Graded Annotations of Word Meaning in Context

Diana McCarthy
Lexical Computing Ltd.
Erasmus Mundus Visiting Scholar at the Universities of Melbourne and Saarlandes

in collaboration with . . .

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

ABBYY Open Seminar May 2011

Outline

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")

Distributional approaches

context	frequency		
	match	*game*	*lighter*
ignite	4	2	2
hold	30	2	22
ticket	27	23	0
watch	15	21	1

Distributional approaches

context	frequency		
	match	*game*	*lighter*
ignite	4	2	2
hold	30	2	22
ticket	27	23	0
watch	15	21	1

Thesaurus (nearest neighbour) output
Word: <closest word> <score> <2nd closest > <score>...
match: game 0.171 tournament 0.166 matchstick 0.149 cigarette 0.131 competition 0.131

Distributional approaches

context	frequency		
	match	*game*	*lighter*
ignite	4	2	2
hold	30	2	22
ticket	27	23	0
watch	15	21	1

Thesaurus (nearest neighbour) output
Word: <closest word> <score> <2nd closest > <score>...
match: game 0.171 tournament 0.166 matchstick 0.149 cigarette 0.131 competition 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 (recall)

task	best system	MFS	ITA
SemEval 2007			
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
SemEval 2010 domain specific all words			
English	55.5	50.5	-
Chinese	55.9	56.2	96
Dutch	52.6	48.0	90
Italian	52.9	46.2	72

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?

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

Alternative word meaning annotations: datasets

to compare different representations of word meaning in context

- SemEval-2007 Lexical Substitution (LEXSUB) (McCarthy and Navigli, 2009)
- SemEval-2010 Cross-Lingual Lexical Substitution (CLLS) (Mihalcea et al., 2010)
- Usage Similarity (Usim) and Graded Word Sense (WSsim) (Erk et al., 2009) and on going ...

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.

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.

201 words (2010 sentences) from the English Internet Corpus (Sharoff, 2006)

LEXSUB 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
top-up with extra fluid before the **match** .

Substitute: game OK

☐ nil ☐ extra responses ☐ name ☑ used a dictionary

Target word is part of phrase:

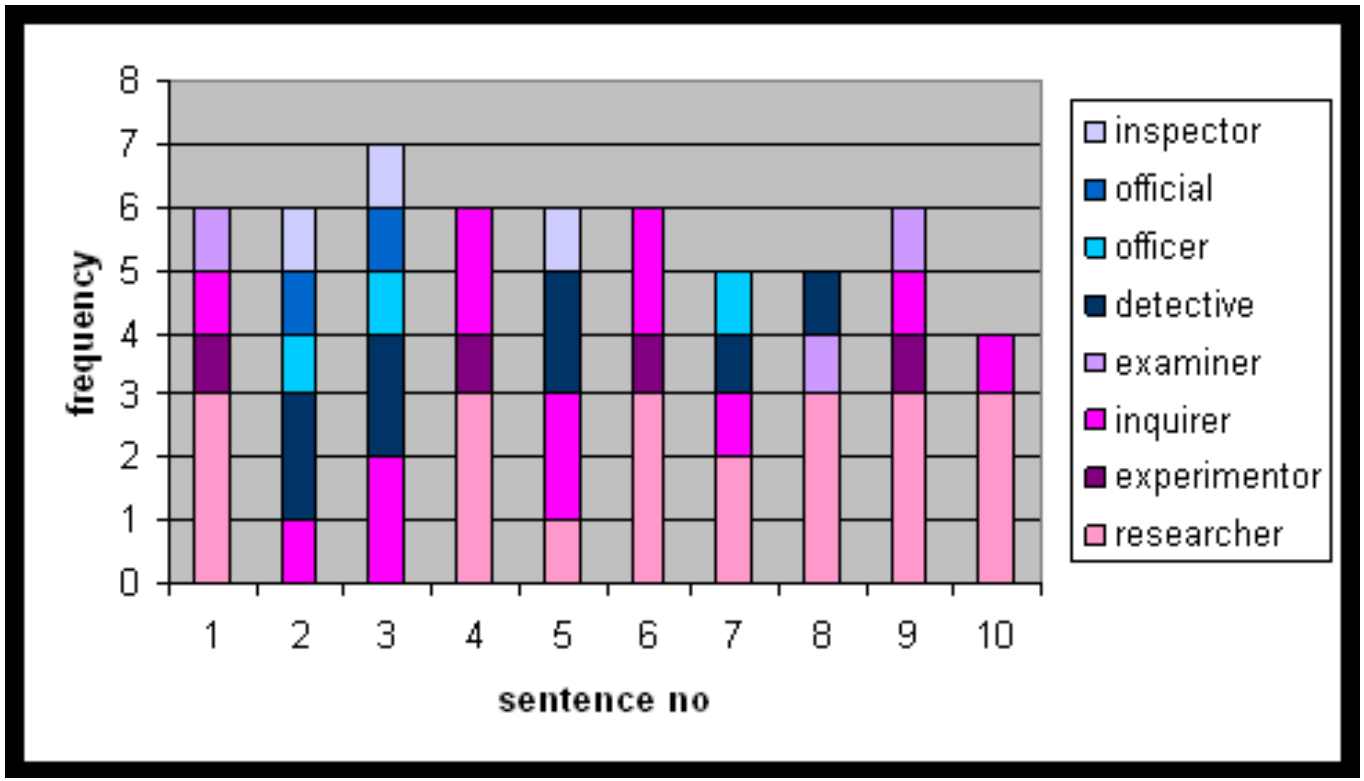
Comments:

