Random Walks for Knowledge-Based Word Sense Disambiguation

Qiuyu Li

Word Sense Disambiguation

1 Supervised

- using labeled training sets
(features and proper sense label)

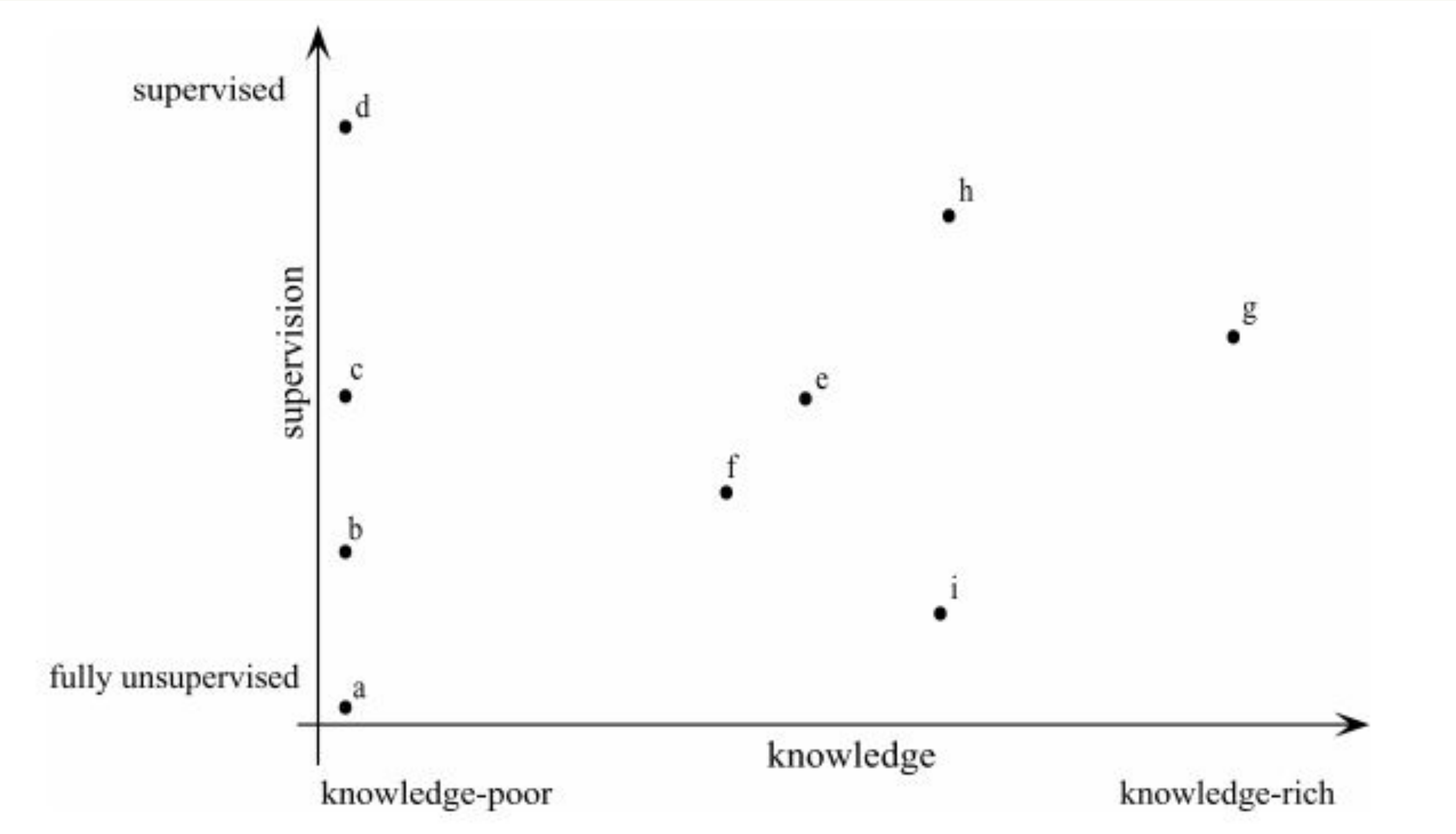
2 Unsupervised

- only use unlabeled corpora without the sense-tagged corpus

3 Knowledge-based

- external lexical resources
(such as machine-readable dictionaries, thesauri and ontologies)

