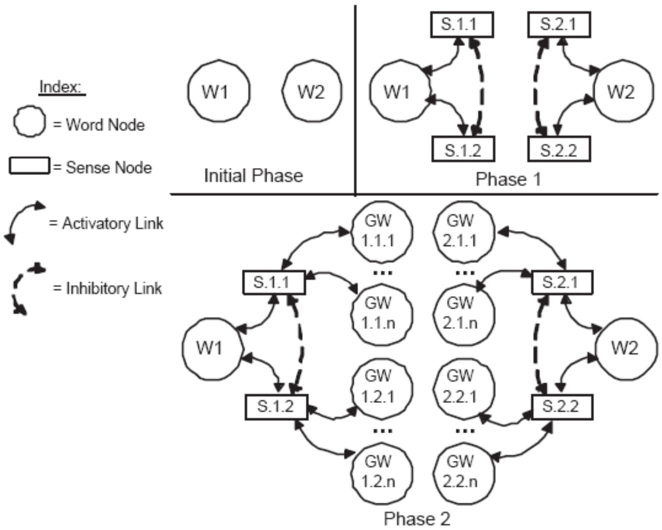
Contributions

- Experimental Evaluation of Unsupervised Graph-based WSD
  - uniform semantic graph-based representation
  - evaluate alternatives
    - Spreading of Activation
    - PageRank
    - HITS
    - P-Rank
  - study space and time complexity
  - analyze inter-agreement at the sense level selection
  - generalize comparison with SoA WSD techniques

Unsupervised Graph-based WSD

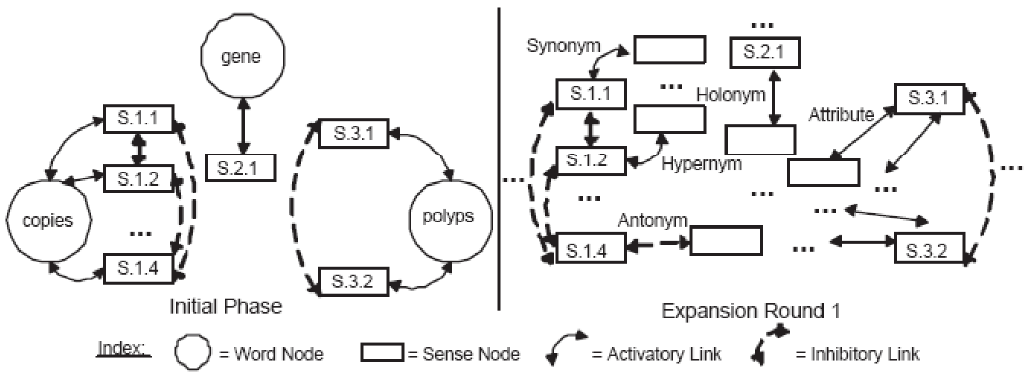
- Graph-based methods demonstrate SoA results among unsupervised WSD methods [Sinha and Mihalcea, 2007].
- An example of an earlier approach: [Veronis and Ide, 1990]

- Spreading of activation (social network analysis) to process the network.

