Alternative Annotations of Word Usage

Diana McCarthy
University of Sussex

in collaboration with ...

- Roberto Navigli (University of Rome "La Sapienza")
- Katrin Erk and Nick Gaylord (University of Texas at Austin)
- Ravi Sinha and Rada Mihalcea (University of North Texas)

SEW June 2009

Manually produced Inventories: e.g. WordNet

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")
- 9. match – (something that resembles or harmonizes with; "that tie makes a good match with your jacket")

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

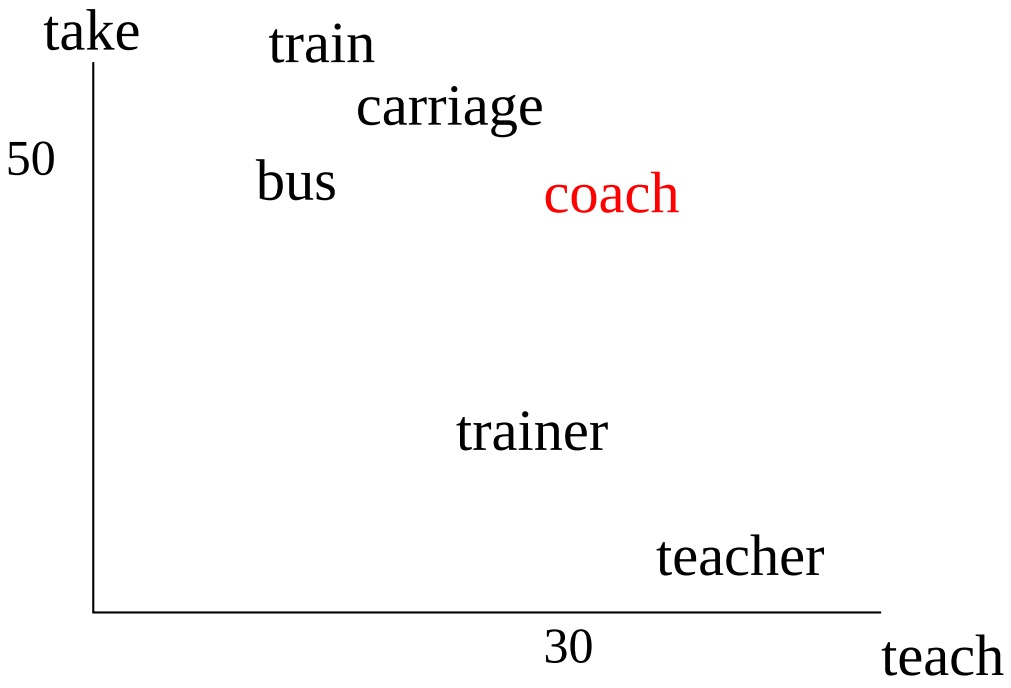
WNs#	gloss
1	a young person
2	a human offspring
3	an immature childish person
4	a member of a clan or tribe