Supervision vs. Knowledge



Overview

1 Introduction

2 Knowledge-based Word Sense Disambiguation (WSD)

3 Lexical Knowledge Bases (LKB) - WordNet

4 Random Walks - PageRank & Personalized PageRank

5 Evaluation

6 Issues & Future Directions

7 Conclusions

Knowledge-based WSD

1 Overlap of sense definitions

- traditional approach, called gloss overlap or the Lesk algorithm

2 Selectional restrictions

- uses selectional preferences to constrain the meanings of a target word in the specific context.

3 Structural approaches

a) similarity measures

- local context

b) graph-based methods

- global context
- lexical chains (eat -> dish -> vegetable -> potato)

WordNet

Synset (each one represents a distinct concept)

- groups nouns, verbs, adjectives and adverbs into sets of synonyms
- over 117,000 synsets

e.g.

<coach#n1, manager#n2, handler#n3>
<coach#n2, private instructor#n1, tutor#n1>
<coach#n3, passenger car#n1, carriage#n1>
<coach#n4, four-in-hand#n2, coach-and-four#n1>
<coach#n5, bus#n1, autobus#n1, charabanc#n1, double-decker#n1, jitney#n1... >
<coach#v1, train#v7>
<coach#v2>

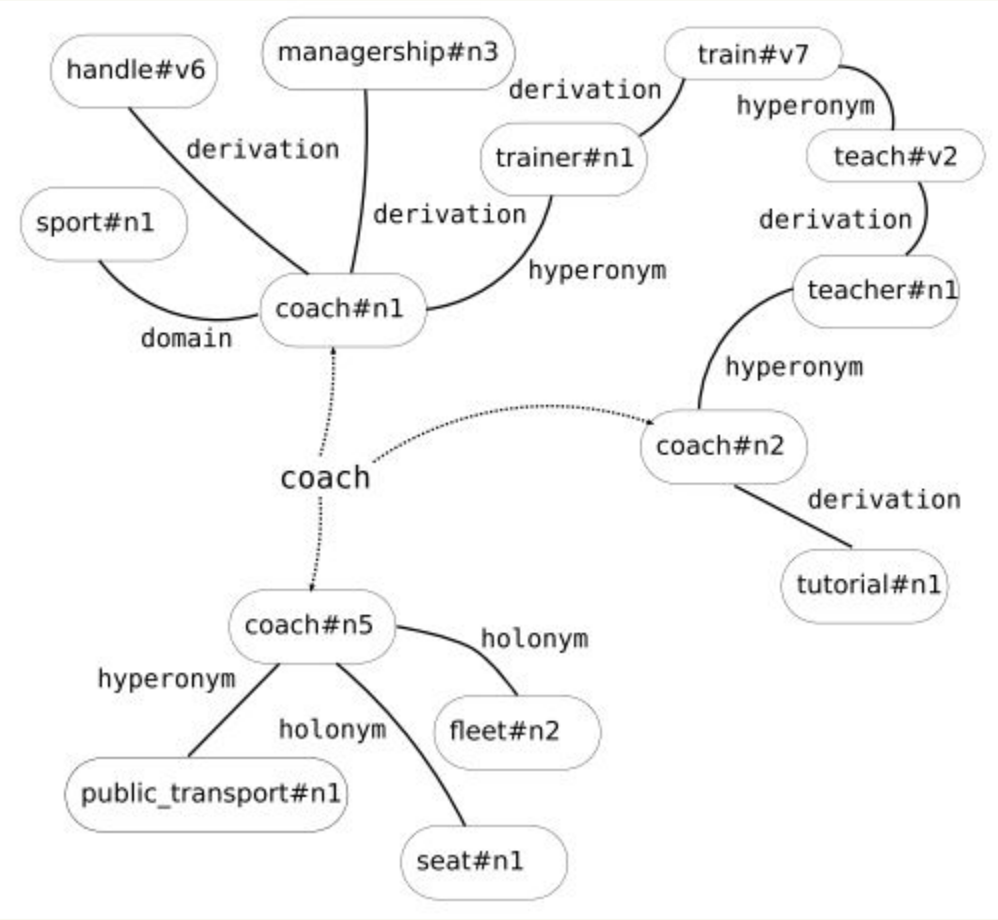
Represent WordNet as a Graph

Dictionary

- Word lemmas linked to the corresponding senses

Concepts and relations

Graph G=(V, E)
- V is the set of nodes
each node represents one sense
- E is the set of edges
each relation between two senses is represented by an edge.



