Evaluating Automatic Approaches for Word Meaning Discovery and Disambiguation using Lexical Substitution

Diana McCarthy[1] Roberto Navigli[2]

[1]University of Sussex, UK

[2]University of Rome "La Sapienza", Italy

The 16th Nordic Conference of Computational Linguistics

Word Meanings and Evaluation

Word Meanings and Evaluation

- word meaning is important for semantic interpretation

Word Meanings and Evaluation

- word meaning is important for semantic interpretation
- what is the right representation to use?

Word Meanings and Evaluation

- word meaning is important for semantic interpretation
- what is the right representation to use?
- how can we compare inventories of word meaning?

Word Meanings and Evaluation

- word meaning is important for semantic interpretation
- what is the right representation to use?
- how can we compare inventories of word meaning?
- the meaning of a word depends on the context

Word Meanings and Evaluation

- word meaning is important for semantic interpretation
- what is the right representation to use?
- how can we compare inventories of word meaning?
- the meaning of a word depends on the context
- we need to find the right meaning in a given context

Word Meanings and Evaluation

- word meaning is important for semantic interpretation
- what is the right representation to use?
- how can we compare inventories of word meaning?
- the meaning of a word depends on the context
- we need to find the right meaning in a given context
- most work on disambiguation uses pre-defined man-made inventory

Word Meanings and Evaluation

- word meaning is important for semantic interpretation
- what is the right representation to use?
- how can we compare inventories of word meaning?
- the meaning of a word depends on the context
- we need to find the right meaning in a given context
- most work on disambiguation uses pre-defined man-made inventory
- how can we compare disambiguation techniques regardless of the inventory used?

Word Sense Disambiguation (WSD)

Given a word in context, find the correct "sense"

> Residents say militants in a station wagon pulled up , doused the building in gasoline , and struck a match.

Word Sense Disambiguation (WSD)

Given a word in context, find the correct "sense"

> Residents say militants in a station wagon pulled up , doused the building in gasoline , and struck a match.

Word Sense Disambiguation (WSD)

Given a word in context, find the correct "sense"

> Residents say militants in a station wagon pulled up , doused the building in gasoline , and struck a match.

match#n#1

Word Sense Disambiguation (WSD)

Given a word in context, find the correct "sense"

> After the match, replace any
> remaining fluid deficit to prevent
> problems of chronic dehydration
> throughout the tournament.

Word Sense Disambiguation (WSD)

Given a word in context, find the correct "sense"

After the match, replace any remaining fluid deficit to prevent problems of chronic dehydration throughout the tournament.

Word Sense Disambiguation (WSD)

Given a word in context, find the correct "sense"

After the match, replace any remaining fluid deficit to prevent problems of chronic dehydration throughout the tournament.

match#n#2

SENSEVAL Evaluation Series

- 1997 ACL-SIGLEX Initial Ideas for Standard Datasets for WSD Evaluation "Tagging Text with Lexical Semantics: Why What and How?"
- SENSEVAL 1998 SENSEVAL-2 2001 SENSEVAL-3 2004
- to be continued ... SemEval 2007
- increase in the range of languages
- man-made inventories used, especially WordNet

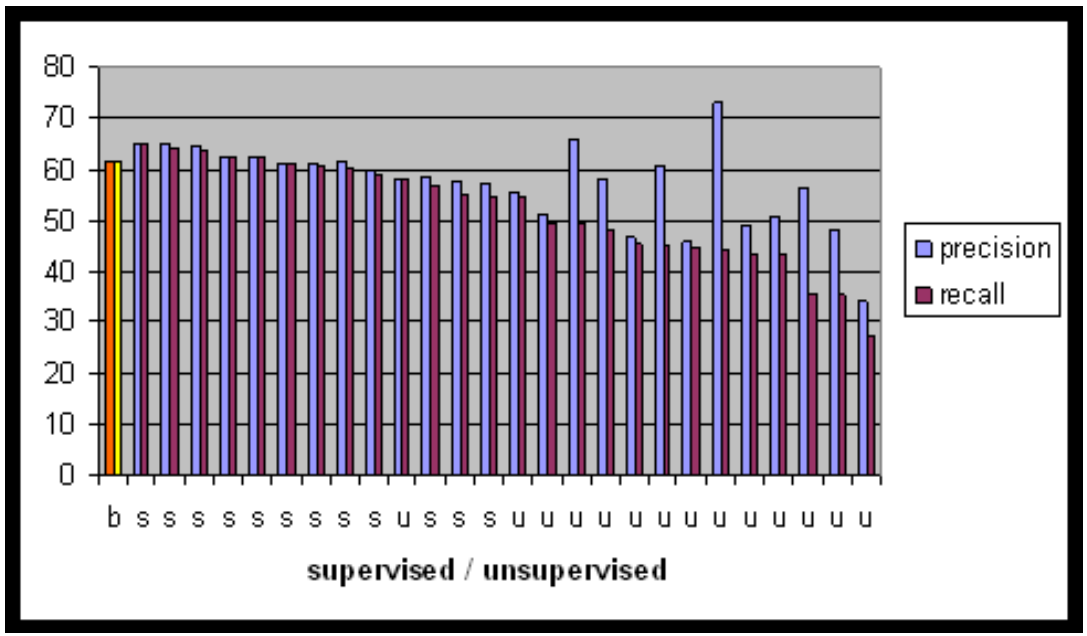
Research in Word Sense Disambiguation (WSD)

Rapid Growth since Inception of SENSEVAL

- 274,000 results Google search "word sense disambiguation"
- 1630 of papers on ACL Anthology
- 5 ACL-SIGLEX Workshops to date
- SemEval workshop ACL 2007

SENSEVAL Lessons

- supervised systems outperform "unsupervised"
- hand-labelled data is costly
- best systems performing just better than first sense heuristic over all words e.g. English all words SENSEVAL-3



Can This Level of Performance Benefit Applications?

- Enough context: WSD comes out in statistical wash
- not enough context and can't do anyway
- IR [Clough and Stevenson, 2004, Schütze and Pederson, 1995] vs [Sanderson, 1994]
- MT [Carpuat and Wu, 2005b, Carpuat and Wu, 2005a]

What is the Right Inventory?