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)

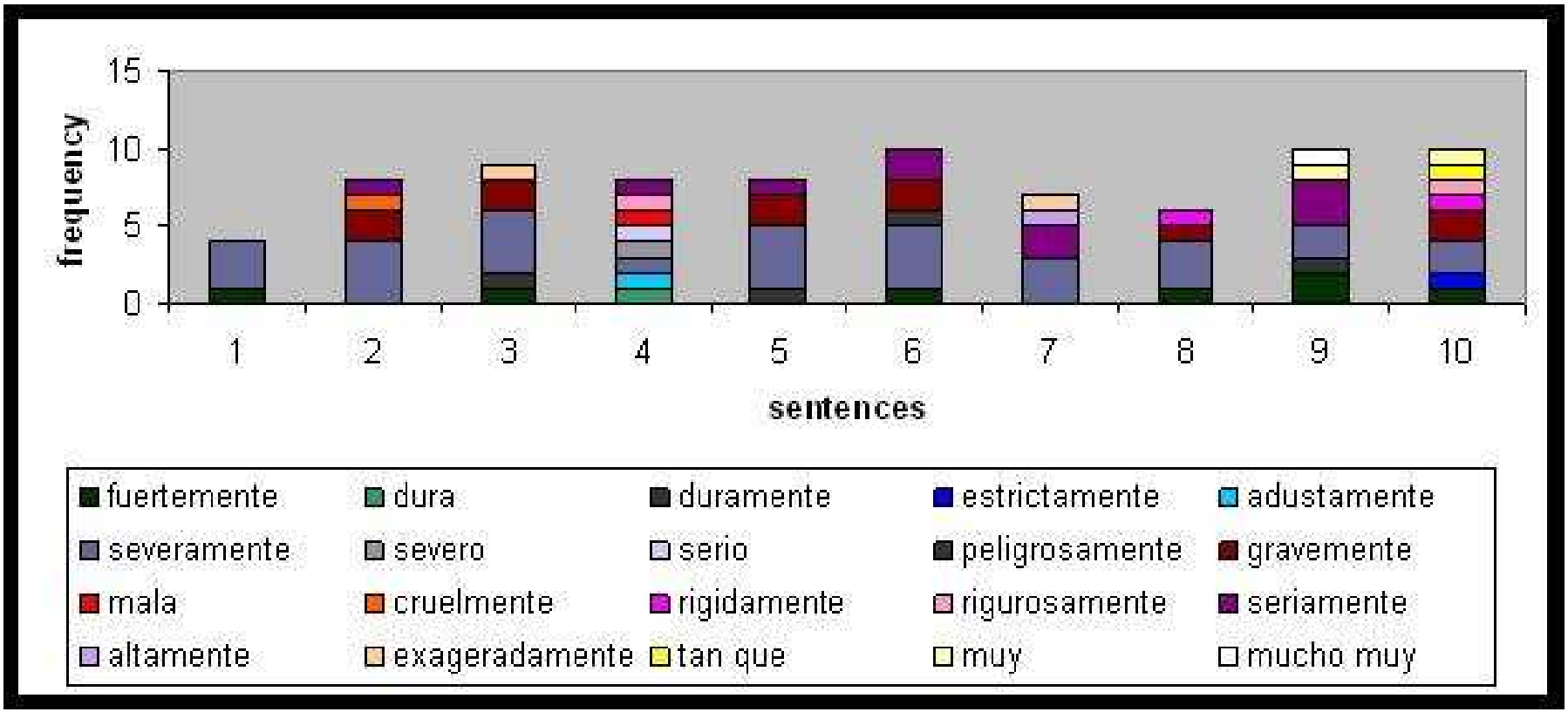
Substitutes for *investigator* (noun)