Random Walk - PageRank

1 Undirected relations between concepts

- symmetric and have inverse counterpart

2 PageRank Random Walk algorithm

- ranks the vertices in a graph in terms of structural relations
- vertex v_i -> v_j, a vote from node i to j, the contribution of node i depends on the i’s rank
- final rank of node i represents the probability of a random walk over the graph ending on node i

Random Walk - PageRank

Given a graph G with N vertices $\{v_1, ..., v_N\}$

d_i - the outdegree of node i

M - N×N transition probability matrix, where

$$M_{ji} = \begin{cases} \frac{1}{d_i} & \text{if a link from i to j exists,} \\ 0 & \text{otherwise.} \end{cases}$$

PageRank Vector **P** over G is calculated by

$$\mathbf{P} = cM\mathbf{P} + (1 - c)\mathbf{v}$$

v - N×1 random vector (initial)

c - damping factor, $c \in [0,1]$, experimentally, $c \in [0.85, 0.95]$

c**MP** - the voting scheme

(1-c)**v** - the probability a random jump (not following any paths)

smoothing factor

Personalized PageRank - PPR

$$\mathbf{P} = cM\mathbf{P} + (1 - c)\mathbf{v}$$

Traditional/Static PageRank

- using uniform vector **v** with all the element values 1/N

Personalized PageRank

- using un-uniform vector **v** (modified)
- assigning v with different initial values makes PageRank algorithm more effective (spreads along the graph during iterations)

Personalized PageRank - PPR

1 Static PageRank (STATIC)

- context-independent ranking (baseline)

2 Personalized PageRank (PPR)

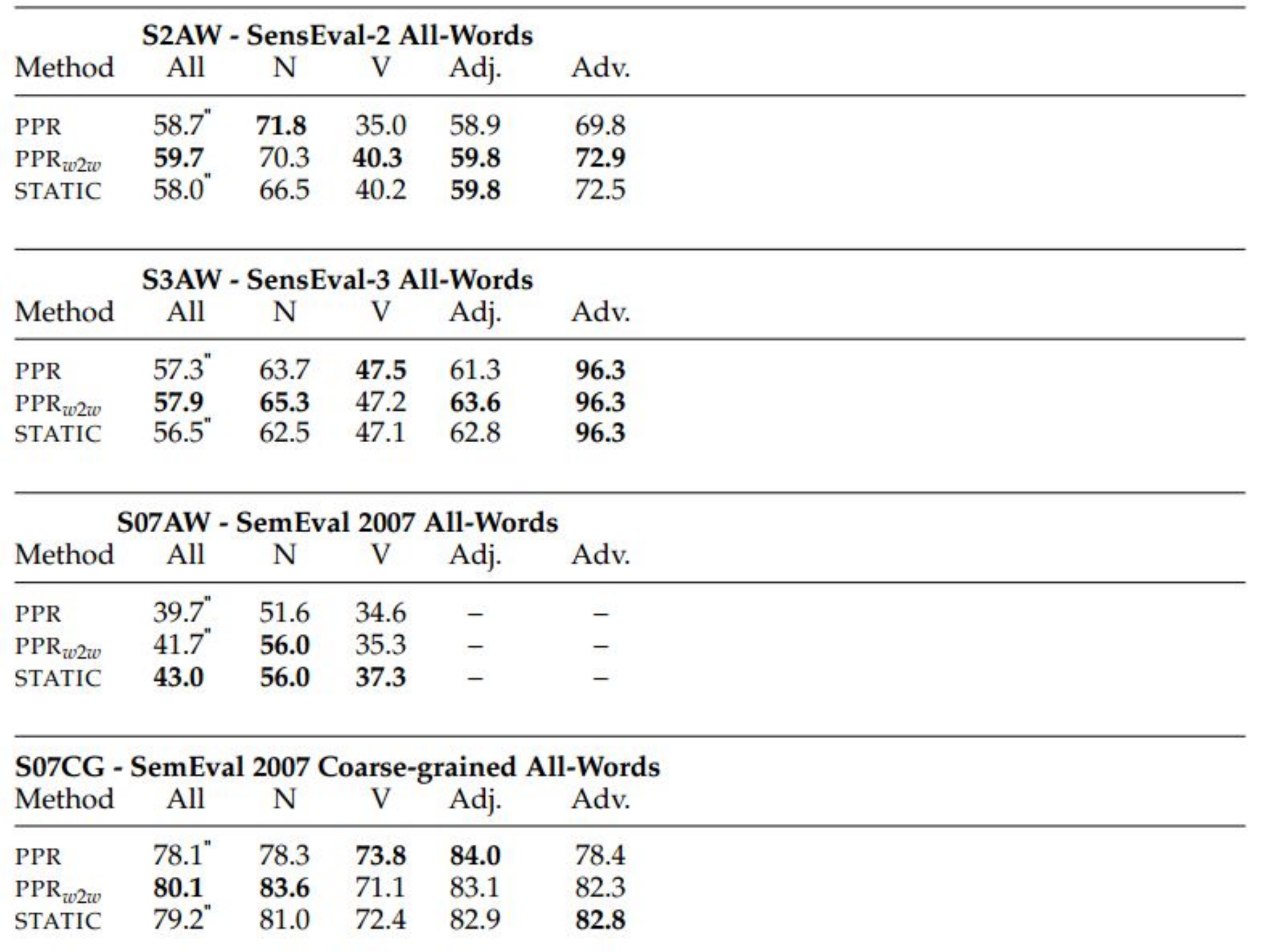
- relate content words to WordNet concepts
- every concept receives a score