- WordNet often used
- granularity is an issue
- but what is the right level of granularity?

match has 9 senses in WordNet including:-

- 1. match, lucifer, friction match – (lighter consisting of a thin piece of wood or cardboard tipped with combustible chemical; ignites with friction; "he always carries matches to light his pipe")
- 3. match – (a burning piece of wood or cardboard; "if you drop a match in there the whole place will explode")
- 6. catch, match – (a person regarded as a good matrimonial prospect)
- 8. couple, mates, match – (a pair of people who live together; "a married couple from Chicago")

What is the Right Inventory?

- many believe we need a coarse-grained level for WSD applications [Ide and Wilks, 2006] (though see [Stokoe, 2005])
- but what is the right way to group senses?

Example *child* WordNet

WNs#	gloss
1	a young person
2	a human offspring
3	an immature childish person
4	a member of a clan or tribe

- for MT use parallel corpora if know target languages
- what about summarising, paraphrasing QA, IR, IE?

What is the Right Inventory?

- many believe we need a coarse-grained level for WSD applications [Ide and Wilks, 2006] (though see [Stokoe, 2005])
- but what is the right way to group senses?

Example *child* WordNet SENSEVAL-2 groups

WNs#	gloss
1	a young person
2	a human offspring
3	an immature childish person
4	a member of a clan or tribe

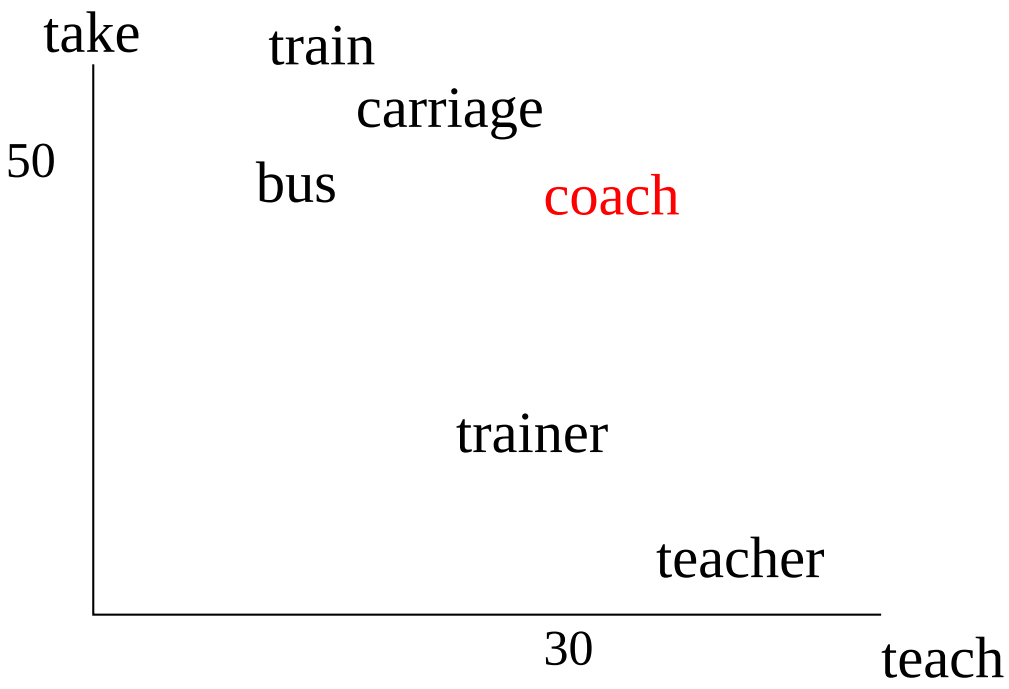
- for MT use parallel corpora if know target languages
- what about summarising, paraphrasing QA, IR, IE?

Distributional Approaches to Finding Word Meanings

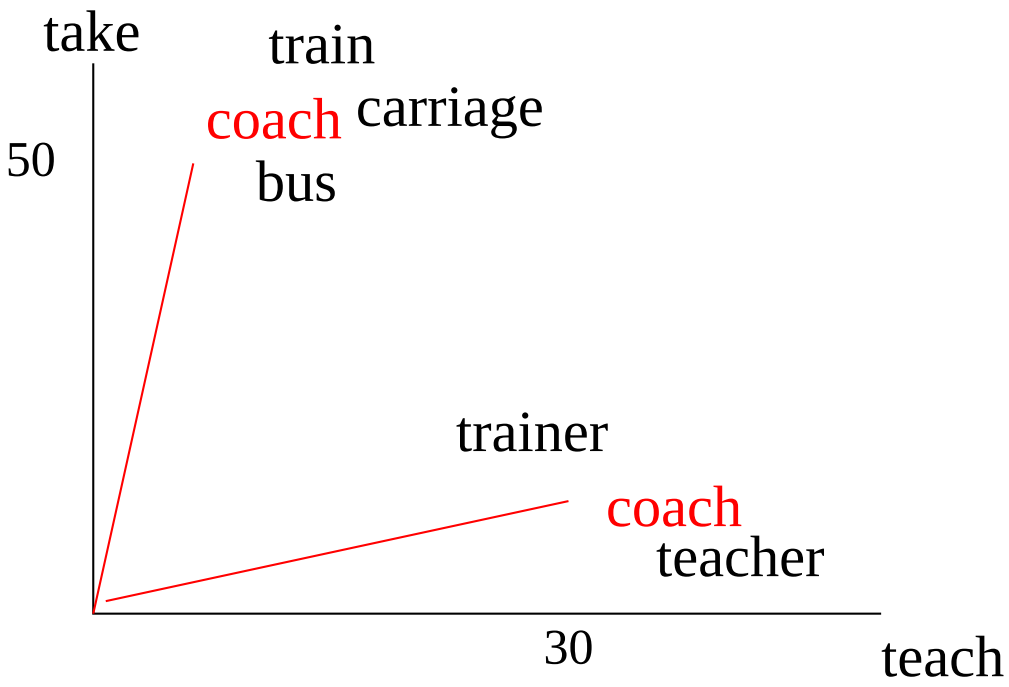
Vector based approaches from raw data e.g. coach

context	frequency		
	coach	*bus*	*trainer*
take	50	60	10
teach	30	2	25
ticket	8	5	0
match	15	2	6

Vector Based Approaches



Vector Based Approaches



Distributional Approaches from Parsed Data

context	frequency		
	coach	*bus*	*trainer*
object take	50	60	10
subject teach	31	2	25
subject instruct	10	0	15
object drive	25	35	10
noun modifier ticket	10	12	0

Output
Word: <closest word> <score> <2nd closest > <score>. . .