CLLS interface

100 words (1000 sentences) from LEXSUB

Some translations for *severely*



CLLS example: *stiff*

1) *Even though it may be able to pump a normal amount of blood out of the ventricles, the stiff heart does not allow as much blood to enter its chambers from the veins.*
3) *One stiff punch would do it.*
7) *In 1968 when originally commissioned to do a cigarstore Indian, he rejected the stiff image of the adorned and phony native and carved " Blue Nose, " replica of a Delaware Indian.*

S	LEXSUB substitutes	CLLS translations
1	rigid 4; inelastic 1; firm 1; inflexible 1	duro 4; tieso 3; rigido 2; agarrotado 1; entumecido 1
3	strong 2; firm 2; good 1; solid 1; hard 1	duro 4; definitivo 1; severo 1; fuerte 1
7	stern 1; formal 1; firm 1; unrelaxed 1; constrained 1; unnatural 1; unbending 1	duro 2; forzado 2; fijo 1; rigido 1; acartonado 1; insipido 1

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:

WSsim Data

- Round 1 (Erk et al., 2009) 3 annotators
  - 8 lemmas (nouns, verbs and adjectives) 50 sentences each from SemCor (Miller et al., 1993) and SENSEVAL-3 English Lexical Sample (SE-3) (Mihalcea et al., 2004)
  - 3 lemmas data from LEXSUB 10 sentences each also in Usim
  - 430 sentences
- Round 2 : 26 lemmas (260 sentences) from LEXSUB, 8 annotators, extra context

In Round 2 we also collected traditional sense annotations (WSbest) and synonyms (SYNbest)

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

ITA (average spearmans) Round 1 $\rho = 0.50$ Round 2 $\rho = 0.60$ ($p < 2.2e - 16$)

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:

Usim Data

- Round 1: (Erk et al., 2009) 3 annotators
  - 34 lemmas (nouns, verbs, adjectives and adverbs) 10 sentences each from LEXSUB
  - 340 sentences
- Round 2 : 26 lemmas (260 sentences). As WSsim round 2 i.e. 8 annotators, extra context.

NB as before In Round 2 we also collected traditional sense annotations (WSbest) and synonyms (SYNbest)

Usim example:

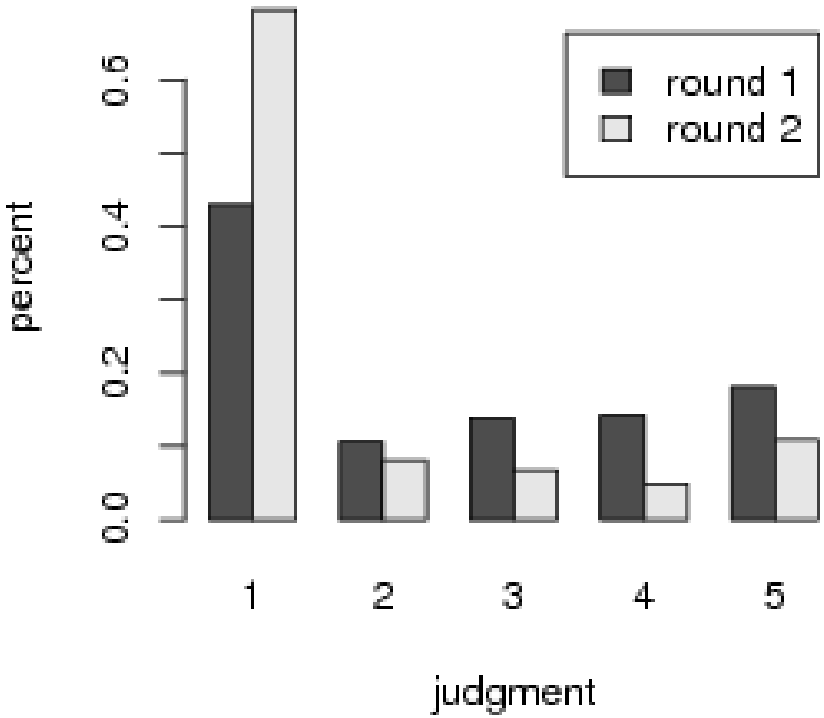
1) We study the methods and concepts that each writer uses to defend the cogency of legal, deliberative, or more generally political prudence against explicit or implicit charges that practical thinking is merely a knack or form of cleverness.