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

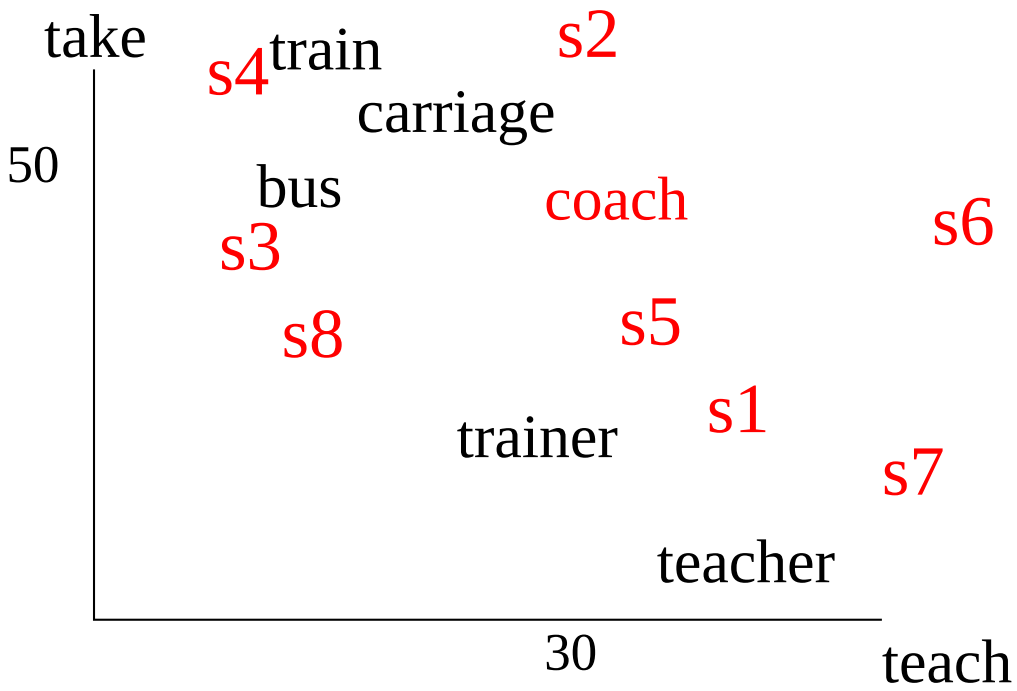
Distributional Approaches

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

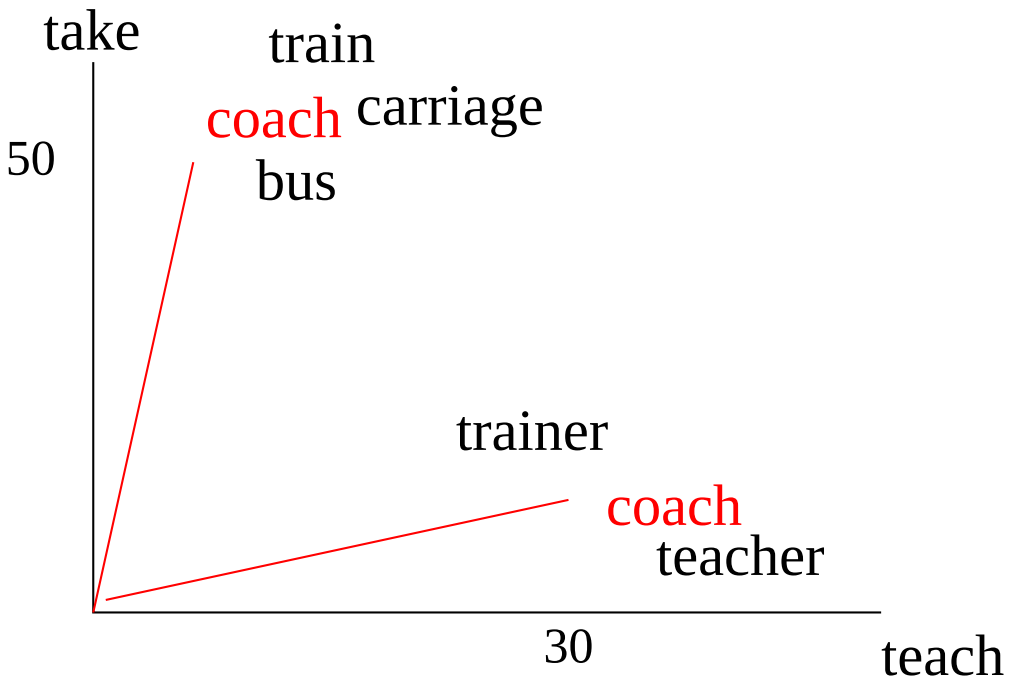
Vector Based Approaches



Vector Based Approaches



Vector Based Approaches



Nearest Neighbour Approaches

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

Nearest Neighbour Approaches

Output
Word: <closest word> <score> <2nd closest > <score>. . .
coach: train 0.171 bus 0.166 player 0.149 captain 0.131 car 0.131

Nearest Neighbour Approaches

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]

Word Sense Disambiguation (WSD)

Given a word in context, find the best-fitting "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 best-fitting "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 best-fitting "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 best-fitting "sense"

This is at least 26 weeks by the week
in which the approved match with
the child is made.

Word Sense Disambiguation (WSD)

Given a word in context, find the best-fitting "sense"

This is at least 26 weeks by the week in which the approved match with the child is made.

Word Sense Disambiguation (WSD)

Given a word in context, find the best-fitting "sense"

This is at least 26 weeks by the week in which the approved match with the child is made.

match#n#9

#9 something that resembles or harmonizes with; "that tie makes a good match with your jacket"

Word Sense Disambiguation (WSD)

Given a word in context, find the best-fitting "sense"

This is at least 26 weeks by the week in which the approved match with the child is made.