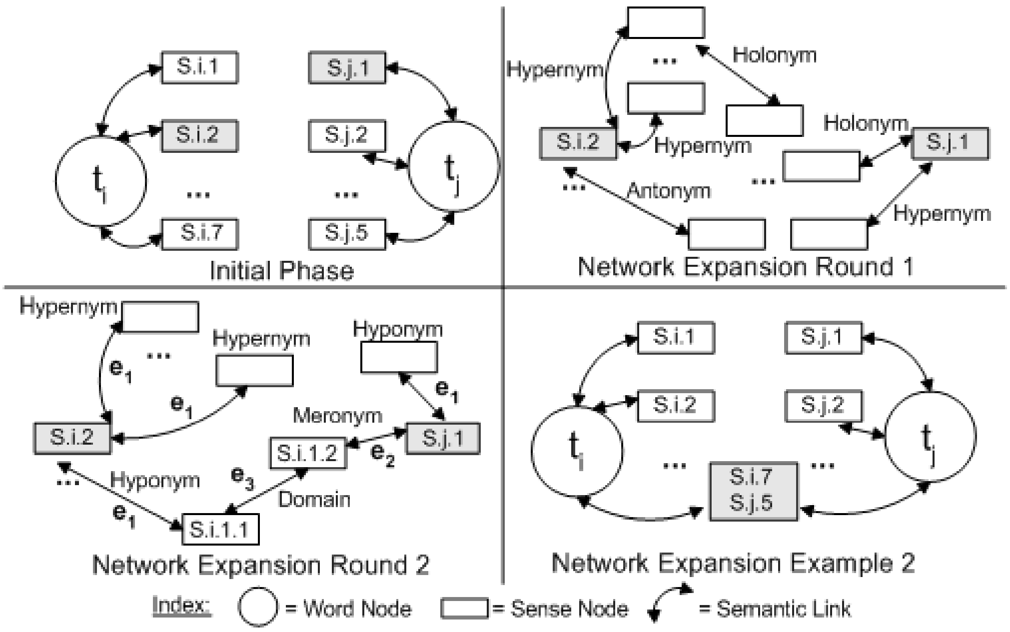


Semantic Networks Creation

- [Tsatsaronis et al., 2007] proposed a new method for constructing semantic networks
  - Use all of the available semantic information from WN
  - Use edges weighting scheme
  - Example: "If both copies of a certain gene were knocked out, benign polyps would develop"



General Example

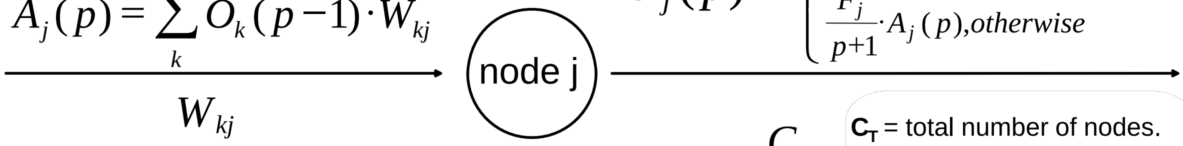


Use of Semantic Networks

- Semantic similarity/relatedness [Budanitsky and Hirst, 2006]
- *Omiotis* measure [Tsatsaronis et al., 2010]
  - Relatedness computation between:
    - Term pairs
    - Sentence pairs
- Publicly available: http://omiotis.hua.gr
- Currently the best lexicon-based measure of semantic relatedness

Spreading of Activation: Weight and Control [IJCAI 2007]

$$A_j(p) = \sum_k O_k(p-1) \cdot W_{kj}$$



$$O_j(p) = \begin{cases} 0, A_j(p) < \tau \\ \frac{F_j}{p+1} \cdot A_j(p), otherwise \end{cases}$$

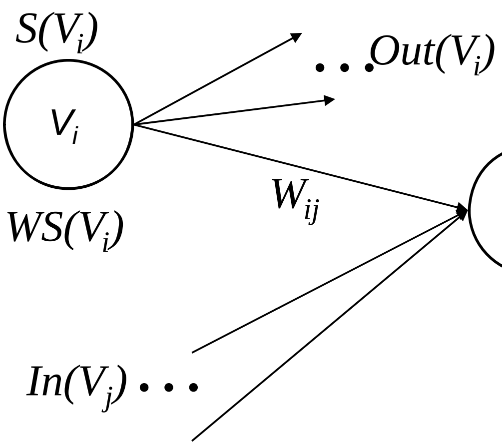
$$F_j = (1 - \frac{C_j}{C_T})$$

C_T = total number of nodes.

C_j = number of nodes directly connected to j via directed edges from j.

- Distance and fan-out constraints implemented to control activation flow
- Control the activation based on: [Crestani, 1997]