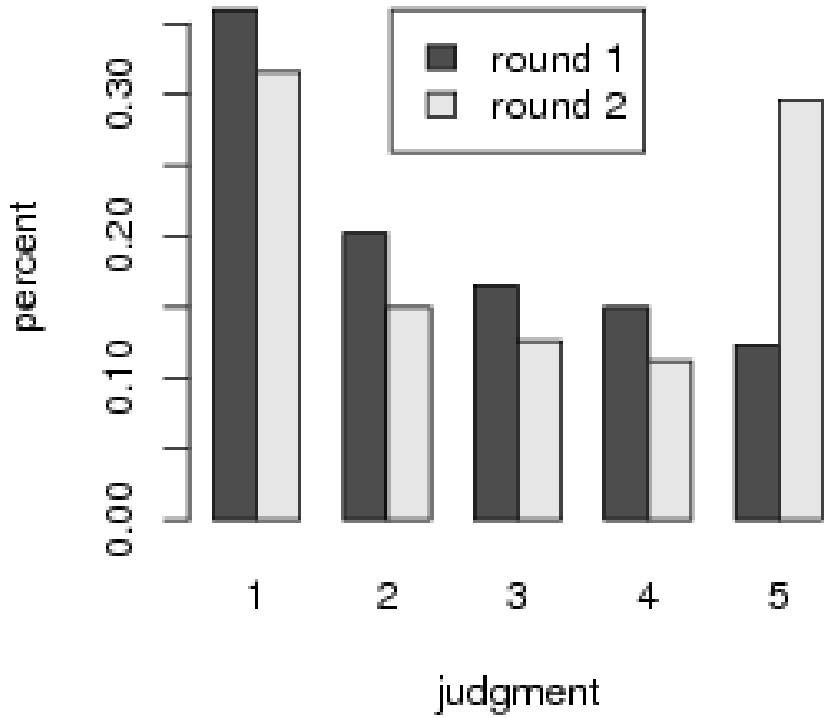
2) Eleven CIRA members have been convicted of criminal charges and others are awaiting trial.

Annotator judgments: 2,3,4

ITA (average spearmans) Round 1 $\rho = 0.55$ Round 2 $\rho = 0.62$ ($p < 2.2e - 16$)

WSsim and Usim R1 and R2 ratings





Outline

Analyses

- Are these datasets correlated?
- Do the WSsim responses suggest coarser groupings?
- Usim, paraphrases and translations correlations: can we predict cases of low inter-tagger agreement?

Calculations

- we use mean judgment from all annotators for Usim and WSsim, we use mode for WSbest
- Similarity/Distance between Sentence Pairs
  - WSsim we use Euclidean distance between vectors for each sentence
  - SYNbest and LEXSUB use overlap of multiset of substitutes to compare to measures on paired sentences

$$\text{Substitute Overlap: } \frac{|\textit{multiset intersection}|}{|\textit{larger multiset}|}$$

e.g. $S_1\{\textit{game, game, game, tournament}\}$

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

Correlation between datasets

tasks	Spearman's ρ
Usim-1 LEXSUB	0.590
Usim-2 SYNbest	0.764
WSsim-2 SYNbest	-0.749
WSsim-1 SemCor	0.426
WSsim-1 SE-3	0.419
WSsim-2 WSbest	0.483
Usim-2 WSsim-2	-0.816