#9 something that resembles or harmonizes with; "that tie makes a good match with your jacket"
#8 a pair of people who live together; "a married couple from Chicago"

match#n#9
or possibly
match#n#8

WSD Performance: SemEval 2007

task	Recall best system	MFS	ITA
English all words fine	59.1	51.4	72/86
English all words coarse	82.5	78.9	93.8
English Lexical sample	88.7	78.0	> 90
Chinese English LS via parallel	81.9	68.9	84/94.7
Chinese English LS	71.7	40.5	84.8
Catalan/Spanish nouns	85.9	84.9	

Can This Level of Performance Benefit Applications?

- Enough context: WSD comes out in 'the statistical wash'
- not enough context and can't do anyway
- IR [Stokoe, 2005, Clough and Stevenson, 2004, Schütze and Pederson, 1995] vs [Sanderson, 1994]
- MT [Carpuat and Wu, 2005b, Carpuat and Wu, 2005a] vs [Chan et al., 2007, Carpuat and Wu, 2007]

Does this Methodology have Cognitive Validity?

- [Kilgarriff, 2006]
  - Word usages often fall between dictionary definitions
  - the distinctions made by lexicographers are not necessarily the ones to make for an application
- [Tuggy, 1993] Word meanings lie on a continuum between ambiguity and vagueness
- [Cruse, 2000] Word meanings don't have discrete boundaries, a more complex *soft* representation is needed

Does this Methodology have Cognitive Validity?

- [Hanks, 2000]
  - Computational procedures for distinguishing homographs are desirable and possible, but...
  - they don't get us far enough for text understanding.
  - Checklist theory at best superficial and at worst misleading.
  - Vagueness and redundancy needed for serious natural language processing
- [McCarthy, 2006] Word meanings between others e.g.

bar	pub	$\leftrightarrow$	counter	$\leftrightarrow$	rigid block of wood
child	young person	$\leftrightarrow$	offspring	$\leftrightarrow$	descendant

Outline

Three Datasets

to compare different representations of word meaning in context

- SemEval-2007 Lexical Substitution (LEXSUB) with Roberto Navigli
- 2 Datasets with Graded Judgments with Katrin Erk and Nick Gaylord
  1. Usage Similarity (Usim)
  2. Word Sense Similarity (WSsim)

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 game.

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

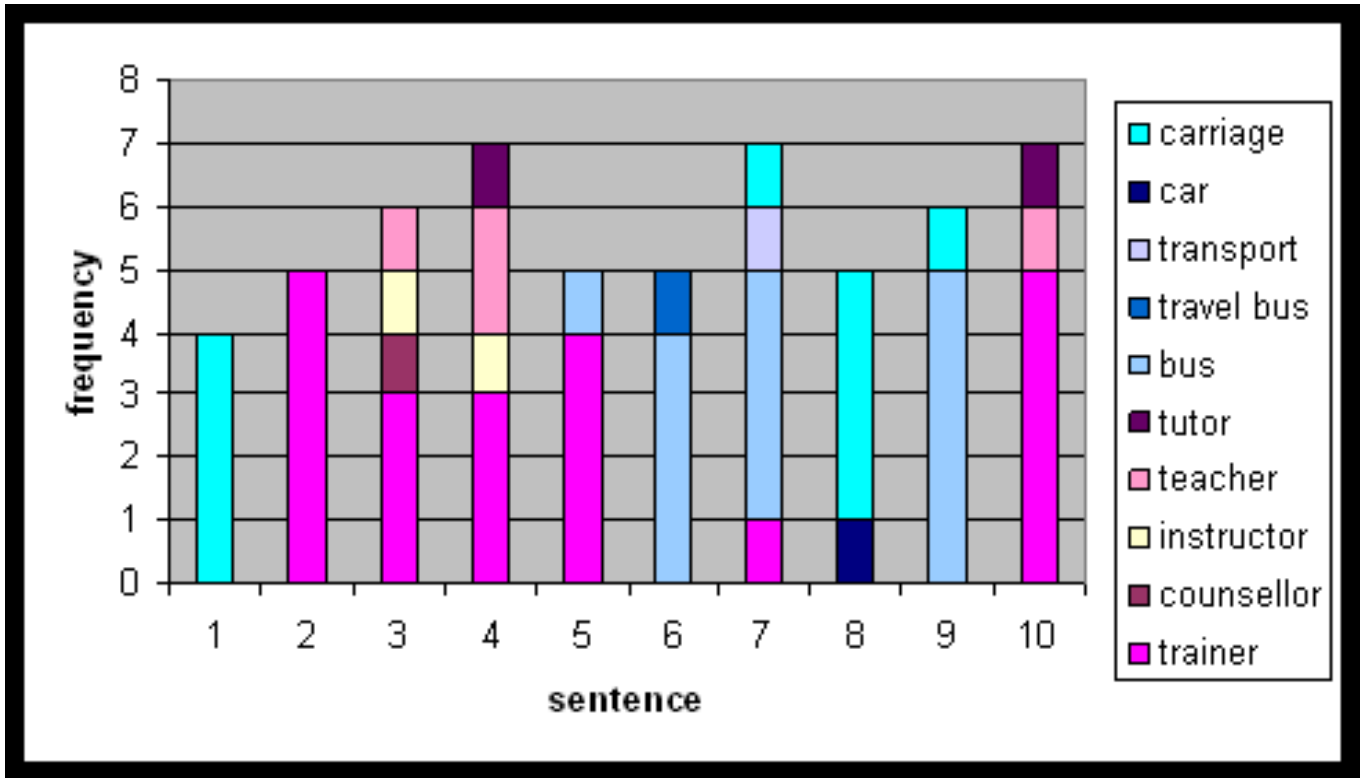
Target word is part of phrase:

Comments:

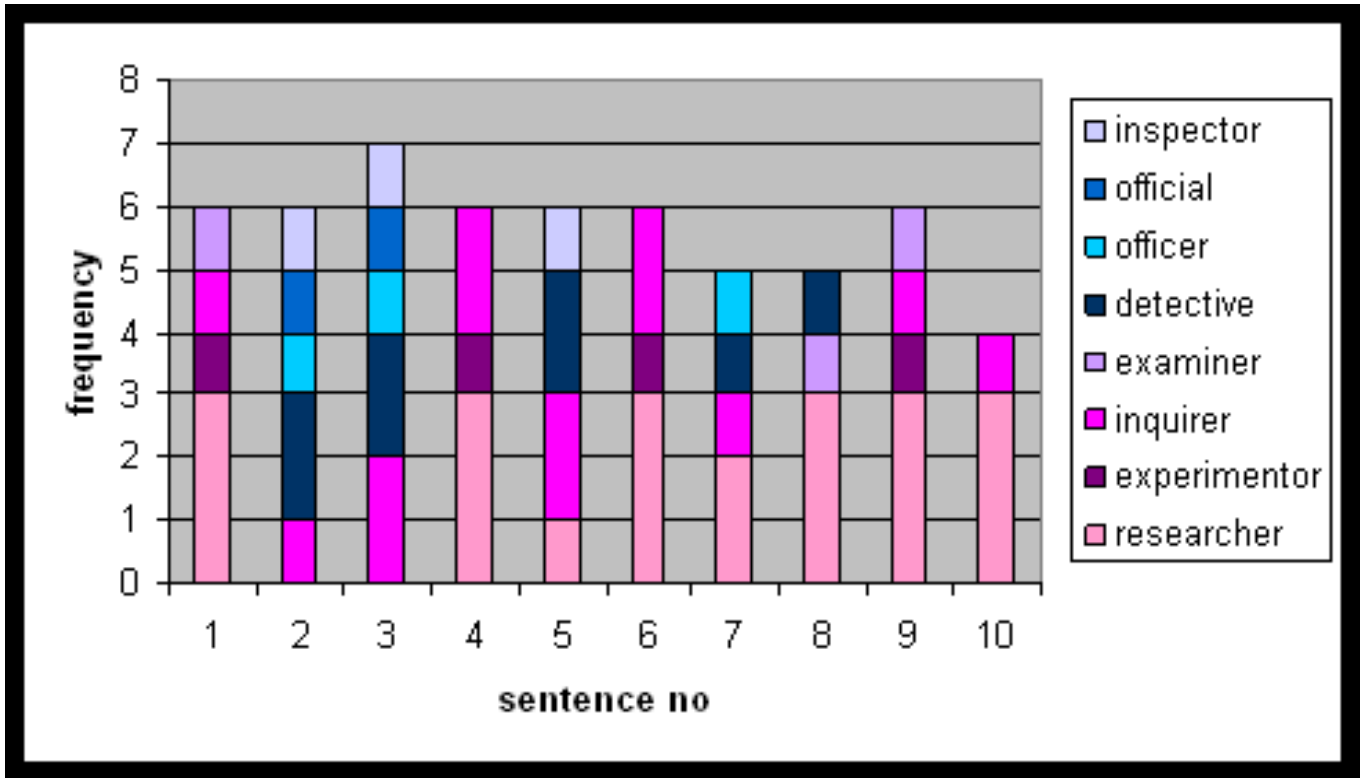
Reminder: "You are free to consult a dictionary or thesaurus if it helps, but not another person. Please tick the dictionary box if you did consult a dictionary for any of the items for this word"