Distributional Approaches from Parsed Data

context	frequency		
	coach	*bus*	*trainer*
object take	50	60	10
subject teach	31	2	25
subject instruct	10	0	15
object drive	25	35	10
noun modifier ticket	10	12	0

Output

Word: <closest word> <score> <2nd closest > <score>. . .

coach: train 0.171 bus 0.166 player 0.149 captain 0.131 car 0.131

Distributional Approaches from Parsed Data

context	frequency		
	coach	*bus*	*trainer*
object take	50	60	10
subject teach	31	2	25
subject instruct	10	0	15
object drive	25	35	10
noun modifier ticket	10	12	0

Output

Word: <closest word> <score> <2nd closest > <score>. . .

coach: train 0.171 bus 0.166 player 0.149 captain 0.131 car 0.131

Grouping similar words [Pantel and Lin, 2002]

Outline

Key Issues

How can we:

- determine the distinctions useful for WSD systems?
- compare disambiguation techniques regardless of the inventories used?
- compare inventories of meaning?

Key Issues

How can we:

- determine the distinctions useful for WSD systems?
- compare disambiguation techniques regardless of the inventories used?
- compare inventories of meaning?

Our idea: lexical substitution

Lexical Substitution

Find a replacement word for a target word in context

Lexical Substitution

Find a replacement word for a target word in context

For example

The ideal preparation would be a light meal about 2-2 1/2 hours pre-match , followed by a warm-up hit and perhaps a top-up with extra fluid before the match.

Lexical Substitution

Find a replacement word for a target word in context

For example

The ideal preparation would be a light meal about 2-2 1/2 hours pre-match , followed by a warm-up hit and perhaps a top-up with extra fluid before the match.

The ideal preparation would be a light meal about 2-2 1/2 hours pre-match , followed by a warm-up hit and perhaps a top-up with extra fluid before the game.

Motivation

- evaluate methods of disambiguating word meanings
- inventory to be determined by task
- permit any inventory without requirement for mapping
- evaluate inventory as well as disambiguation
- task which has potential impact for applications
- no hand-labelled training data

SemEval

see http://nlp.cs.swarthmore.edu/semeval/tasks/index.shtml

- evaluation run during March
- results sent out in April
- Workshop at ACL Prague
- 18 tasks including:

- WSD tasks
- web people search
- affective text
- time event
- semantic relations between nominals
- word sense induction
- metonymy resolution

English Lexical Substitution Task Set Up

- 201 words (nouns, verbs, adjectives and adverbs)
- words selected
  - manually 70
  - automatically 131
- each word with 10 sentences
- 2010 sentences
- 300 trial set 1710 test set
- English Internet Corpus [Sharoff, 2006]
- sentences selected
  - manually for 20 words in each PoS
  - rest selected automatically

Annotators

- 5 native English speakers from the UK
- range of backgrounds
  - 3 some background in linguistics
  - 2 other backgrounds
- all subjects annotated the entire dataset

Instructions

- the substitute should preserve the meaning of the target word as much as possible
- use a dictionary or thesaurus if necessary
- supply up to 3 substitutes if they all fit the meaning equally well
- use NIL if you cannot think of a substitute
- pick a substitute that is close in meaning even if it doesn't preserve the meaning (aim for one that is more general)
- use a phrase if you can't think of a single word substitute
- use "name" for proper names
- indicate if the target word is an integral part of a phrase, and what the phrase is

The Annotation Interface

Please replace the word in bold with a substitute which preserves the meaning of the sentence:

Sentence #671:
The ideal preparation would be a light meal about 2-2 1/2 hours pre-match , followed by a warm-up hit and perhaps a top-up with extra fluid before the **match** .

Substitute: game OK

☐ nil ☐ extra responses ☐ name ☑ used a dictionary

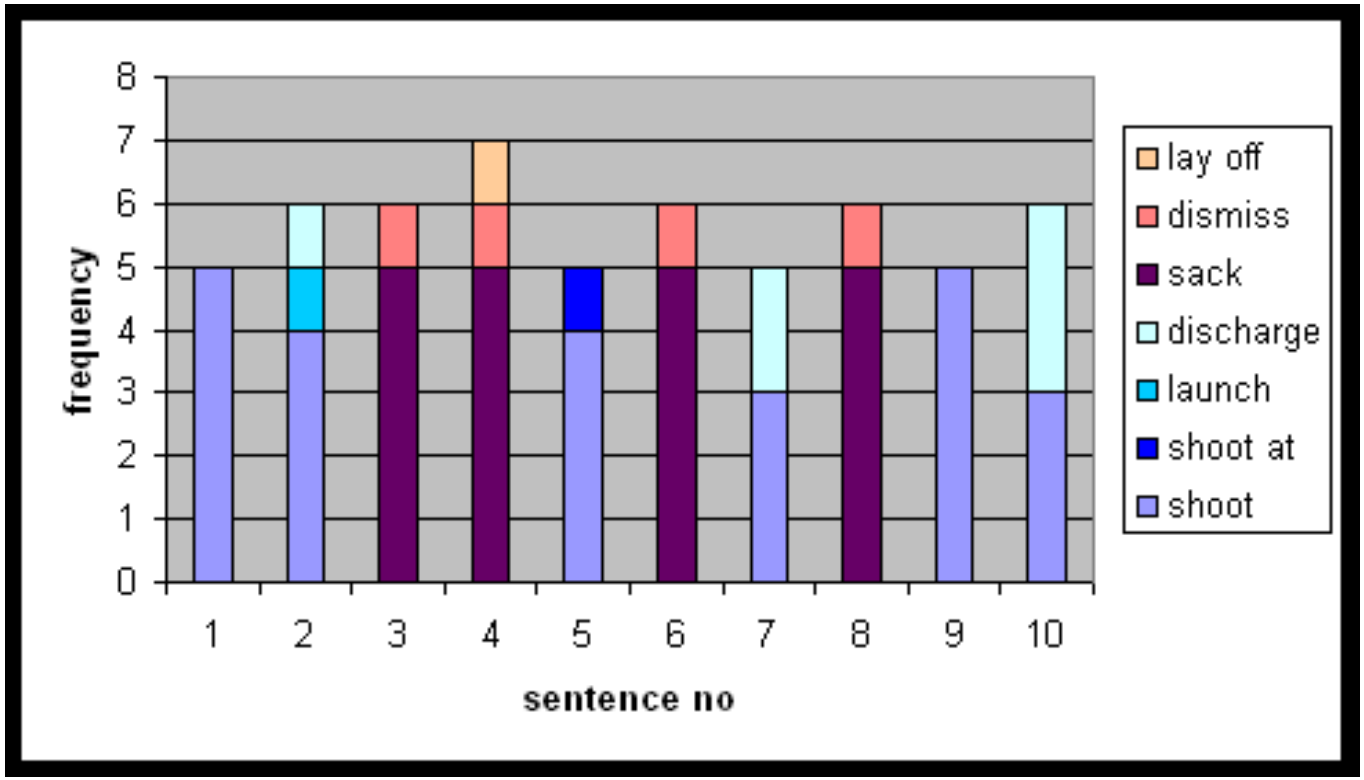
Target word is part of phrase:

Comments:

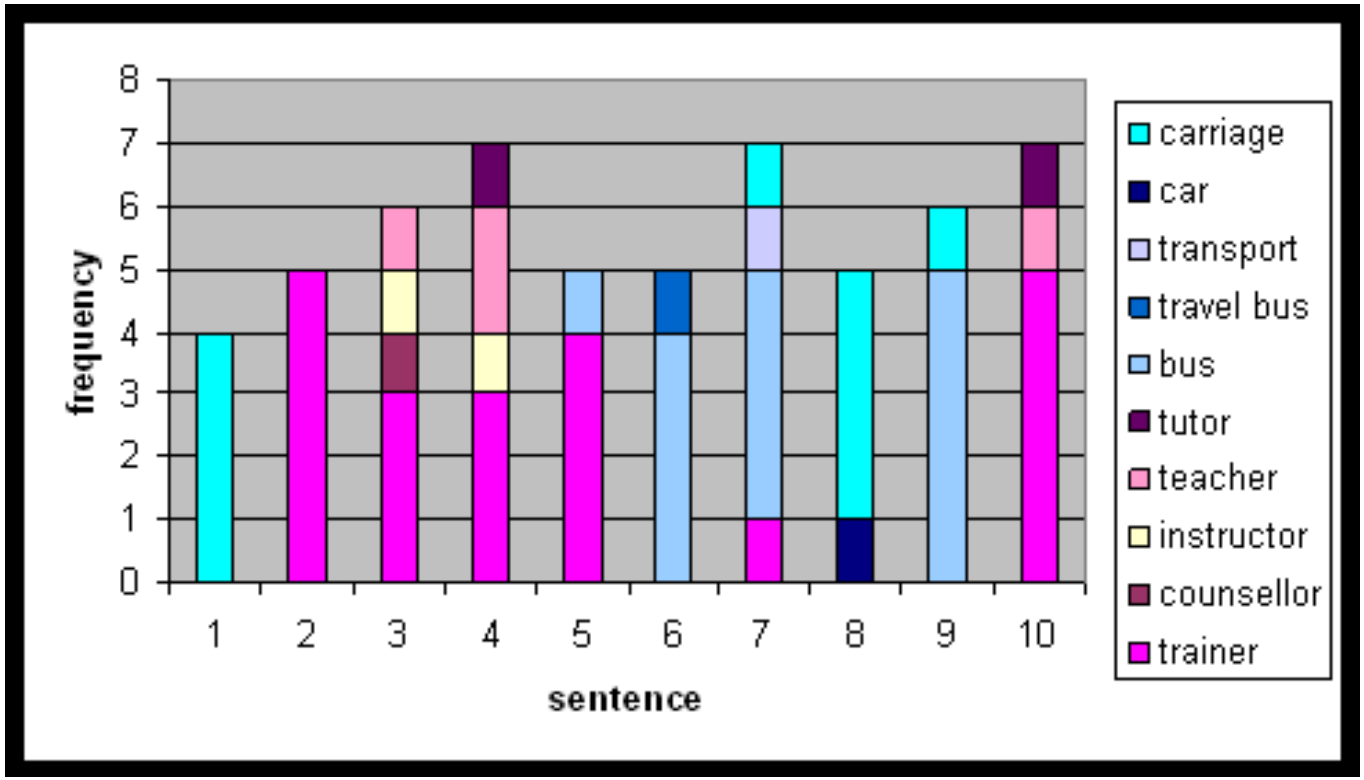
Reminder: "You are free to consult a dictionary or thesaurus if it helps, but not another person. Please tick the dictionary box if you did consult a dictionary for any of the items for this word"

< **previous** | **next** > | **summaries** | **instructions** | **logout**

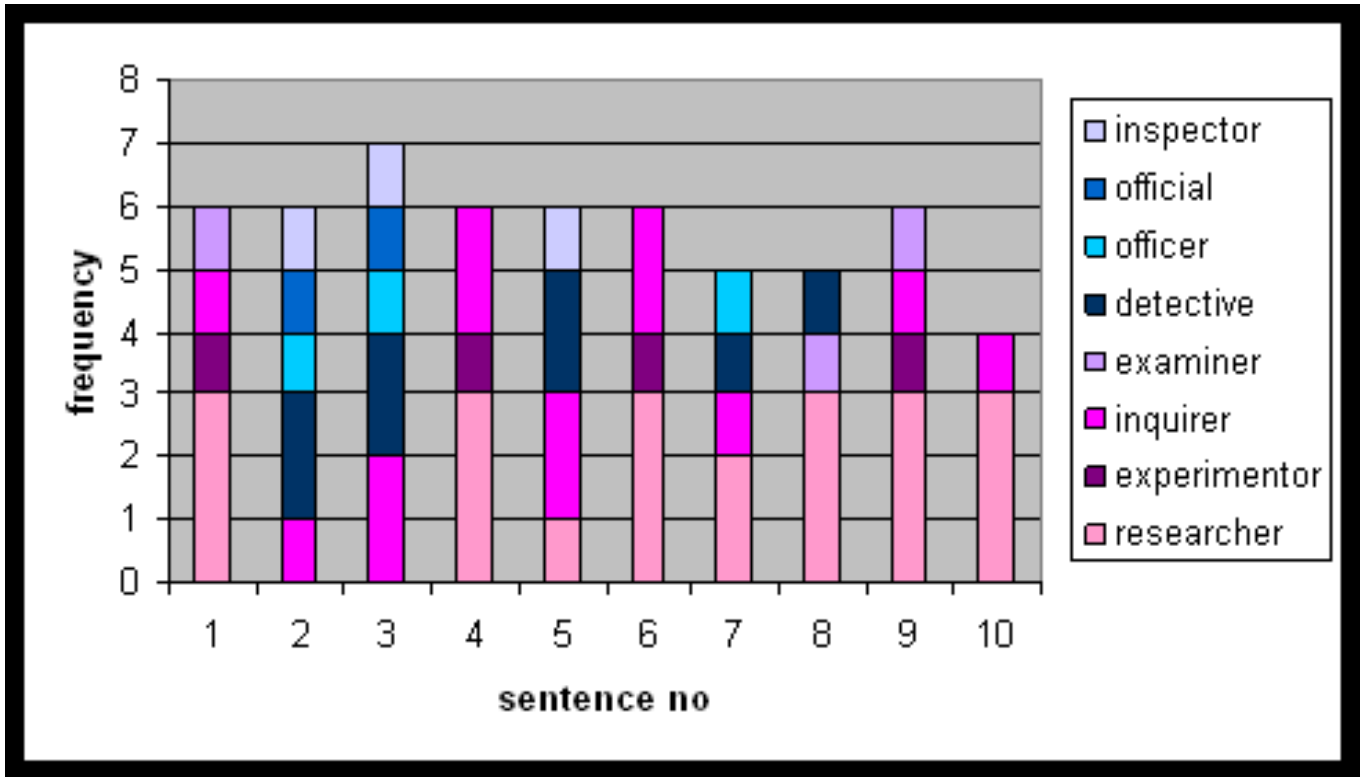
Substitutes for *fire* (verb)



