Graded Annotations of Word Meaning in Context

Diana McCarthy

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

University of Melbourne, July 2011

Outline

Introduction

Alternative Word Meaning Annotations

- Graded Judgments (Usim and WSsim)

Analyses

- Correlation Between Datasets
- Sense Groupings
- Usim, Paraphrases and Translations
- Computational Models

Conclusions

Word Sense Representation and Disambiguation

Some conclusions so far

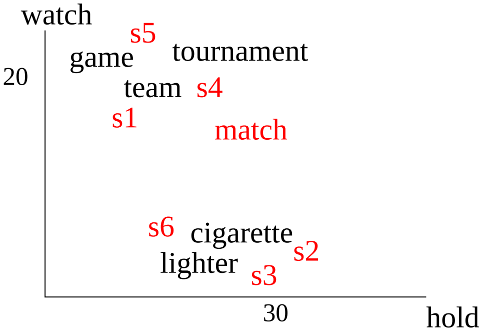
- what is the right inventory?
- how can we compare different representations?
- how to paraphrases and substitutes relate to sense annotations?
- are we right to assume groupings of word senses?

Manually produced inventories: e.g. WordNet

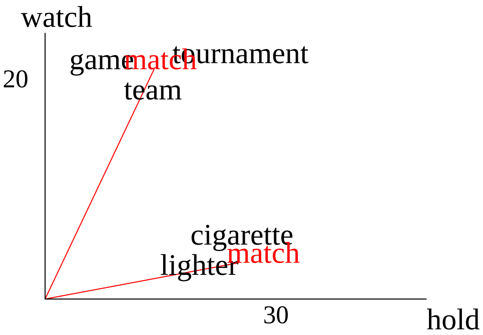
match has 9 senses in WordNet including:-

- 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")

Vector based models



Vector based models



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.

- 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")

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.

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

match#n#9

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

What is the right inventory?

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

- should we enumerate senses?
- will it help applications?
- how can we test different inventories?

What is the right inventory?

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

- should we enumerate senses?
- will it help applications?
- how can we test different inventories?

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	↔	counter	↔	rigid block of wood
child	young person	↔	offspring	↔	descendant

Alternative word meaning annotations: datasets

to compare different representations of word meaning in context

- SemEval-2007 Lexical Substitution (LEXSUB) (McCarthy and Navigli, 2007)(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 . . .

LEXSUB and 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.*

LEXSUB and 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 and Usim

- new datasets to explore subtler representations of *sense*
- modelled as psycholinguistic experiment: no right or wrong answer
- use multiple annotators and check consensus
- WSsim (word sense similarity) for a given context of a word, rate every sense in terms of its relevance on a graded scale (1-5)
- Usim (usage similarity) for a pair of contexts of a word, rate the pair in terms of similarity of use on a graded scale (1-5)

WSsim and Usim: motivations

- compare to existing annotations, paraphrases and translations
- WSsim
  - explore the extent that multiple senses apply with less bias to annotators
  - explore whether graded annotations are explained by sense groupings
- Usim
  - examine phenomena without a predefined sense inventory

Annotation

- 2 rounds
- all annotators native English speakers
- nouns, verbs, adjectives, adverbs (1st round adverbs only Usim)

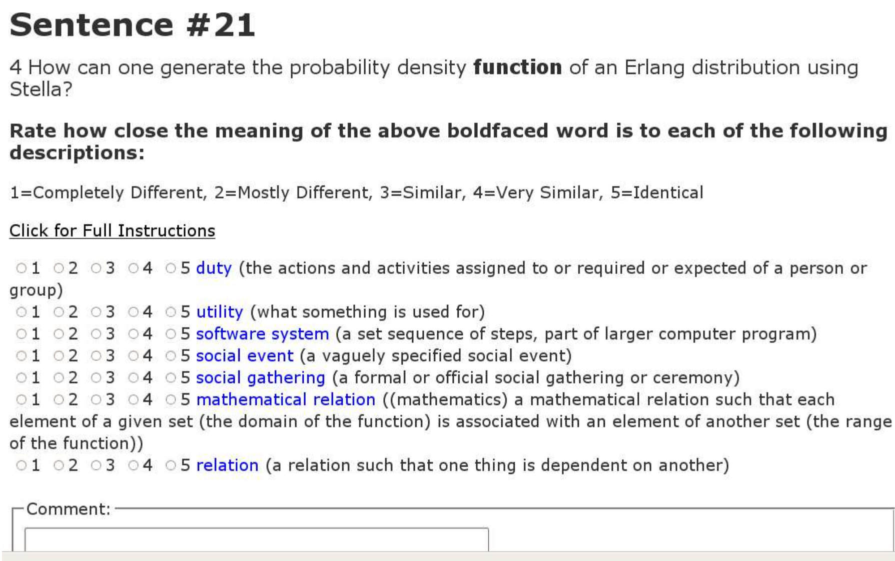
Round 1 Erk et al. (2009)