< previous | next > | summaries | instructions | logout

Substitutes for *coach* (noun)



Substitutes for *investigator* (noun)



LEXSUB Systems and Findings

- all participants used 1 or more hand-crafted inventories
- WordNet and Roget most popular
- contextual disambiguation with n-grams
- baselines with distributional techniques
- task is hard because of inherent variability
- post-hoc analysis annotators preferred substitutes from humans
- scope for evaluating inventories AND/OR disambiguation
- look at word meaning using synonym overlap

Outline

Data

- Usim
  - 34 lemmas (nouns, verbs, adjectives and adverbs) 10 sentences each from LEXSUB
  - 340 sentences
- WSsim
  - 8 lemmas (nouns, verbs and adjectives) 50 sentences each from SemCor and SENSEVAL-3 English Lexical Sample (SE-3)
  - 3 lemmas data from LEXSUB 10 sentences each also in Usim
  - 430 sentences

Usim interface

Rate how similar in meaning the two boldfaced words below are:

This is sentence pair number **9**

(1) This more upright position is most easily and affordably achieved through slapping a riser bar on your setup, and only requires you to buy a bar instead of a **bar** and stem.
(2) For twelve hours Livewire will be broadcasting live from the blue **bar** of Union House at UEA in an attempt to raise as much money as possible for a very worthy cause.

- 1: Completely different
- 2: Mostly Different
- 3: Similar
- 4: Very Similar
- 5: Identical
- Cannot Decide

Click for Full Instructions

Comment:

WSsim interface

Sentence #21

4 How can one generate the probability density **function** of an Erlang distribution using Stella?

Rate how close the meaning of the above boldfaced word is to each of the following descriptions:

1=Completely Different, 2=Mostly Different, 3=Similar, 4=Very Similar, 5=Identical

Click for Full Instructions

○1 ○2 ○3 ○4 ○5 duty (the actions and activities assigned to or required or expected of a person or group)
○1 ○2 ○3 ○4 ○5 utility (what something is used for)
○1 ○2 ○3 ○4 ○5 software system (a set sequence of steps, part of larger computer program)
○1 ○2 ○3 ○4 ○5 social event (a vaguely specified social event)
○1 ○2 ○3 ○4 ○5 social gathering (a formal or official social gathering or ceremony)
○1 ○2 ○3 ○4 ○5 mathematical relation ((mathematics) a mathematical relation such that each element of a given set (the domain of the function) is associated with an element of another set (the range of the function))
○1 ○2 ○3 ○4 ○5 relation (a relation such that one thing is dependent on another)

Comment:

Usim example:

1) We study the methods and concepts that each writer uses to defend the cogency of legal, deliberative, or more generally political prudence against explicit or implicit charges that practical thinking is merely a knack or form of cleverness.

2) Eleven CIRA members have been convicted of criminal charges and others are awaiting trial.

Annotator judgments: 2,3,4

WSsim example

Sentence	Senses						
	1	2	3	4	5	6	7
This question provoked **arguments** in America about the	1	4	4	2	1	1	3
Norton Anthology of Literature by Women, some of the	4	5	4	2	1	1	4
contents of which were said to have had little value as	1	4	5	1	1	1	1
literature.							