PageRank

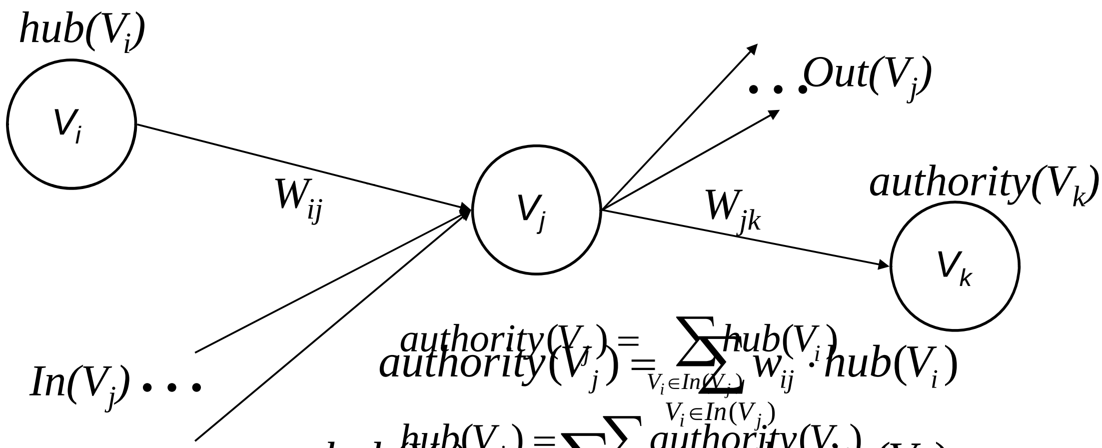


$$S(V_j) = (1-d) + d \cdot \sum_{i \in In(V_j)} \frac{S(V_i)}{|Out(V_i)|}$$

$$WS(V_j) = (1-d) + d \cdot \sum_{V_i \in In(V_j)} \frac{w_{ij} \cdot WS(V_i)}{\sum_{V_k \in Out(V_i)} w_{ik}}$$

- [Brin and Page, 1998]
- [Mihalcea et al., 2004]

HITS

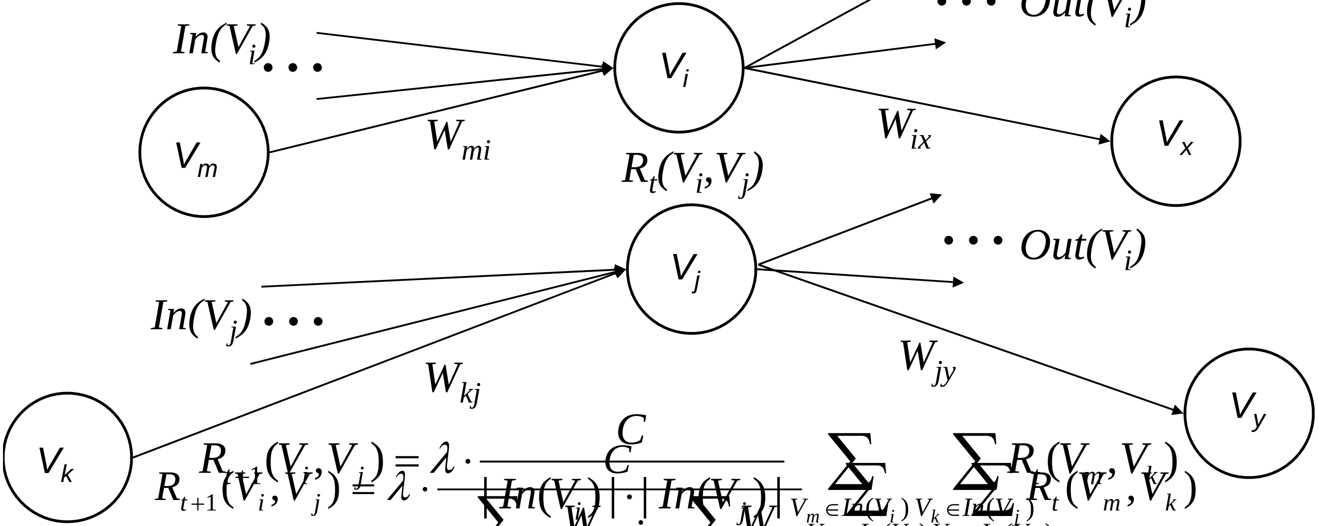


$$authority(V_j) = \sum_{V_i \in In(V_j)} w_{ij} \cdot hub(V_i)$$

$$hub(V_j) = \sum_{V_k \in Out(V_j)} w_{jk} \cdot authority(V_k)$$

- [Kleinberg, 1999]

P-Rank



$$R_{t+1}(V_i,V_j) = \lambda \cdot \frac{C}{\sum_{V_m \in In(V_i)} W_{mi} \cdot \sum_{V_k \in In(V_j)} W_{kj}} \sum_{V_m \in In(V_i)} \sum_{V_k \in In(V_j)} R_t(V_m,V_k)$$

$$+ (1-\lambda) \cdot \frac{C}{\sum_{V_x \in Out(V_i)} W_{ix} \cdot \sum_{V_y \in Out(V_j)} W_{jy}} \sum_{V_x \in Out(V_i)} \sum_{V_y \in Out(V_j)} R_t(V_x,V_y)$$