- 3 annotators for Usim, and 3 for WSsim (1 did both)
- no particular expertise (ages, undergrad → early 50s, all women)
- one sentence of context for each target instance
- data released (http://www.katrinerk.com/graded-sense-and-usage-annotation)

Round 2

- 8 annotators , all doing all for tasks
- one phd comp linguistics (rest not, but 2 had done round 1)
- 4 men, 4 women (ages 18-early 50s)
- Usim WSsim, traditional word sense tagging WSbest, lexical substitution SYNbest
  - group 1: Usim, SYNbest, WSsim, WSbest
  - group 2: Usim, SYNbest, WSbest,WSsim
- 2 sentences of context for each instance, an extra sentence either side of that with target
- data to be released on publication (from http://www.dianamccarthy.co.uk/)
- part of Usim-2 released already (Cicling 2011, with R code)

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:

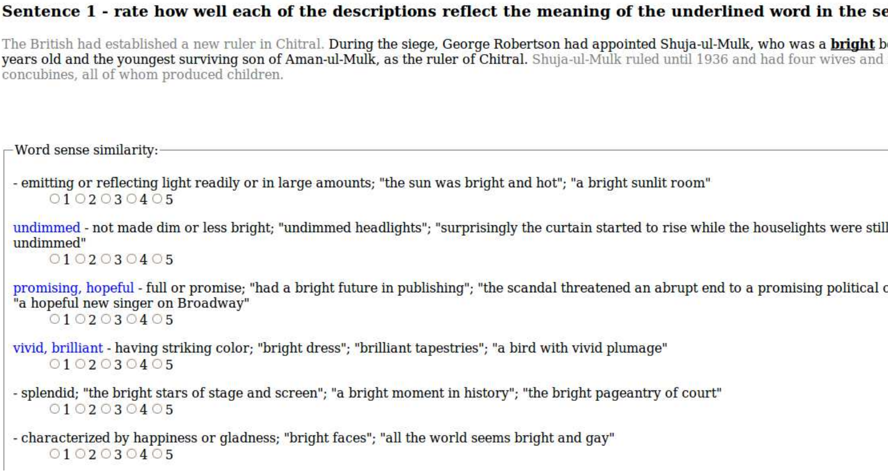
1=Completely Different, 2=Mostly Different, 3=Similar, 4=Very Similar, 5=Identical

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-2 interface



Sentence 1 - rate how well each of the descriptions reflect the meaning of the underlined word in the se

The British had established a new ruler in Chitral. During the siege, George Robertson had appointed Shuja-ul-Mulk, who was a **bright** b years old and the youngest surviving son of Aman-ul-Mulk, as the ruler of Chitral. Shuja-ul-Mulk ruled until 1936 and had four wives and concubines, all of whom produced children.

Word sense similarity:

- emitting or reflecting light readily or in large amounts; "the sun was bright and hot"; "a bright sunlit room"
○1 ○2 ○3 ○4 ○5

undimmed - not made dim or less bright; "undimmed headlights"; "surprisingly the curtain started to rise while the houselights were still undimmed"
○1 ○2 ○3 ○4 ○5

promising, hopeful - full or promise; "had a bright future in publishing"; "the scandal threatened an abrupt end to a promising political c "a hopeful new singer on Broadway"
○1 ○2 ○3 ○4 ○5

vivid, brilliant - having striking color; "bright dress"; "brilliant tapestries"; "a bird with vivid plumage"
○1 ○2 ○3 ○4 ○5

- splendid; "the bright stars of stage and screen"; "a bright moment in history"; "the bright pageantry of court"
○1 ○2 ○3 ○4 ○5

- characterized by happiness or gladness; "bright faces"; "all the world seems bright and gay"
○1 ○2 ○3 ○4 ○5

WSsim Data

- Round 1 (Erk et al., 2009)
  - 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,