Substitutes for *coach* (noun)



Substitutes for *investigator* (noun)



Agreement

pairwise agreement between every possible pairing (P)

PoS	#	$p\ a$	% with modes	agreement with mode
noun	497	28.4	74.4	52.2
verb	440	25.2	72.3	48.6
adjective	468	24.0	72.7	47.4
adverb	298	36.4	77.5	56.1
all	1703	27.7	73.9	50.7

Some More Statistics

Average Number of Substitutes and
Spread of Substitute over Sentences for that Word and PoS

PoS	#	avg # per item	spread
noun	497	5.7	1.9
verb	440	6.5	1.8
adjective	468	6.4	2.0
adverb	298	6.4	2.3
all	1703	6.2	1.9

Scoring

best systems provide best answers and credit is divided by number of answers

oot systems provide 10 answers and credit is not divided by number of answers

mw systems are scored for detecting where the target word is part of a "multiword" and for identifying what that multiword is

details at http://nlp.cs.swarthmore.edu/semeval/tasks/task10/task10documentation.pdf

best scores

- precision and recall against frequency distribution of substitutes
- systems can produce more than 1 answer but scores are divided by the number of guesses as well as by number of gold standard substitutes for that item
- Mode precision and recall: score first item against mode

oot scores

- precision and recall against frequency distribution of substitutes
- systems produce 10 answers and the scores are not divided by the number of answers
- Mode precision and recall: score to see if mode is in top 10 answers

mw scores

An item is considered a "multiword" if there is a majority vote by more than one annotator for the same multiword

- detection : does the system find a multiword at the same sentences as the annotators
- identification : does the system find the same multiword as the annotators
- precision : how many items are correct from the ones the system finds
- recall : how many items are correct from the ones the annotators find

Baselines: From WordNet

1. synonyms from the first synset of the target word, and ranked with frequency data obtained from the BNC [Leech, 1992].
2. synonyms from the hypernyms (verbs and nouns) or closely related classes (adjectives) of that first synset, ranked with the frequency data.
3. synonyms from all synsets of the target word, and ranked using the BNC frequency data.
4. synonyms from the hypernyms (verbs and nouns) or closely related classes (adjectives) of all synsets of the target, ranked with the BNC frequency data.