- [Zhao et al., 2009]

Sense Selection

- Per Word Node:
  - **SAN:** The most active sense node after activation ceases spreading
  - **PageRank:** The sense node with the highest PageRank score
  - **HITS:** The sense node with the highest authority score
  - **P-Rank:** The sense node with the highest similarity to the respective word node

Complexity Comparison

	Space	Time (Network Creation + Execution)
SAN	$O(n^2 \cdot k^{2l+3})$	$O(n \cdot k^{l+1}) + O(n^2 \cdot k^{2l+3})$
PageRank (PR)	$O(n^2 \cdot k^{2l+3})$	$O(n \cdot k^{l+1}) + O(n^2 \cdot k^{\frac{3}{2}l+3})$
HITS	$O(n^2 \cdot k^{2l+3})$	$O(n \cdot k^{l+1}) + O(n^2 \cdot k^{\frac{3}{2}l+3})$
P-Rank	$O(n^2 \cdot k^{2l+3})$	$O(n \cdot k^{l+1}) + O(n^4)$

Experimental Evaluation

Method	Senseval 2 N	Senseval 2 V	Senseval 2 Adj.	Senseval 2 All	Senseval 3 N	Senseval 3 V	Senseval 3 Adj.	Senseval 3 All
SAN	53.9	31.7	59.0	49.5	50.8	36.5	58.0	46.8
PR	69.5	37.2	59.0	58.8	61.8	47.3	60.6	56.7
HITS	69.1	36.6	59.1	58.3	69.2	40.4	66.7	57.4
P-Rank	51.3	27.31	57.4	45.6	60.6	29.9	67.8	52.1
Mih05	57.5	36.5	56.7	52.0	*n/a*	*n/a*	*n/a*	51.8
Agi09	70.4	38.9	58.3	59.5	64.1	46.9	62.6	57.4
Nav07	*n/a*	*n/a*	*n/a*	*n/a*	61.9	36.1	62.8	52.5
FS	74.0	42.4	63.1	63.7	70.9	50.7	59.7	61.3

- SAN, PR and HITS show stable performance for all POS in both data sets
- P-Rank: More unstable and usually significantly lower performance
- All unsupervised methods lose by the First Sense heuristic but have narrowed the gap.

Inter-Agreement

Pair	Senseval 2 N	Senseval 2 V	Senseval 2 Adj.	Senseval 2 All	Senseval 3 N	Senseval 3 V	Senseval 3 Adj.	Senseval 3 All
SAN - PR	51.51	35.74	54.16	47.86	53.17	49.48	49.83	51.21
SAN - HITS	52.42	23.89	57.55	39.51	50.6	40.38	50.16	46.68
SAN - P-Rank	50.84	27.16	63.46	46.77	66.52	32.94	69.04	55.37
PR - HITS	62.56	34.93	64.32	55.54	60.36	44.64	66.88	55.57
PR - P-Rank	50.55	30.95	67.3	48.1	68.2	30.58	71.42	55.78
HITS - P-Rank	53.88	23.8	59.61	46.83	67.78	31.76	69.04	54.17

- Inter-agreement in all cases always lower than 70%
- Very low inter-agreement in the VERB POS
- Evaluating the union of the correct assignments for method pairs:
  - SAN-PR leads to an upper bound of 69.73% in Senseval 2 and 63.36% in Senseval 3.
  - Similar findings with other method pairs.

Overall Comparison with SoA

- SenseLearner: [Mihalcea and Csomai, 2005]
- Simil-Prime: [Kohomban and Lee, 2005]
- SSI: [Navigli, 2006]
- WE: [Hoste et al., 2002]

Dataset	SenseLearner	Simil-Prime	SSI	WE	FS	PR	HITS	Agi09
Senseval2	64.82	65.00	n/a	63.2	63.7	58.8	58.3	59.5
Senseval3	63.01	65.85	60.4	n/a	61.3	56.7	57.4	57.4

- Unsupervised methods have narrowed the gap from supervised to almost 8%
- State of the art supervised methods have limitations:
  - Simil-Prime resides to the FS for the disambiguation of adjectives and adverbs
  - Usually bounded to words that have previously been seen in the training corpus
  - FS performs well in Senseval 2 and 3, but in domain-specific data sets, it might need re-training