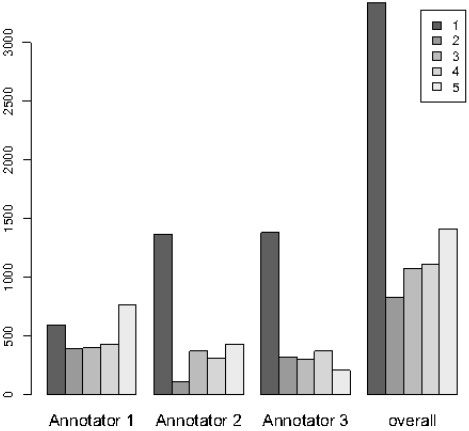
WSsim example

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

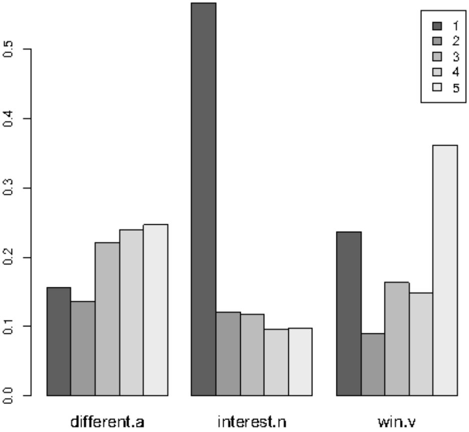
The senses are: 1:statement, 2:controversy, 3:debate, 4:literary argument, 5:parameter, 6:variable, 7:line of reasoning

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

WSsim number of times each judgment was used, by annotator and summed over all annotators (R1)



Usim percentage of times each judgment was used for the lemmas *different.a*, *interest.n* and *win.v* summed over 3 annotators (R1)



Percentage of items with multiple senses assigned

Orig: in the original SemCor/SE-3 data. WS*sim judgment*: items with judgments at or above the specified threshold. R1

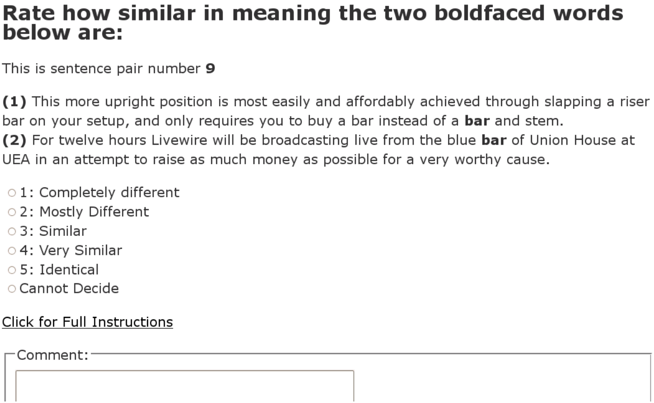
		WSsim judgment		
Data	Orig.	≥ 3	≥ 4	5
WSsim/SemCor	0.0	80.2	57.5	28.3
WSsim/SE-3	24.0	78.0	58.3	27.1
All WSsim		78.8	57.4	27.7

Overall, 0.3% of tokens in SemCor have multiple labels, and 8% of tokens in SE-3, so the multiple label assignment in our sample is not an underestimate.

WSsim multiple senses having highest response

	Proportion of sentences with multiple senses having highest response
WSsim-1	0.46
WSsim-2	0.30
WSsim-2 group 1	0.36
WSsim-2 group 2	0.23

Usim interface



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

This is sentence pair number **9**

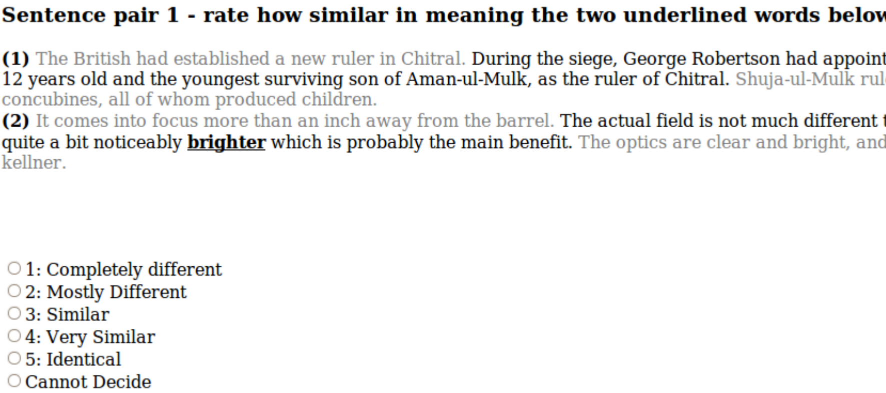
(1) This more upright position is most easily and affordably achieved through slapping a riser bar on your setup, and only requires you to buy a bar instead of a **bar** and stem.
(2) For twelve hours Livewire will be broadcasting live from the blue **bar** of Union House at UEA in an attempt to raise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



Sentence pair 1 - rate how similar in meaning the two underlined words belo

(1) The British had established a new ruler in Chitral. During the siege, George Robertson had appoint
12 years old and the youngest surviving son of Aman-ul-Mulk, as the ruler of Chitral. Shuja-ul-Mulk rul
concubines, all of whom produced children.
(2) It comes into focus more than an inch away from the barrel. The actual field is not much different
quite a bit noticeably **brighter** which is probably the main benefit. The optics are clear and bright, and
kellner.