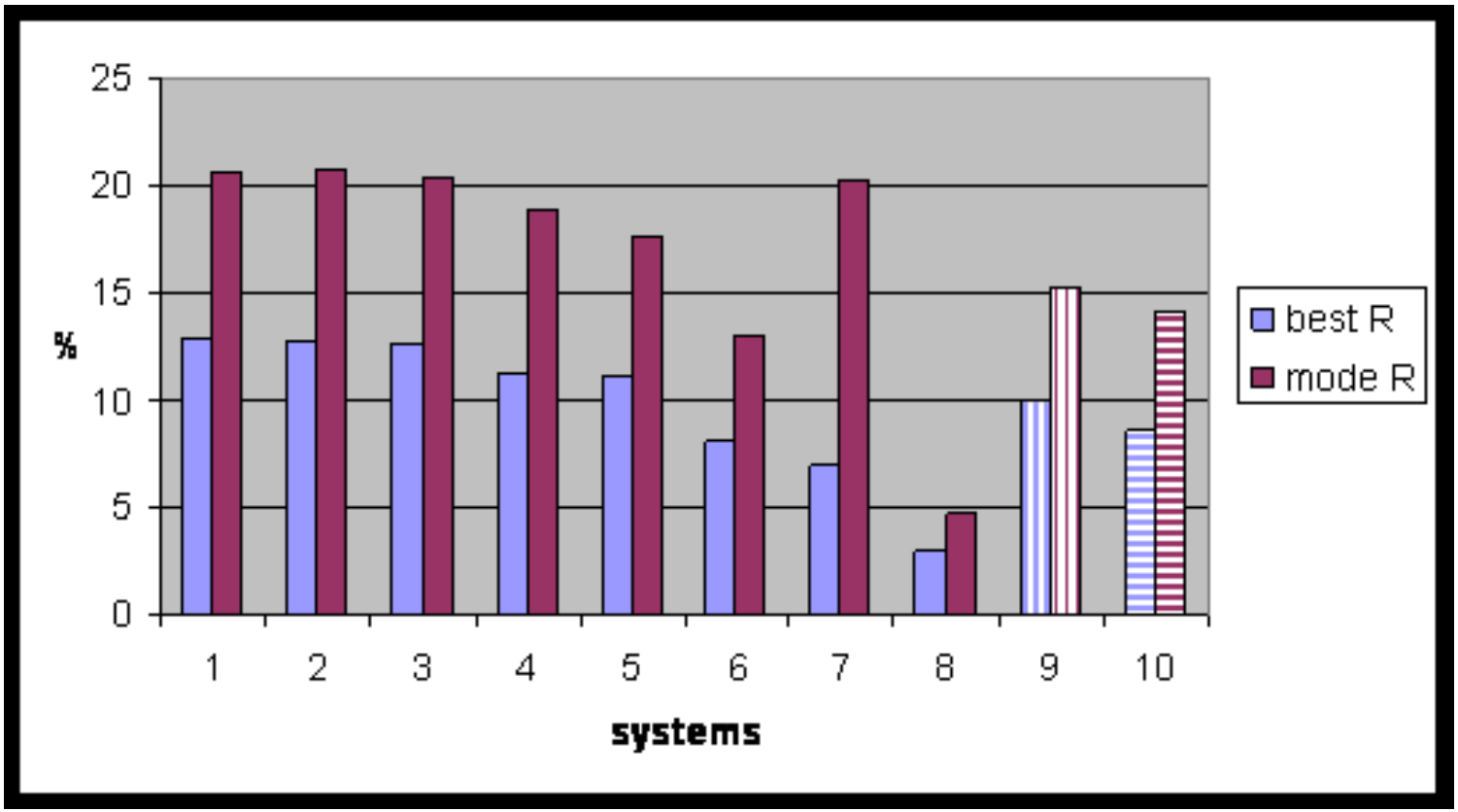
Baselines: Using Distributional Scores

- Lin [Lin, 1998]
- Jaccard
- L1
- cosine
- $\alpha\mathrm{SD}$ [Lee, 1999]

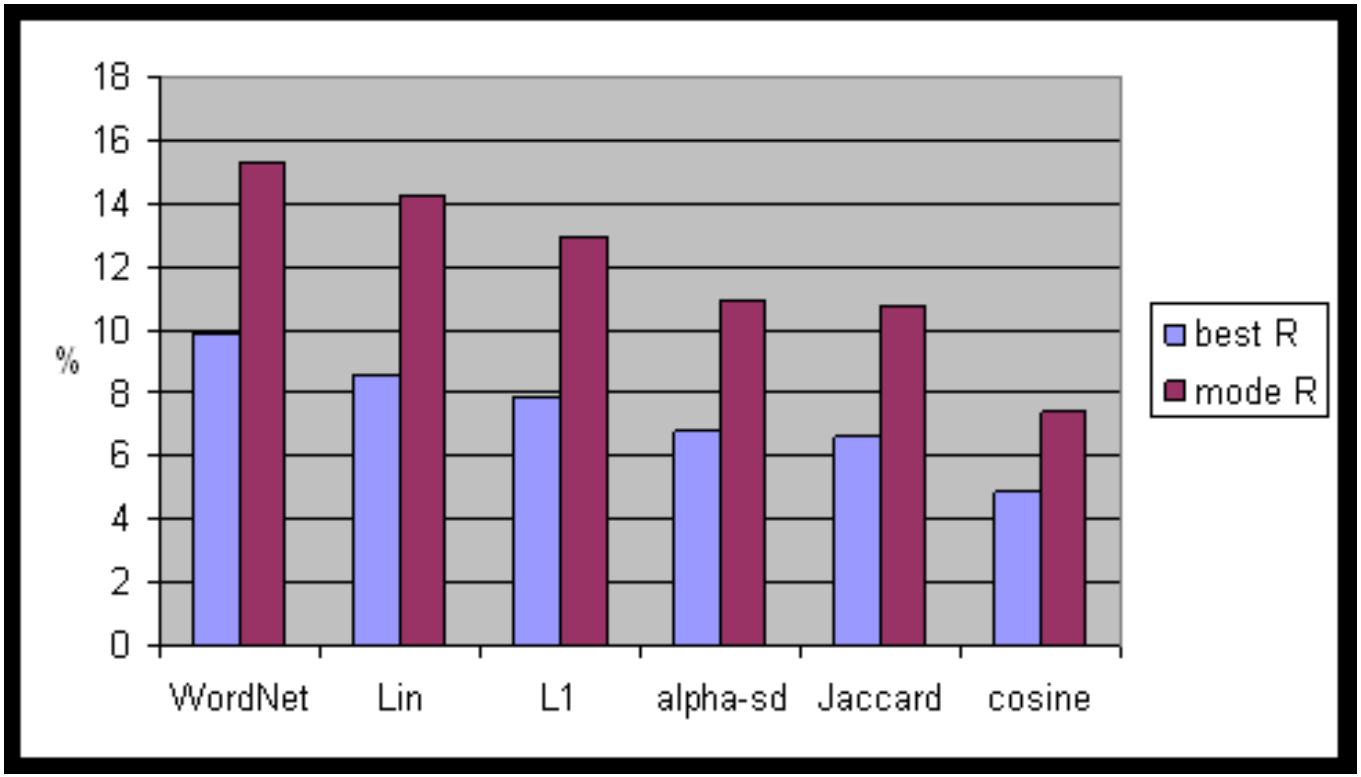
Systems

- 8 teams with 10 systems
- all used 1 or more manually defined inventories (5 in total)
- most use web data for disambiguation
- sense-labelled data only used by 2 systems
  - 1 for filtering synonyms
  - one for semi-supervised learning
- multiword task: 1 system, longest multiword in WordNet in window of 2 words either side of target