The senses are: 1:statement, 2:controversy, 3:debate, 4:literary argument, 5:parameter, 6:variable, 7:line of reasoning

Graded Annotation

1. LEXSUB overlapping substitutes for two instances
2. Usim ratings do these correlate with 1)
3. WSsim ratings do these correlate with 1) 2) and previous WordNet annotations (SemCor SE-3)

Inter-Tagger Agreement: Spearman's ρ

Usim

$\rho = 0.502$, 0.641 and 0.501 (average 0.548) all $p < 2.2e\text{-}16$

WSsim

$\rho =$0.506, 0.466 0.540 (average 0.504) all $p < 2.2e\text{-}16$.

Do Judgments Correlate with LEXSUB Substitutes?

Synonym Overlap: $\frac{|multiset\ \ intersection|}{size\ \ of\ \ larger\ \ multiset}$

e.g.

$S_1\{game, game, game, tournament\}$

$S_2\ \{\ game, game, competition, tournament\} = \frac{3}{4}$

WSsim: $ED(J_1, J_2) = \sqrt{(\sum_{i=1}^{n}(J_1[i] - J_2[i])^2)}$

Annotator correlation with LEXSUB substitute overlap

Usim All		Usim W∩U	WSsim W∩U	
ann.	ρ	ρ	ann.	ρ
4	0.383	0.330	1	-0.520
5	0.498	0.635	2	-0.503
6	0.584	0.631	3	-0.463

Comparison to Previous Resources: multiple sense assignments

Data	Orig.	WSsim judgment ≥ 3	≥ 4	5
WSsim/SemCor	0.0	80.2	57.5	28.3
WSsim/SE-3	24.0	78.0	58.3	27.1
All WSsim		78.8	57.4	27.7

WSsim Correlation to Previous Resources

converted SemCor and SE-3 to 1 or 5 judgment for each sense

data	individual ρ	average ρ	significance
SemCor	0.234 0.448, 0.390	0.357	$p <$ 2.2e-16
SE-3	0.346, 0.449, 0.338	0.378	$p <$ 2.2e-16

WSsim Sense Correlations

Percentage of sense pairs that were significantly positively (pos) or negatively (neg) correlated at $p < 0.05$ and $p < 0.01$, shown by annotator.

	$p < 0.05$		$p < 0.01$	
	pos	neg	pos	neg
Ann. 1	30.8	11.4	23.2	5.9
Ann. 2	22.2	24.1	19.6	19.6
Ann. 3	12.7	12.0	10.0	6.0

Uncorrelated Sense with High Ratings

Percentage of sentences in which at least two uncorrelated ($p > 0.05$) or negatively correlated senses have been annotated with judgments at the specified threshold.

	$j \geq 3$	$j \geq 4$	$j = 5$
Ann. 1	71.9	49.1	8.1
Ann. 2	55.3	24.7	8.1
Ann. 3	42.8	24.0	4.9

Summary

- substitutes and graded judgments as alternative annotations of word meaning
- reflect underlying meanings in context and allow relationships between usages
- annotations with senses show highly significant correlation to substitutes
- graded sense annotations show highly significant correlation to best-fitting sense annotation (SemCor, SE-3)
- NB high similarity given to uncorrelated senses for same instance