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

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:

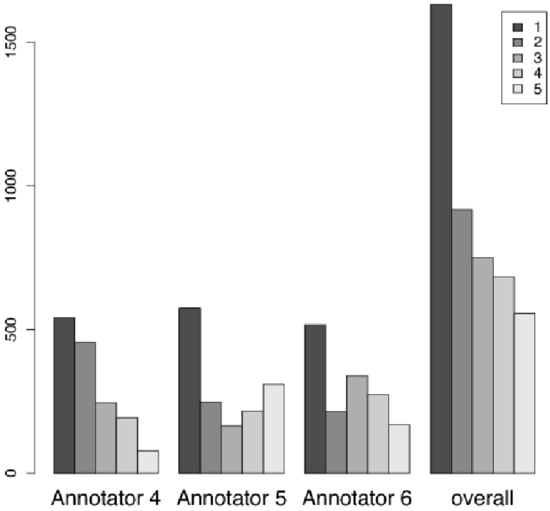
1) We study the methods and concepts that each writer uses to defend the cogency of legal, deliberative, or more generally political prudence against explicit or implicit charges that practical thinking is merely a knack or form of cleverness.

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

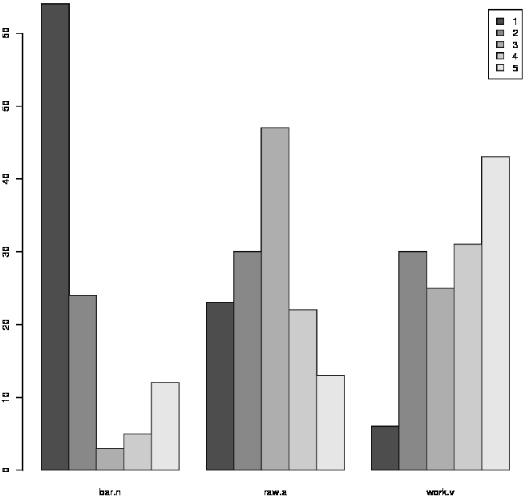
Annotator judgments: 2,3,4

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

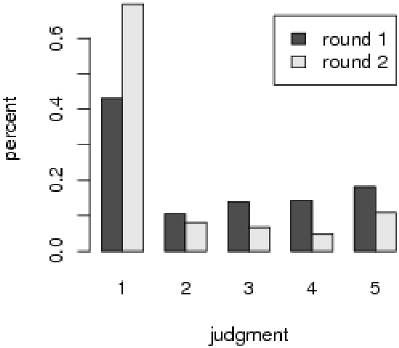
Usim number of times each judgment was used, by annotator and summed over all annotators (R1)

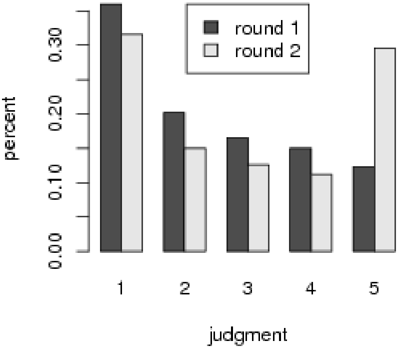


Usim number of times each judgment was used for *bar.n*, *work.v* and *raw.a* (R1)



WSsim and Usim R1 and R2 ratings

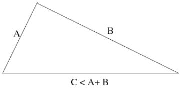




The relative frequency of the annotations at each judgment from all annotators

	Judgment				
Exp	1	2	3	4	5
WSsim-1	0.43	0.106	0.139	0.143	0.181
WSsim-2	0.696	0.081	0.067	0.048	0.109
WSsim-2 group 1	0.664	0.099	0.069	0.048	0.12
WSsim-2 group 2	0.727	0.063	0.065	0.048	0.097
Usim-1	0.360	0.202	0.165	0.150	0.123
Usim-2	0.316	0.150	0.126	0.112	0.296

