Evaluating Word Sense Inventories and Disambiguation using Lexical Substitution

Diana McCarthy

University of Melbourne, July 2011

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

- most work on disambiguation uses pre-defined man-made inventory
- there is widespread concern that the distinctions are not appropriate
- how can we compare the merits of disambiguation techniques without specifying the inventory

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.

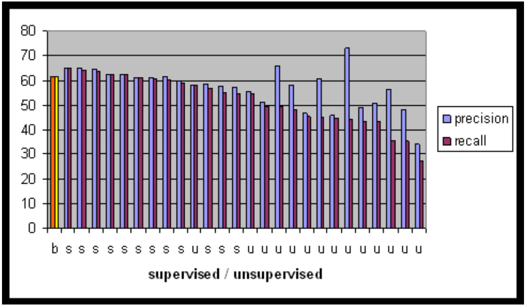
match#n#2

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
- increase in the range of languages
- man-made inventories used, especially WordNet

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] vs [Chan et al., 2007, Carpuat and Wu, 2007]

What is the Right Inventory?

- WordNet often used
- 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?

What about distributional similarity representations?

- disambiguation tasks require mapping to gold standard inventory, but is the gold inventory appropriate?
- task-based methods e.g. information retrieval ([Schütze, 1998]) avoid the need to agree an inventory
  - pro: inventory is relevant to the task
  - cons: conflating evaluation of inventory/representation with evaluation of disambiguation
  - many applications will require complex systems which
    - favour large teams at the expense of individual researchers/students
    - mask the impact of disambiguation due to the numerous other components

Outline

Key Issues

How can we:

- determine the distinctions useful for WSD systems?
- compare inventories of meaning?
- compare disambiguation techniques without fixing the inventory?

Key Issues

How can we:

- determine the distinctions useful for WSD systems?
- compare inventories of meaning?
- compare disambiguation techniques without fixing the inventory?

Our idea: lexical substitution

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.

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 NB NO training data
- English Internet Corpus [Sharoff, 2006]
- sentences selected
  - manually for 20 words in each PoS
  - rest selected automatically

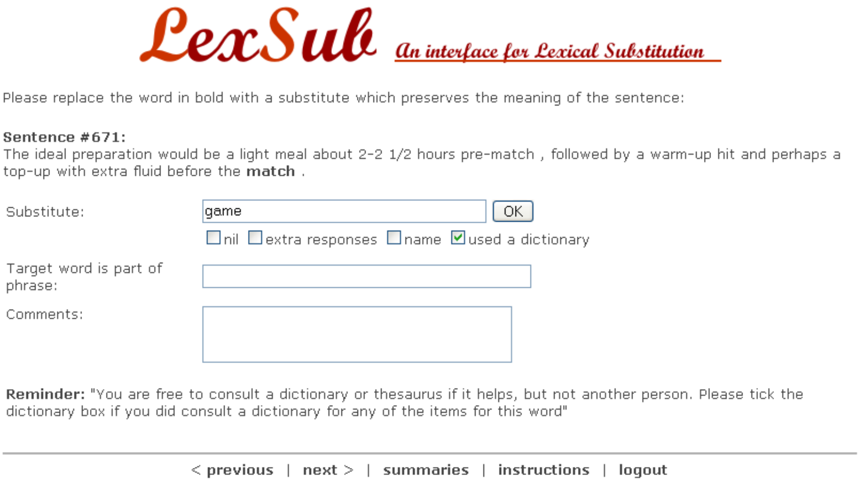
Annotators

- 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



LexSub An interface for Lexical Substitution

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"