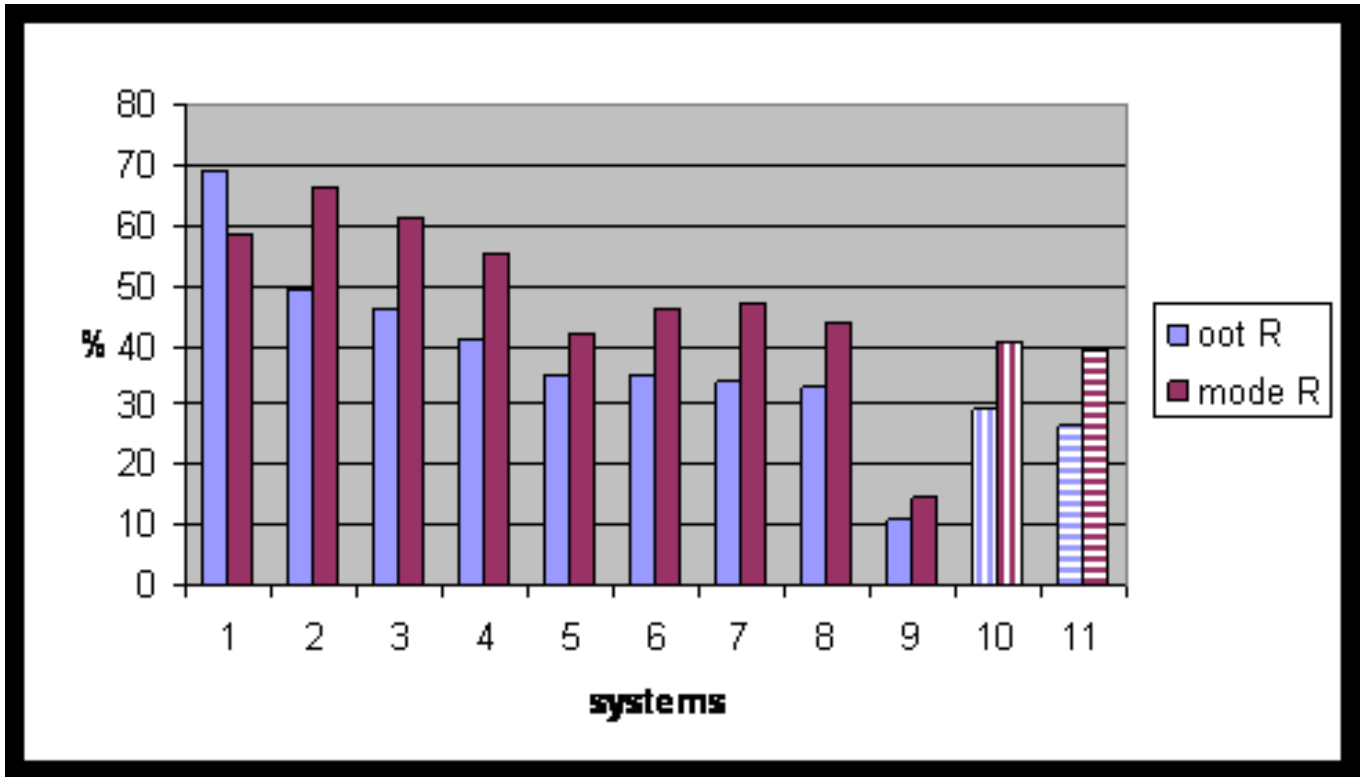
best results



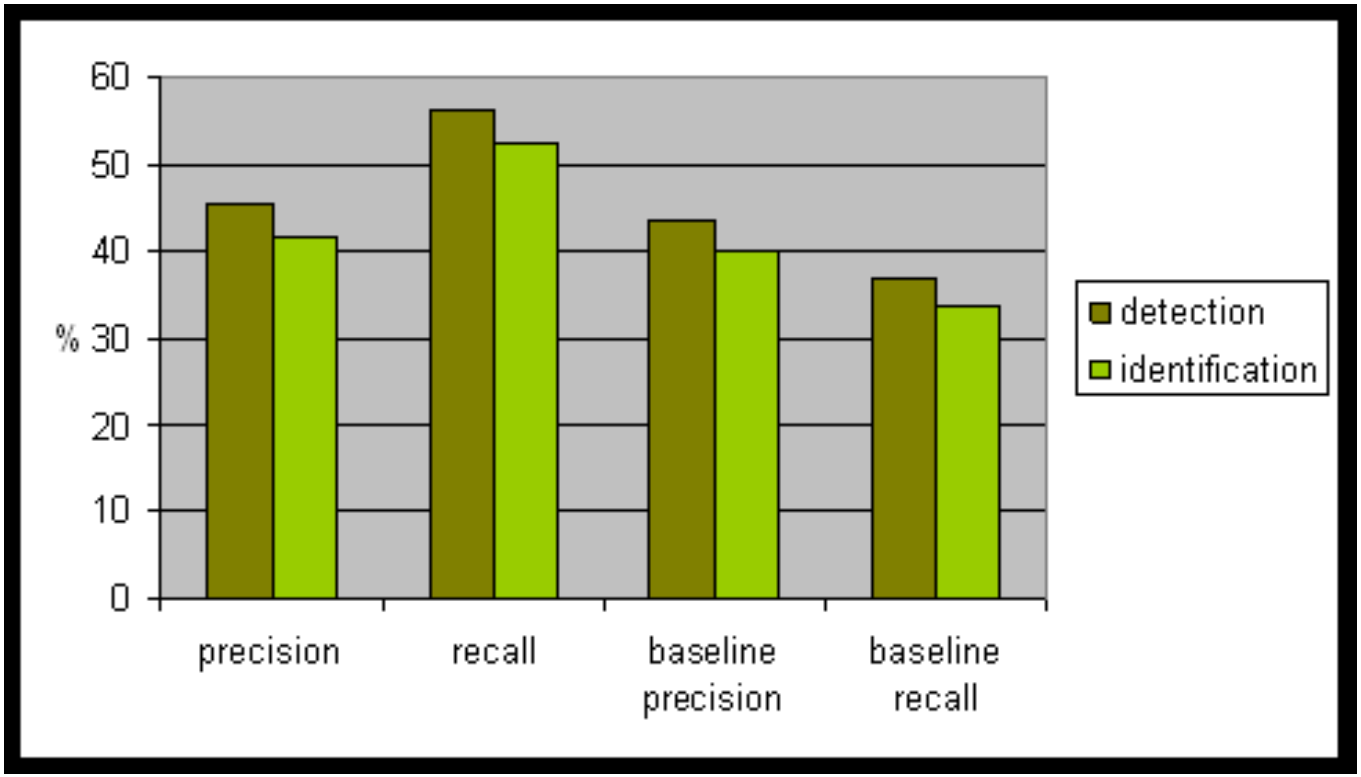
best Baseline Results



oot Results



mw Results


60
50
40
% 30
20
10
0
precision
recall
baseline precision
baseline recall
detection
identification

Post-Hoc Evaluation

- 3 new native English speakers from the UK
  - 1 some background in linguistics
  - 2 other backgrounds
- 100 randomly selected sentences (with substitutes)
- categorised substitutes from original annotators and systems
- good, reasonable, bad

The Post-Hoc Annotation Interface

LexSub POST-HOC
An interface for Lexical Substitution

Please rate the quality of the candidate substitutes for the word in bold in the sentence below:

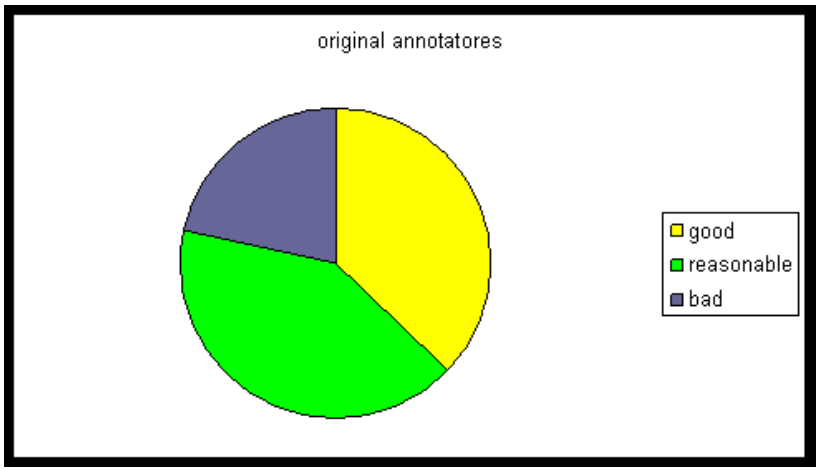
Sentence #675:
Other costs (**match** day , ground and administration) were down by 12 % on 2001/02 levels .

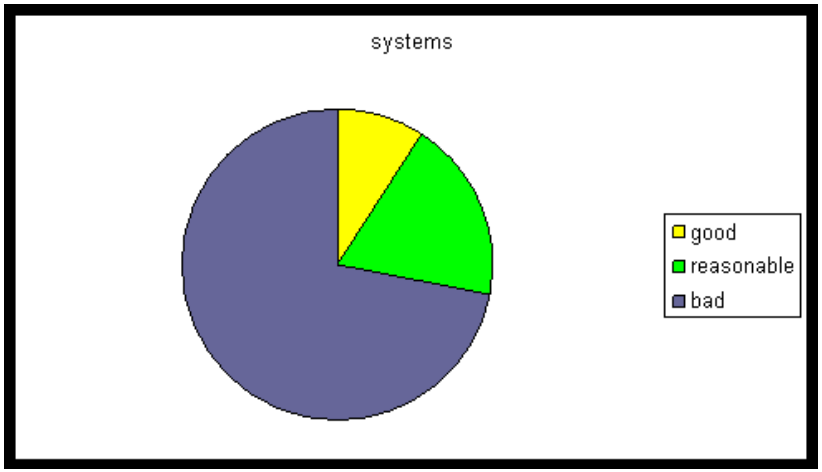
candidate substitutes:

fire	bad
event	bad
equal	bad
couple	bad
tournament	bad
family	bad
contest	bad
game	bad
test	bad

Post-Hoc Results

Percentage of majority decisions by 3 post-hoc annotators





Conclusions

- lexical substitution task successful
  - no training data and no fixed inventory
  - 8 teams 10 systems
  - participants positive about task

Conclusions

- lexical substitution task successful
  - no training data and no fixed inventory
  - 8 teams 10 systems
  - participants positive about task
- human substitutes has higher proportion of substitutes ranked as good or reasonable by post-hoc annotators
- participants used a range of man-made inventories
- most systems use web data for disambiguation
- lots of scope for unsupervised systems
- performance is better on non-multiwords

Future Work

- further exploration of multiword data for extraction and compositionality detection
- multiwords: which are identified and which not?
- analyse system results by PoS
- examine if lexicographer decisions correlate with substitutions
- look at word meaning overlap using synonym overlaps
- separate evaluation of inventory and disambiguation
- try contextual disambiguation with distributional inventories

References

available at:
http://www.informatics.susx.ac.uk/research/nlp/mccarthy/files/NODALIDA07references.pdf

task web site:
http://www.informatics.sussex.ac.uk/research/nlp/mccarthy/task10index.html

Thanks for support from
UK Royal Society
& INTEROP NoE (508011, 6^{th} EU FP)

References

available at:
http://www.informatics.susx.ac.uk/research/nlp/mccarthy/files/NODALIDA07references.pdf

task web site:
http://www.informatics.sussex.ac.uk/research/nlp/mccarthy/task10index.html

Thanks for support from
UK Royal Society
& INTEROP NoE (508011, 6^{th} EU FP)

Thank you!


Carpuat, M. and Wu, D. (2005a).
Evaluating the word sense disambiguation performance of statistical machine translation.
In *Proceedings of the Second International Joint Conference on Natural Language Processing (IJCNLP)*, Jeju, Korea. Association for Computational Linguistics.


Carpuat, M. and Wu, D. (2005b).
Word sense disambiguation vs. statistical machine translation.
In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, Ann Arbor, Michigan. Association for Computational Linguistics.


Clough, P. and Stevenson, M. (2004).
Evaluating the contribution of eurowordnet and word sense disambiguation to cross-language retrieval.

In *Second International Global WordNet Conference (GWC-2004)*, pages 97–105.


Ide, N. and Wilks, Y. (2006).
Making sense about sense.
In Agirre, E. and Edmonds, P., editors, *Word Sense Disambiguation, Algorithms and Applications*, pages 47–73. Springer.


Lee, L. (1999).
Measures of distributional similarity.
In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 25–32.


Leech, G. (1992).
100 million words of English: the British National Corpus.
Language Research, 28(1):1–13.


Lin, D. (1998).
An information-theoretic definition of similarity.
In *Proceedings of the 15th International Conference on Machine Learning*, Madison, WI.


Pantel, P. and Lin, D. (2002).
Discovering word senses from text.
In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 613–619, Edmonton, Canada.


Sanderson, M. (1994).
Word sense disambiguation and information retrieval.
In *17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 142–151. ACM Press.


Schütze, H. and Pederson, J. O. (1995).

Information retrieval based on word senses.
In *Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, Las Vegas, NV.


Sharoff, S. (2006).
Open-source corpora: Using the net to fish for linguistic data.
International Journal of Corpus Linguistics, 11(4):435–462.


Stokoe (2005).
Differentiating homonymy and polysemy in information retrieval.
In *Proceedings of the joint conference on Human Language Technology and Empirical methods in Natural Language Processing*, Vancouver, B.C., Canada.