Triangular inequality



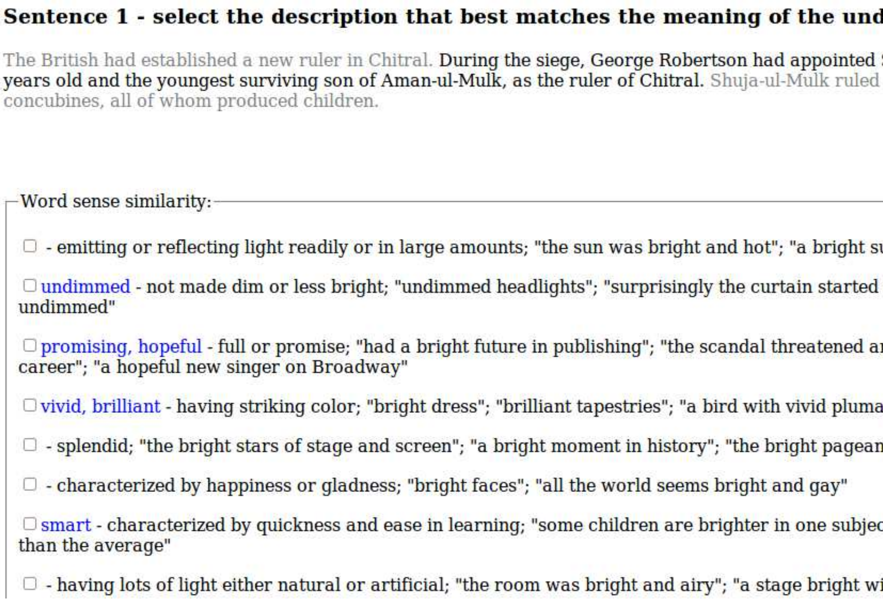
$$missed\ \ by =$$

$$max(length(longest) - (length(second\ \ longest) + length(shortest))0)$$

i.e. 0 where the triangular inequality holds.

	% obey	missed by (if missed)
Usim-1	99.2	0.520
Usim-2	100	-

wsbest interface



Sentence 1 - select the description that best matches the meaning of the und

The British had established a new ruler in Chitral. During the siege, George Robertson had appointed
years old and the youngest surviving son of Aman-ul-Mulk, as the ruler of Chitral. Shuja-ul-Mulk ruled
concubines, all of whom produced children.

Word sense similarity:

□ - emitting or reflecting light readily or in large amounts; "the sun was bright and hot"; "a bright su

□ undimmed - not made dim or less bright; "undimmed headlights"; "surprisingly the curtain started undimmed"

□ promising, hopeful - full or promise; "had a bright future in publishing"; "the scandal threatened a career"; "a hopeful new singer on Broadway"

□ vivid, brilliant - having striking color; "bright dress"; "brilliant tapestries"; "a bird with vivid pluma

□ - splendid; "the bright stars of stage and screen"; "a bright moment in history"; "the bright pagean

□ - characterized by happiness or gladness; "bright faces"; "all the world seems bright and gay"

□ smart - characterized by quickness and ease in learning; "some children are brighter in one subjec than the average"

□ - having lots of light either natural or artificial; "the room was bright and airy"; "a stage bright wi

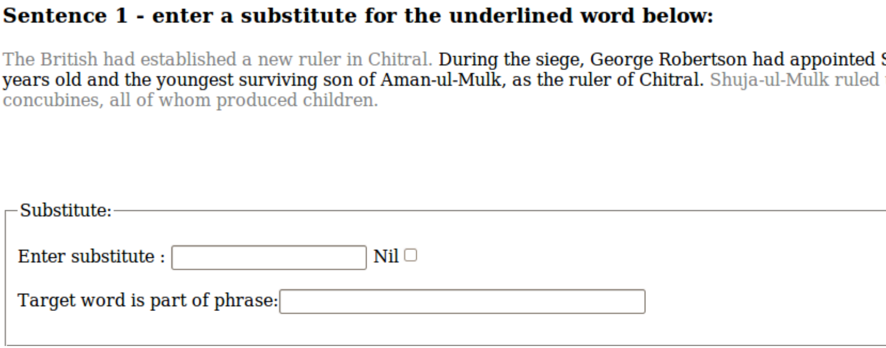
WSbest annotations

	sense selected		Proportion with
	n	y	multiple choice
WSbest	19599	2401	0.13
WSbest group 1	9779	1221	0.15
WSbest group 2	9820	1180	0.11

$$ITA \text{ WSbest} = \sum_{i \in I} \frac{\sum_{\{a_i, a'_i\} \in P_i} \frac{a_i \cap a'_i}{max(|a_i|,|a'_i|)}}{|P_i| \cdot |I|}$$

ITA 0.574 or 0.626 for items with 1 response from both in pair

SYNbest interface



Sentence 1 - enter a substitute for the underlined word below:

The British had established a new ruler in Chitral. During the siege, George Robertson had appointed S
years old and the youngest surviving son of Aman-ul-Mulk, as the ruler of Chitral. Shuja-ul-Mulk ruled
concubines, all of whom produced children.