3 Word-to-word Heuristic (PPR_{w2w})

- run Personalized PageRank separately for each target word in the context
- let surrounding words determine the most relavent sense (avoid the influence comes from the target word)

PPR_{w2w} does not disambiguate all target words of the context in a single run, which makes it less efficient

Evaluation - F1 over different Datasets



S2AW - SensEval-2 All-Words

Method	All	N	V	Adj.	Adv.
PPR	58.7$^{''}$	**71.8**	35.0	58.9	69.8
PPR$_{w2w}$	**59.7**	70.3	**40.3**	**59.8**	**72.9**
STATIC	58.0$^{''}$	66.5	40.2	**59.8**	72.5

S3AW - SensEval-3 All-Words

Method	All	N	V	Adj.	Adv.
PPR	57.3$^{''}$	63.7	**47.5**	61.3	**96.3**
PPR$_{w2w}$	**57.9**	**65.3**	47.2	**63.6**	**96.3**
STATIC	56.5$^{''}$	62.5	47.1	62.8	**96.3**

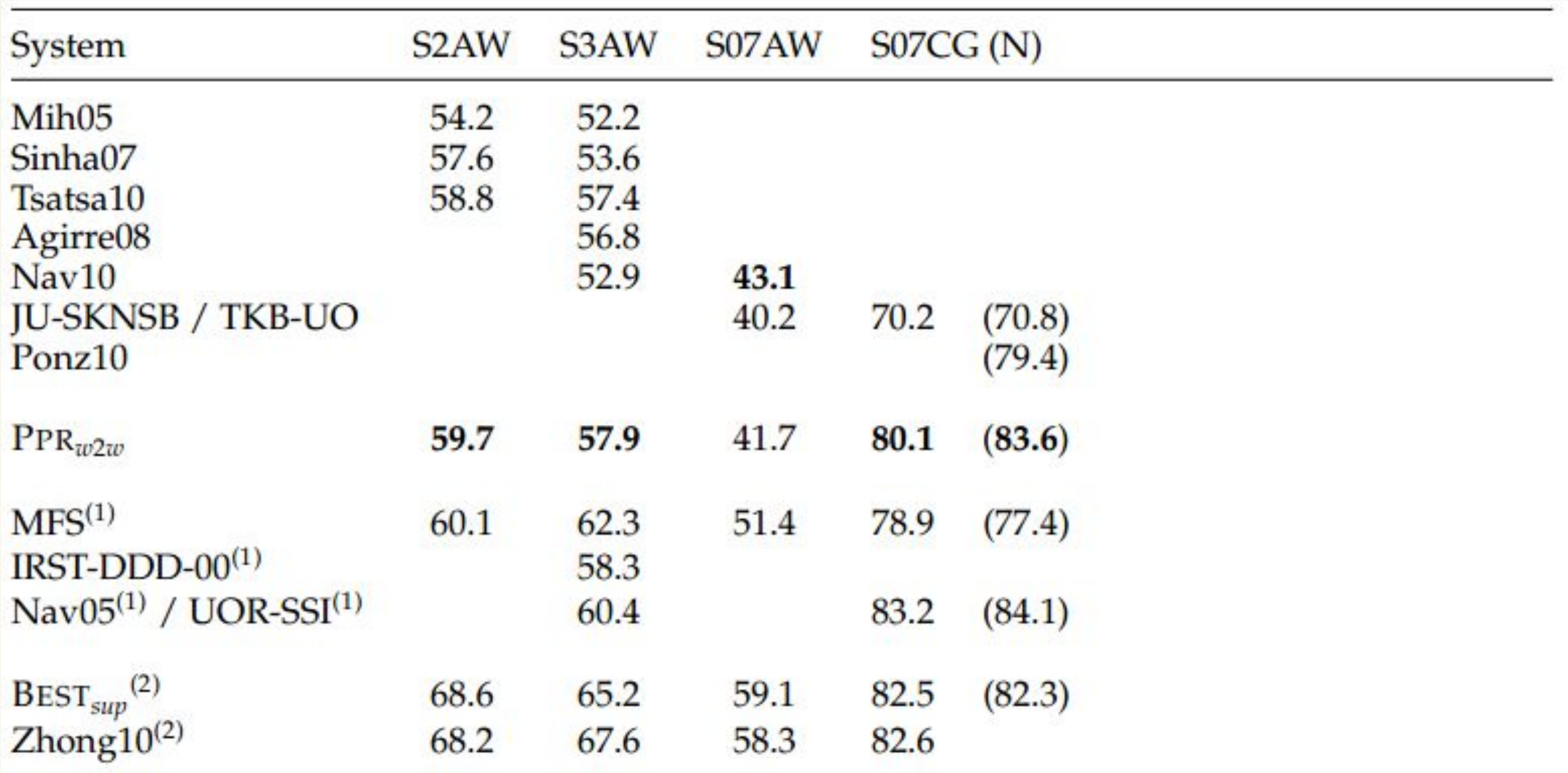
S07AW - SemEval 2007 All-Words

Method	All	N	V	Adj.	Adv.
PPR	39.7$^{''}$	51.6	34.6	–	–
PPR$_{w2w}$	41.7$^{''}$	**56.0**	35.3	–	–
STATIC	**43.0**	**56.0**	**37.3**	–	–