Correlating senses: WSsim of two senses of *account*

WordNet sense	Sentence									
	1	2	3	4	5	6	7	8	9	10
account%1:10:00::	1.0	2.3	1.1	4.3	1.1	1.0	1.1	1.1	1.1	4.3
account%1.10:04::	1.5	3.0	1.3	2.9	1.5	1.5	1.6	1.0	1.4	3.9

Percentage of sense pairs that were significantly positively (pos) or negatively (neg) correlated

	$p < 0.05$		$p < 0.01$	
	pos	neg	pos	neg
Rd. 1	30.3	22.2	21.1	16.8
Rd. 2	14.3	11.1	8.0	4.6

Percentage of sentences with two uncorrelated or negatively correlated senses have judgments above a threshold

	$j \geq 3$	$j \geq 4$	$j = 5$
Rd. 1	69.3	33.0	9.1
Rd. 2	50.1	20.0	4.6

Lemmas in WSsim having coarse grained mappings

	R1		R2	
lemma	ON (Hovy et al., 2006)	EAW (Navigli et al., 2007)	ON	EAW
account.n			√	√
add.v	√			
ask.v	√	√		
call.v			√	√
coach.n			√	
different.a		√		
dismiss.v			√	√
fire.v			√	
fix.v			√	
hold.v			√	√
lead.n				√
new.a				√
order.v	√		√	
paper.n		√		
rich.a				√
shed.n			√	
suffer.v			√	√
win.v	√	√		

Sentences with positive judgments for senses in different coarse groupings

J.	OntoNotes Rd. 1		OntoNotes Rd. 2		EAW Rd. 1		EAW Rd. 2	
≥ 3	28%	(42)	52%	(52)	78%	(157)	62%	(50)
≥ 4	13%	(19)	16%	(16)	41%	(82)	22%	(18)
5	3%	(5)	3%	(3)	8%	(17)	6%	(5)

Sentences that have widely different judgments for pairs of senses in the same coarse grouping

		OntoNotes				EAW			
J1	J2	Rd. 1		Rd. 2		Rd. 1		Rd. 2	
≤ 2	≥ 4	35%	(52)	30%	(30)	20%	(39)	60%	(48)
≤ 2	5	11%	(16)	4%	(4)	2%	(4)	15%	(12)

Average Usim for R2 where WSbest annotations suggested the same or different coarse grouping

	ON		EAW	
	same	different	same	different
	4.0	1.9	4.1	2.0
	by lemma			
account.n	4.0	1.6	4.0	1.5
call.v	4.3	1.4	4.3	1.4
coach.n	4.6	2.3	-	-
dismiss.v	3.8	2.6	3.8	2.6
fire.v	4.6	1.2	-	-
fix.v	4.2	1.1	-	-
hold.v	4.5	2.0	3.8	1.9
lead.v	-	-	2.9	1.5
new.a	-	-	4.6	4.6
order.v	4.3	1.7	-	-
rich.a	-	-	4.6	2.0
shed.v	2.9	3.3	-	-
suffer.v	4.2	-	4.2	-

Paraphrases, translations and Usim analysis

- data common to CLLS, Usim-1 or -2 and LEXSUB
- 32 lemmas (Usim-1) + 24 lemmas (Usim-2) (4 lemmas in both)
- Usim take the mean judgments (as above)
- overlap in paraphrases and translations (as above)

Correlation between datasets

datasets	ρ
LEXSUB-CLLS	0.519
LEXSUB-Usim-1	0.576
LEXSUB-Usim-2	0.724
CLLS-Usim-1	0.531
CLLS-Usim-2	0.624

Correlation between datasets . . . by lemma

lemma	LEXSUB CLLS	LEXSUB Usim	CLLS Usim	Usim MID	Usim IAA
account.n	0.322	0.524	0.488	0.389	0.66
bar.n	0.583	0.624	0.624	0.296	0.35
bright.a	0.402	0.579	0.137	0.553	0.53
call.v	0.708	0.846	0.698	0.178	0.65
. . .	. . .	. . .	. . .	. . .	. . .