Substitute:

Enter substitute : ▭ Nil □

Target word is part of phrase: ▭

PA = 0.261 (LEXSUB 0.278)

Outline

Introduction

Alternative Word Meaning Annotations

- Graded Judgments (Usim and WSsim)

Analyses

- Correlation Between Datasets
- Sense Groupings
- Usim, Paraphrases and Translations
- Computational Models

Conclusions

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 for Comparing Datasets

- we use mean judgment from all annotators for Usim and WSsim, we use mode for WSbest
- for traditional WSD methodology we assume scores of 1 and 5 (no match vs match)
- 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

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

$$\text{e.g. } S_1\{\mathit{game}, \mathit{game}, \mathit{game}, \mathit{tournament}\}$$

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

Correlation of WSsim with traditional methodology

	Original Gold Standard	
Exp	SemCor	SE-3
WSsim-1 Ann1 ρ	0.234	0.346
WSsim-1 Ann2 ρ	0.448	0.449
WSsim-1 Ann3 ρ	0.390	0.338
WSsim-1 Average Ind ρ	0.357	0.378
WSsim-1 mean ρ	0.426	0.419

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

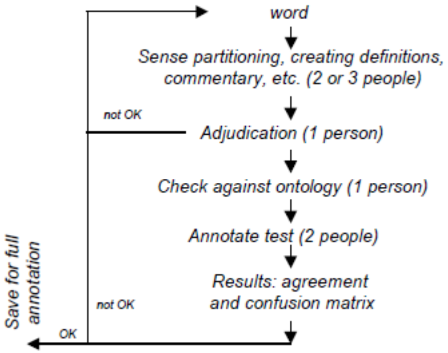
	$j \geq 3$	$j \geq 4$	$j = 5$
Rd. 1	69.3	33.0	9.1
Rd. 2	50.1	20.0	4.6

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	√	√		

Ontonotes Annotation Procedure

This figure is from Hovy et al. (2006)


word
Sense partitioning, creating definitions, commentary, etc. (2 or 3 people)
not OK
Adjudication (1 person)
Check against ontology (1 person)
Annotate test (2 people)
Results: agreement and confusion matrix
not OK
OK
Save for full annotation

WordNet 2.1 senses of the noun *account*, and groups in OntoNotes (ON) and EAW (ODE)

WordNet sense	WordNet key	ON group	EAW group
business relationship: ”he asked to see the executive who handled his account”	account%1:26:00::	1.1	5
report: ”by all accounts they were a happy couple”	account%1:10:05::	1.2	2
explanation: ”I expected a brief account”	account%1:10:04::	1.2	2
history, story: ”he gave an inaccurate account of the plot [...]”	account%1:10:00::	1.3	2
report, story: ”the ac-	account%1:10:03::	1.3	2

Sentences with positive judgments for senses in different coarse groupings

	OntoNotes				EAW			
J.	Rd. 1		Rd. 2		Rd. 1		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

	LEXSUB	LEXSUB	CLLS	Usim	Usim
lemma	CLLS	Usim	Usim	MID	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	LEXSUB	CLLS	Usim	Usim
CLLS	Usim	Usim	rev MID	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	LEXSUB	CLLS	Usim	Usim
CLLS	Usim	Usim	rev MID	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	

WSsim Computational Models: motivations

- could classic models be used to predict graded ratings?
- would vector space models outperform these if provided with training data to partition senses?