S07CG - SemEval 2007 Coarse-grained All-Words

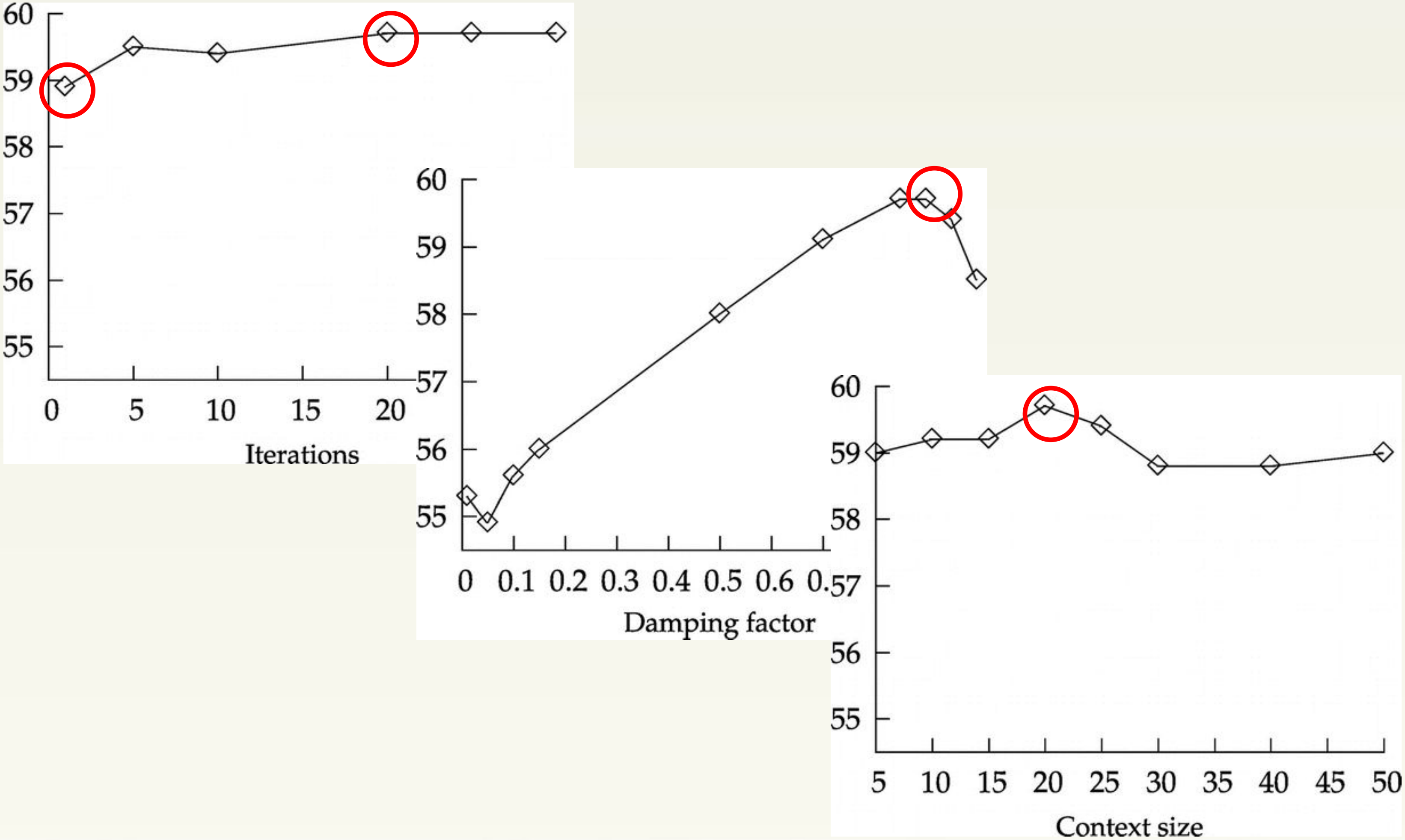
Method	All	N	V	Adj.	Adv.
PPR	78.1$^{''}$	78.3	**73.8**	**84.0**	78.4
PPR$_{w2w}$	**80.1**	**83.6**	71.1	83.1	82.3
STATIC	79.2$^{''}$	81.0	72.4	82.9	**82.8**