Correlation between datasets . . . by lemma

LEXSUB CLLS	LEXSUB Usim	CLLS Usim	Usim rev MID	Usim IAA
throw.v	lead.n	new.a	fresh.a	new.a
neat.a	hard.r	throw.v	raw.a	function.n
work.v	new.a	work.v	strong.a	fresh.a
strong.a	put.v	hard.r	special.a	investigator.n
. . .	. . .	. . .	. . .	. . .
dismiss.v	fire.v	rude.a	post.n	severely.r
coach.n	rich.a	coach.n	call.v	flat.a
fire.v	execution.n	fire.v	fire.v	fire.v

Correlation between datasets . . . by lemma

LEXSUB CLLS	LEXSUB Usim	CLLS Usim	Usim rev MID	Usim IAA
throw.v	lead.n	new.a	fresh.a	new.a
neat.a	hard.r	throw.v	raw.a	function.n
work.v	new.a	work.v	strong.a	fresh.a
strong.a	put.v	hard.r	special.a	investigator.n
. . .	. . .	. . .	. . .	. . .
dismiss.v	fire.v	rude.a	post.n	severely.r
coach.n	rich.a	coach.n	call.v	flat.a
fire.v	execution.n	fire.v	fire.v	fire.v
0.424	0.528	0.674	-0.486	

Summary

- Word meaning annotations using substitutes, translations, graded sense annotations and similarity judgments
- Annotations reflect underlying meanings in context and allow relationships between usages
- WSsim annotations indicate groupings are not straightforward for all lemmas
- Usim judgments alongside traditional WSD annotations might highlight difficult lemmas

...

Summary contd.

- Annotations of similarity of usage show highly significant correlation to substitutes and translations
- Correlation is not evident for all lemmas
- Correlation between these annotations by lemma itself correlates with Usim inter-tagger agreement
- Proportion of Usim mid scores by lemma is a useful indicator of low inter-tagger agreement and issues with separability of senses

Ongoing and future work

- Datasets available for evaluating different representations of meaning
- . . . particularly fully unsupervised
- Analysis of the extent that paraphrases and translations can be clustered

Credits

Thank you

Credits

Thank you

and thanks also to . . .
Collaboration with Roberto Navigli
and Katrin Erk and Nick Gaylord
and Rada Mihalcea, Ravi Sinha
and Huw McCarthy

Credits

Thank you

and thanks also to . . .
Collaboration with Roberto Navigli
and Katrin Erk and Nick Gaylord
and Rada Mihalcea, Ravi Sinha
and Huw McCarthy

- LEXSUB task web site:
  http://www.dianamccarthy.co.uk/task10index.html
- CLLS web site:
  http://lit.csci.unt.edu/index.php/Semeval_2010
- Usim and WSsim from websites of Katrin Erk and Diana

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.

Erk, K., McCarthy, D., and Gaylord, N. (2009). Investigations on word senses and word usages. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, Suntec, Singapore. Association for Computational Linguistics.

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

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2006). Ontonotes: The 90% solution. In *Proceedings of the HLT-NAACL 2006 workshop on Learning word meaning from*

non-linguistic data, New York City, USA. Association for Computational Linguistics.

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.

McCarthy, D. and Navigli, R. (2009). The English lexical substitution task. *Language Resources and Evaluation Special Issue on Computational Semantic Analysis of Language: SemEval-2007 and Beyond*, 43(2):139–159.

Mihalcea, R., Chklovski, T., and Kilgarriff, A. (2004). The SENSEVAL-3 english lexical sample task. In Mihalcea, R. and Edmonds, P., editors, *Proceedings SENSEVAL-3 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 25–28, Barcelona, Spain.

Mihalcea, R., Sinha, R., and McCarthy, D. (2010). Semeval-2010 task 2: Cross-lingual lexical substitution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 9–14, Uppsala, Sweden. Association for Computational Linguistics.

Miller, G. A., Leacock, C., Tengi, R., and Bunker, R. T. (1993). A semantic concordance. In *Proceedings of the ARPA Workshop on Human Language Technology*, pages 303–308. Morgan Kaufman.

Navigli, R., Litkowski, Kenneth, C., and Hargraves, O. (2007). SemEval-2007 task 7: Coarse-grained english all-words task. In *Proceedings of the 4th International Workshop on Semantic*

Evaluations (SemEval-2007), pages 30–35, Prague, Czech Republic.

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.

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

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.