Preliminary Modelling of WSsim

- Gold standard provides vector of ratings, one for each sense
- mapped judgments 1-5 → 0-1
- Traditional vs Prototype models
- experiment with WSsim-1 lemmas in SemCor and SENSEVAL

Lemmas in this Study

lemma (PoS)	# senses	# training SemCor	 SE-3
add (v)	6	171	238
argument (n)	7	14	195
ask (v)	7	386	236
different (a)	5	106	73
important (a)	5	125	11
interest (n)	7	111	160
paper (n)	7	46	207
win (v)	4	88	53
total training sentences		1047	1173

Models

- Classic Binary (one classifier per sense)
- Max Entropy `http://maxent.sourceforge.net/` (n-ary slightly worse)
- 2 models:
  - best (traditional 0 vs 1)
  - conf (confidence used as rating)

Models: feature representation

feature representation of a sentence. e.g. features for *add* in BNC occurrence *For sweet-sour sauce, cook onion in oil until soft.* **Add** *remaining ingredients and bring to a boil.*

Cx/2 (Cx/50): context of size 2 (size 50) either side of the target.

Ch: children of target.

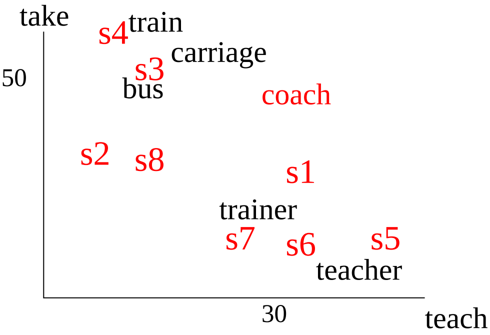
Cx/2	until, IN, soft, JJ, remaining, VBG, ingredient, NNS
Cx/50	for, IN, sweet-sour, NN, sauce, NN, . . . , to, TO, a, DT, boil, NN
Ch	OA, OA/ingredient/NNS

Models: traditional

- Use traditional best fitting training data to obtain probabilistic WSD models
  - Best: best fitting senses
  - Conf: probability over senses

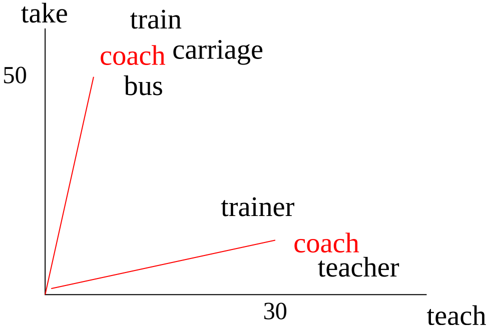
Models: Vector Space-Based

Use vector space models which take best fitting training data
Instead of:



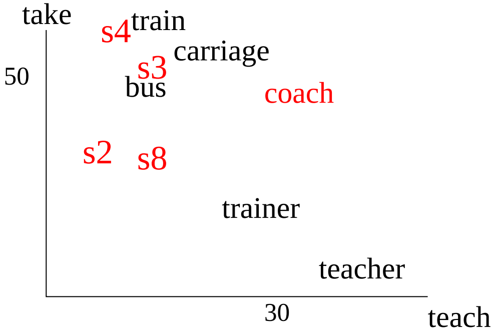
Models: Vector Space-Based

Use vector space models which take best fitting training data
Instead of:



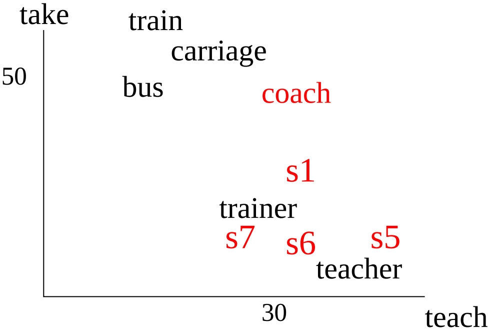
Models: Vector Space-Based

Use vector space models which take best fitting training data



Models: Vector Space-Based

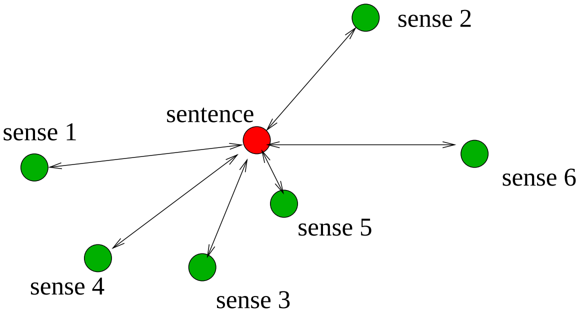
Use vector space models which take best fitting training data



Models: Vector Space-Based

- use training data to create prototypes
- the DV package, `http://www.nlpado.de/~sebastian/dv.html`, to compute the vector space.
- one prototype per sense
- same feature representation of a sentence as traditional models
- centroid of vectors for sense (not using ‘negative’ evidence for different senses)
- classify an occurrence by distance to *each* sense