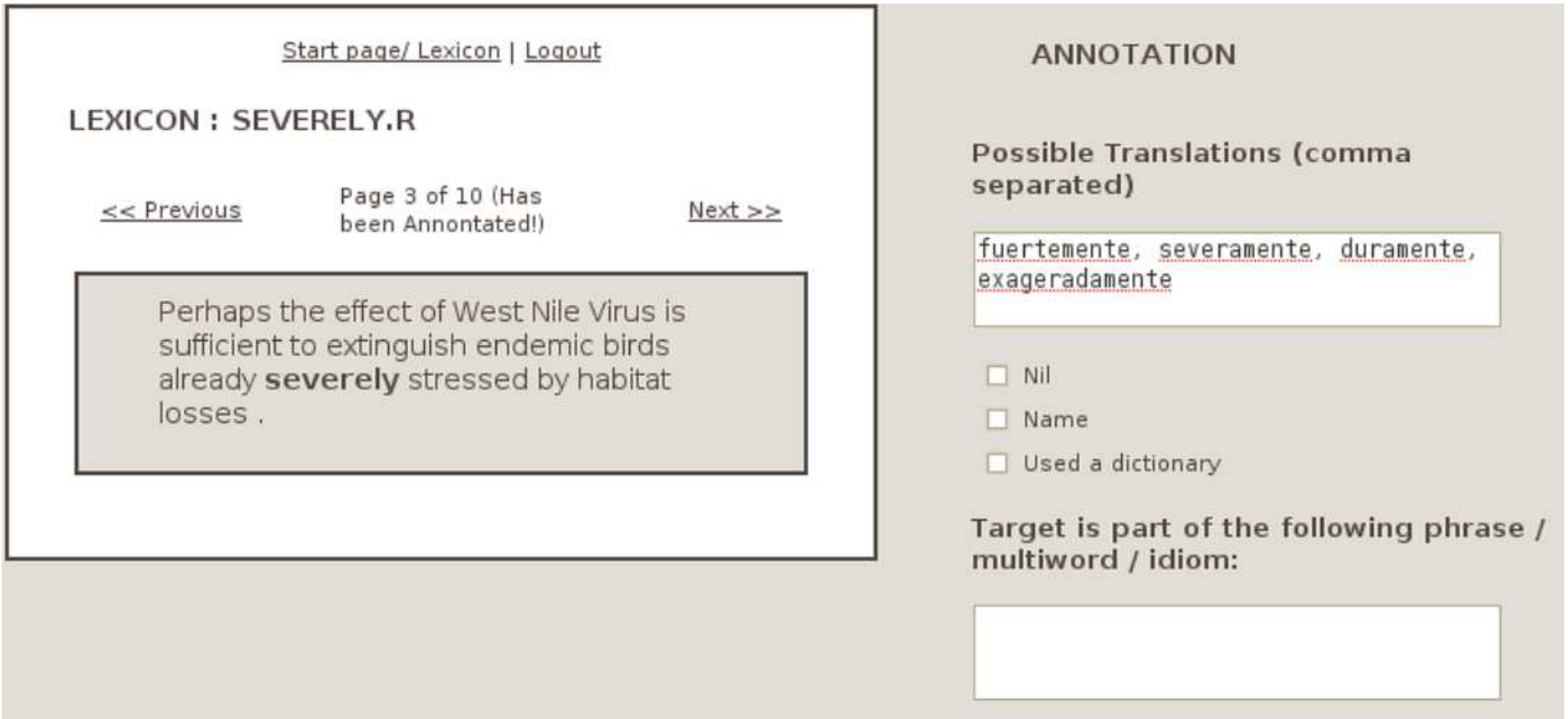
Future Work

- build soft meaning representations (non discrete representations)
- compare WSD systems
- can graded output help applications?
- use more annotators