Evaluation - with other Systems

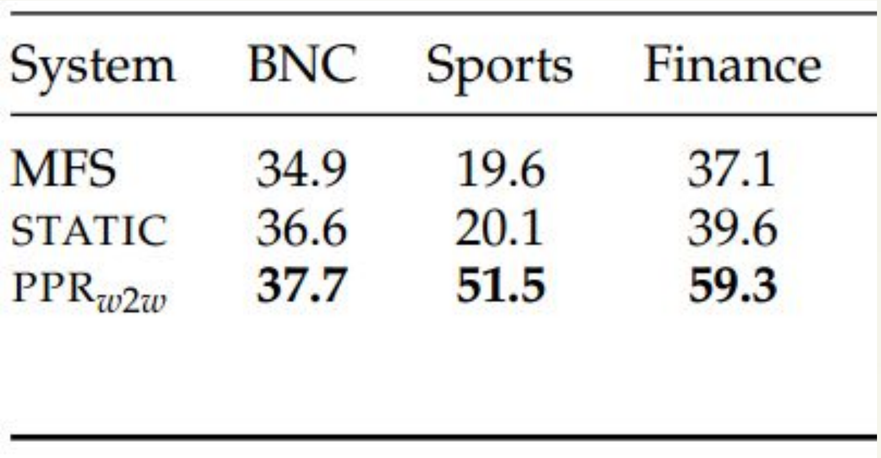


System	S2AW	S3AW	S07AW	S07CG	(N)
Mih05	54.2	52.2			
Sinha07	57.6	53.6			
Tsatsa10	58.8	57.4			
Agirre08		56.8			
Nav10		52.9	**43.1**		
JU-SKNSB / TKB-UO			40.2	70.2	(70.8)
Ponz10					(79.4)
PPR$_{w2w}$	**59.7**	**57.9**	41.7	**80.1**	**(83.6)**
MFS(1)	60.1	62.3	51.4	78.9	(77.4)
IRST-DDD-00(1)		58.3			
Nav05(1) / UOR-SSI(1)		60.4		83.2	(84.1)
BEST$_{sup}$(2)	68.6	65.2	59.1	82.5	(82.3)
Zhong10(2)	68.2	67.6	58.3	82.6	

Evaluation - PageRank Parameters

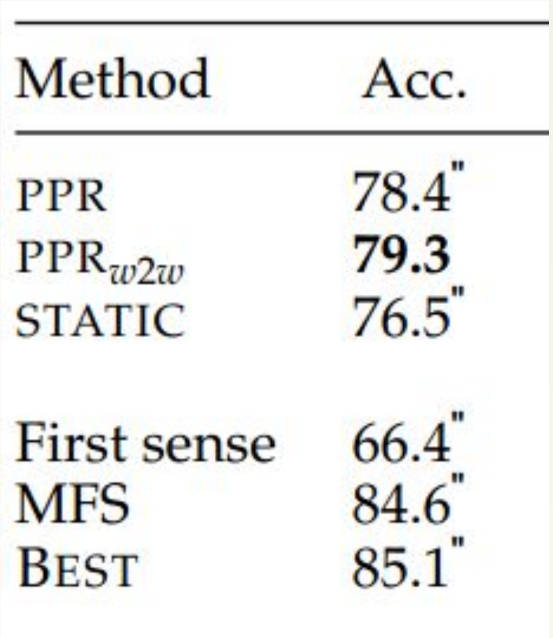


Evaluation - Domain Specific & Spanish



System	BNC	Sports	Finance
MFS	34.9	19.6	37.1
STATIC	36.6	20.1	39.6
PPR_{w2w}	**37.7**	**51.5**	**59.3**

General-domain: British National Corpus (BNC)
Domain-specific: Sports & Finance corpora



Method	Acc.
PPR	78.4"
PPR_{w2w}	**79.3**
STATIC	76.5"
First sense	66.4"
MFS	84.6"
BEST	85.1"

Other Evaluations

Results on English data sets (F1)
Comparison to State-of-the-Art Systems
Comparison with Related Algorithms
PageRank Parameters
Size of Context Window
Using Different WordNet Versions
Using xwn vs. WN3.0 Gloss Relations
Analysis of relation types
Correlation between systems, gold tags, and MFS
Results on three subcorpora(BNC, Sports & Finance corpora)
Combination with MFS (F1)
Efficiency of Full Graphs vs. Subgraphs
Experiments on Spanish

Issues & Future Directions

1 "Knowledge acquisition bottleneck"
- Automatic enrichment of knowledge resources

2 Global weights of the edges in the random walk calculations

3 Combine PPR with other WordNet related resources

Conclusions

1 Knowledge-based WSD based on random walks

- over relations in a LKB (WordNet)

2 Full Graph of WordNet

3 PageRank & Personalized PageRank (PPR)

- Static PageRank **(STATIC)**
- Personalized PageRank **(PPR)**
- Word-to-word Heuristic **($\mathbf{PPR_{w2w}}$)**

4 Other Language - Spanish

- only requirement of having a WordNet

5 Reproducible Experiments

THANK YOU

Any Questions?

References

1 E. Agirre, O. L. de Lacalle, and A. Soroa, “Random walks for knowledge-based word sense disambiguation,”Computational Linguistics, vol. 40, no. 1, pp. 57–84, 2014.
2 R. Navigli, “Word sense disambiguation: A survey,” ACM Computing Surveys (CSUR), vol. 41, no. 2, p. 10, 2009.