Models: Vector Space-Based



Models: Vector Space-Based

- *Prototype* first order, counts words in sentence
- Prototype-2 second order for each sentence
  - compute vector for each word
  - sentence vector is centroid of word vectors
- prototype-n prototype-2n normalised judgments for each sentence $(\frac{assigned}{sum \ \ for \ \ all \ \ senses \ \ for \ \ that \ \ item})$

Correlation between Gold-Standard and Model

lemma: for each lemma $\ell \in L$, compute correlation between $G|_{lemma=\ell}$ and $A|_{lemma=\ell}$

sense: for each lemma ℓ and each sense number $i \in S_\ell$, compute correlation between $G|_{lemma=\ell,senseno=i}$ and $A|_{lemma=\ell,senseno=i}$

token: for each lemma ℓ and sentence number $t \in T$, compute correlation between $G|_{lemma=\ell,sentence=t}$ and $A|_{lemma=\ell,sentence=t}$

Jenson Shannon divergence

Symmetric version of kullback-Leibler divergence of probabilities

$$JS(p,q) = \frac{1}{2}(D(p||\frac{p+q}{2}) + D(q||\frac{p+q}{2})$$

Compare distributions given lemma and sentence

Graded precision and recall

$$P_\ell = \frac{\sum_{i \in S_\ell, t \in T} \min(\text{gold}_{\ell,i,t}, \text{assigned}_{\ell,i,t})}{\sum_{i \in S_\ell, t \in T} \text{assigned}_{\ell,i,t}}$$

and

$$R_\ell = \frac{\sum_{i \in S_\ell, t \in T} \min(\text{gold}_{\ell,i,t}, \text{assigned}_{\ell,i,t})}{\sum_{i \in S_\ell, t \in T} \text{gold}_{\ell,i,t}}$$

- macro averaged by lemma
- precision decrease if model overshoots
- recall decreases as model undershoots
- classical precision and recall if data is categorial.

Experimental set up

- training:
  - SemCor (minus WSsim)
  - SE-3 (minus WSsim)
- human ceiling : evaluate performance of one annotator against other two
- baseline: most frequent sense from corpus

Human ceiling: one annotator vs. average of the other two annotators

Avg: average annotator performance

	lemma	sense	token				
Ann	ρ	ρ	ρ	J/S	P	R	F
Ann.1	0.517	0.407	0.482	0.131	50.6	87.5	64.1
Ann.2	0.587	0.403	0.612	0.153	75.5	62.4	68.3
Ann.3	0.528	0.41	0.51	0.165	82.4	52.3	64.0
Avg	0.544	0.407	0.535	0.149	69.5	67.4	65.5

Evaluation: computational models, and baseline.

Model	l ρ	s ρ	t ρ	J/S	P	R	F
best	0.267	0.053	0.28	0.39	58.7	25.5	35.5
conf	0.396	0.177	0.401	0.164	81.8	37.1	51.0
Prototype	0.245	0.053	0.396	0.173	58.4	78.3	66.9
Prototype/2	0.292	0.086	0.478	0.164	68.2	63.3	65.7
Prototype/N	0.396	0.137	0.396	0.173	82.2	29.9	43.9
Prototype/2N	0.465	0.168	0.478	0.164	82.6	30.9	45.0
baseline	0.338	0.0	0.355	0.167	79.9	34.5	48.2

Average judgment for individual annotators (transformed) and average rating for models

Ann.	avg	Model	avg
Ann.1	0.540	*WSD/single*	0.163
Ann.2	0.345	*WSD/conf*	0.173
Ann.3	0.285	*Prototype*	0.558
		Prototype/N	0.143
		Prototype/2	0.375
		Prototype/2N	0.143
		baseline	0.167

Computational Modelling of Usim

- Contrast vector space models with WordNet
- Vector space model using DV package, `http://www.nlpado.de/~sebastian/dv.html`
  - minipar parses of BNC
  - frequency, relative frequency, pmi
  - centroid or best (closest vector of words in sentence to target)
  - correlation with average judgement best higher correlation some significance but ρ really small

Computational Modelling of Usim

- WordNet: lesk
- all words (max WordNet similarity in two sentences)
- best (WordNet similarity between 2 words that are closest to target
- Results show no correlation or wrong direction

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

Thank You

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

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 McCarthy

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.

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. (2007). SemEval-2007 task 10: English lexical substitution task. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53, Prague, Czech Republic.

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.

Tuggy, D. H. (1993). Ambiguity, polysemy and vagueness. *Cognitive linguistics*, 4(2):273–290.