Conclusions

- Unsupervised Graph-based WSD methods are now closer in performance to supervised methods
- They usually present low inter-agreement rate (i.e., lower than 70%)
- An ensemble of those approaches can boost performance
- Rich thesauri like WordNet offer the opportunity to create semantic networks across POS and allow for many options in graph-based techniques

Future Directions

- Combine lexical resources to enrich the semantic representation (i.e., YAGO)
  - This may affect the graph creation method
- Design ensembles of graph-based methods
  - Take advantage of the relatively low inter-agreement rate
  - New ensemble strategies: learn to select the most proper WSD method, rather than the most proper sense
- Unsupervised Domain-biased WSD
  - This may affect both graph creation and processing

Questions

Thank you very much for your attention!

Thank you Alexander, Corina, and the whole local organizing team for a wonderful CICLing 2010!

Questions/Comments?

References

- [Agirre and Soroa, 2009] Agirre, E. and Soroa, A.. Personalizing pagerank for word sense disambiguation. In *Proc. of EACL*, pages 33–41, 2009.
- [Brin and Page, 1998] Brin, S. and Page, L.. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30:1–7, 1998.
- [Budanitsky et al. 2006] Budanitsky, A. and Hirst, G, *Evaluating WordNet-based Measures of Lexical Semantic Relatedness*, Computational Linguistics, 32(1), pp. 13-47,, 2006.
- [Crestani, 1997] Crestani, F., Application of spreading activation techniques in information retrieval. Artificial Intelligence Review, 11:453–482, 1997.
- [Haliday and Hasan, 1976] Halliday, M. and Hasan, R., Cohesion in English. Longman 1976.
- [Hoste et al., 2002] V. Hoste, W. Daelemans, I. Hendrickx, and A. van den Bosch. Evaluating the results of a memory-based word-expert approach to unrestricted word sense disambiguation. In *Proc. of the ACL Workshop on Word Sense Disambiguation*, 2002.
- [Kleinberg, 1999] Kleinberg, J., Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999
- [Kohomban and Lee, 2005] Kohomban, U. and Lee, W.. Learning semantic classes for word sense disambiguation. In *Proc. of ACL*, pages 34–41, 2005.
- [Mihalcea and Csomai, 2005] Mihalcea, R. and Csomai, A.. Senselearner: Word sense disambiguation for all words in unrestricted text. In *Proc. of ACL*, pages 53–56, 2005.
- [Mihalcea, 2005] Mihalcea, R.. Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling. In *HLT*, 2005
- [Mihalcea et al., 2004] Mihalcea, R., Tarau, P., and Figa, E.. Pagerank on semantic networks with application to word sense disambiguation. In *Proc. of COLING*, 2004
- [Navigli, 2006] Navigli, R. Online word sense disambiguation with structural semantic interconnections. In *Proc. of EACL*, 2006.
- [Navigli and Lapata, 2007] Navigli, R. and Lapata, M.. Graph connectivity measures for unsupervised word sense disambiguation. In *Proc. of IJCAI*, pages 1683–1688, 2007.
- [Sinha and Mihalcea, 2007] Sinha, R. and Mihalcea, R., Unsupervised graph-based word sense disambiguation using measures of word semantic similarity. In *Proc. of ICSC*, 2007
- [Snyder and Palmer, 2004]. Snyder, B. and Palmer, M., The english all-words task. In *Proc. of Senseval-3*, pages 41–43, 2004.
- [Tsatsaronis et al., 2007] Tsatsaronis, G., Vazirgiannis, M., and Androutsopoulos, I.. Word sense disambiguation with spreading activation networks generated from thesauri. In *Proc. of IJCAI*, pages 1725–1730, 2007.
- [Tsatsaronis et al., 2010] Tsatsaronis, G., Varlamis, I., and Vazirgiannis, M., "Text Relatedness Based on a Word Thesaurus", JAIR, vol. 37, pp. 1-39.
- [Veronis and Ide, 1990] Veronis, J. and Ide, N., *Word Sense Disambiguation with very large neural networks extracted from machine readable dictionaries*, COLING 1990.
- [Zhao et al., 2009] Zhao, P., Han, J. and Sun, Y.. P-Rank: a comprehensive structural similarity measure over information networks. In *Proc. of CIKM*, pages 553–562, 2009.