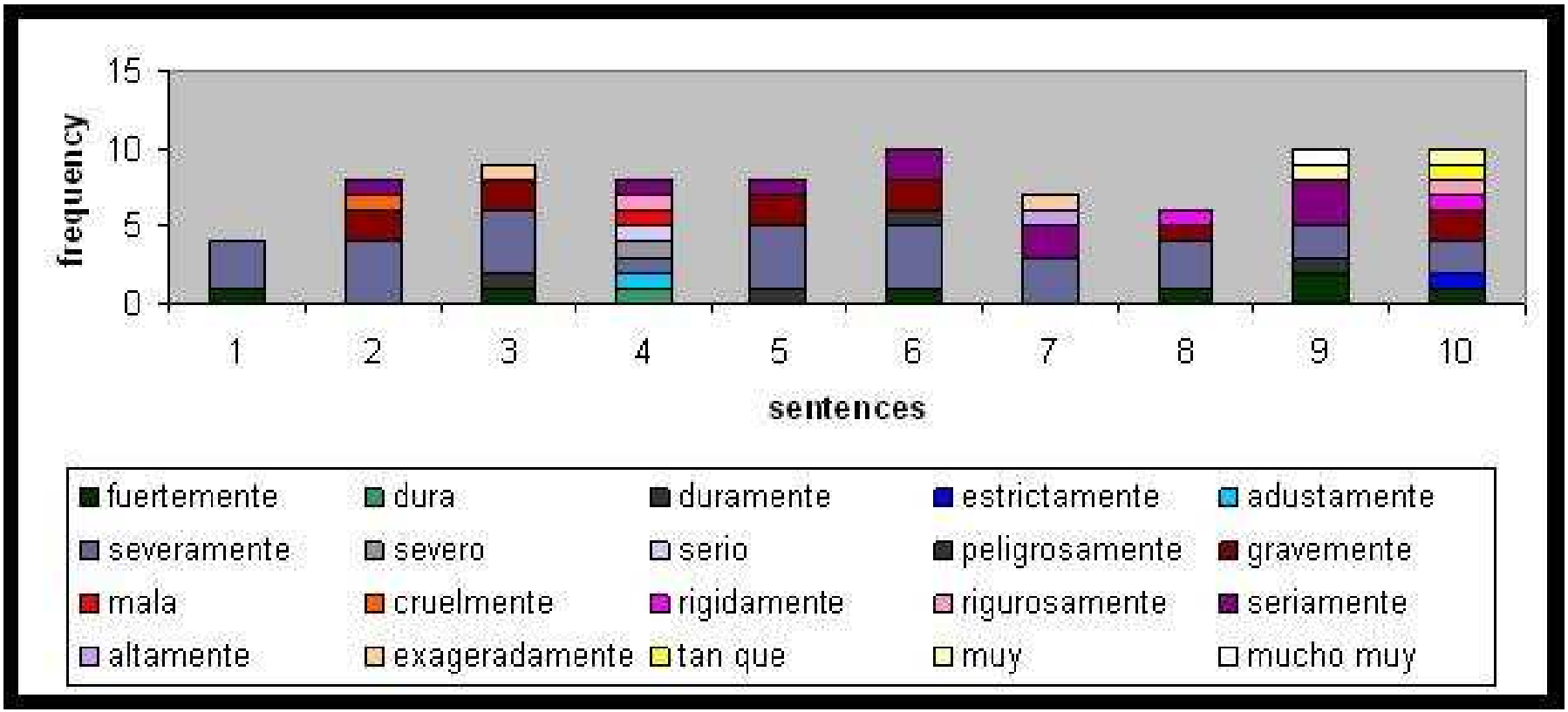
Cross-Lingual Lexical Substitution (CLLS)

- coming soon . . . 2010
- Rada Mihalcea's idea do substitution task with translations
- Not fixing resources/inventory
- Not restricting number of translations for an instance
- Not expecting discrete senses
- applications for foreign language learners, tool for human or machine translation

CLLS interface



Some Translations for *severely*



Credits

Collaboration with Roberto Navigli
and Katrin Erk and Nick Gaylord
and Rada Mihalcea, Ravi Sinha

Support from


THE ROYAL SOCIETY

&

INTEROP NoE
(508011, 6^{th} EU FP)

- LEXSUB task web site:
  http://www.informatics.sussex.ac.uk/research/nlp/mccarthy/task10index.html
- CLLS web site:
  http://lit.csci.unt.edu/index.php/Semeval_2010
- Usim and WSsim data to be released soon

And finally . . .

And finally . . .

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.


Carpuat, M. and Wu, D. (2007).
Improving statistical machine translation using word sense disambiguation.

In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)*, pages 61–72, Prague, Czech Republic. Association for Computational Linguistics.

Chan, Y. S., Ng, H. T., and Chiang, D. (2007).
Word sense disambiguation improves statistical machine translation.
In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 33–40, Prague, Czech Republic. 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.


Cruse, D. A. (2000).
Aspects of the microstructure of word meanings.
In Ravin, Y. and Leacock, C., editors, *Polysemy: Theoretical and Computational Approaches*, pages 30–51. OUP, Oxford, UK.


Hanks, P. (2000).
Do word meanings exist?
Computers and the Humanities. Senseval Special Issue, 34(1–2):205–215.


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.


Kilgarriff, A. (2006).

Word senses.
In Agirre, E. and Edmonds, P., editors, *Word Sense Disambiguation, Algorithms and Applications*, pages 29–46. Springer.


McCarthy, D. (2006).
Relating wordnet senses for word sense disambiguation.
In *Proceedings of the EACL 06 Workshop: Making Sense of Sense: Bringing Psycholinguistics and Computational Linguistics Together*, pages 17–24, Trento, Italy.


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.


Stokoe, C. (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*, pages 403–410, Vancouver, B.C., Canada.


Tuggy, D. H. (1993).

Ambiguity, polysemy and vagueness.
Cognitive linguistics, 4(2):273–290.