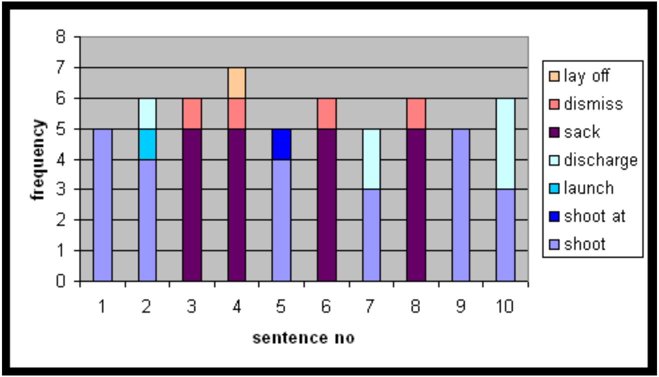
< previous | next > | summaries | instructions | logout

Example

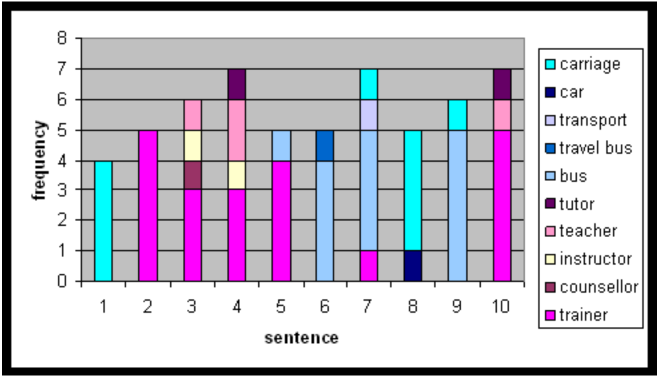
In the event of a chemical spill, 3/4's of the children know that they should evacuate (leave area) as advised on radio, TV, or by people in charge.

Annotator	1	2	3	4	5
substitutes	control, command	control	authority	power	command

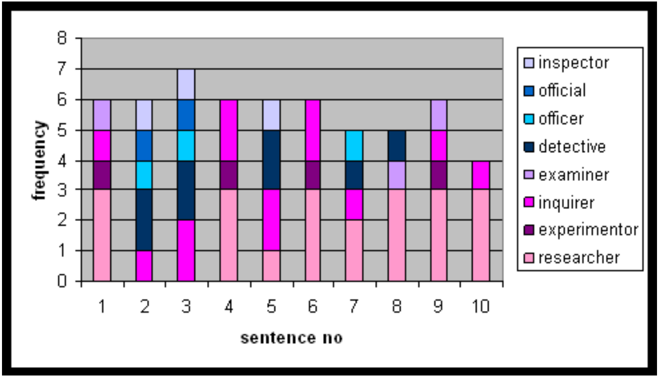
Substitutes for *fire* (verb)



Substitutes for *coach* (noun)



Substitutes for *investigator* (noun)



pairwise agreement

The average proportion of all the paired responses for which the two paired annotators gave the same response.

- I is set of instances
- Pairwise agreement between every possible pairing of annotators (P_i) for each item
- h_i is a set of substitutes from one annotator in the pairing.
- HI_m is set of all non empty h_i for items in I_m (those with a mode)

$$pa = \sum_{i \in I} \frac{\sum_{\{h_i, h_i'\} \in P_i} \frac{h_i \cap h_i'}{h_i \cup h_i'}}{|P_i| \cdot |I|} \times 100 = 27.7 \ (31.13)$$

$$pa_m = \frac{\sum_{i \in I_m} \sum_{h_i \ : \ h \in H} \frac{1 \ if \ m_i \in h_i}{|h_i|}}{|HI_m|} \times 100 = 50.7 \ (64.7)$$

Agreement

pairwise agreement between every possible pairing (P)

PoS	#	pa	% 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

Baselines: From WordNet

For a target word:

1. synonyms from the first synset (ranked with frequency data from the BNC)
2. synonyms from closely related classes of that first synset (ranked with the BNC frequency)
3. synonyms from all synsets (ranked using the BNC frequency)
4. synonyms from all closely related classes of all synsets of the target (ranked with the BNC frequency)

Baselines: Using Distributional Scores

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

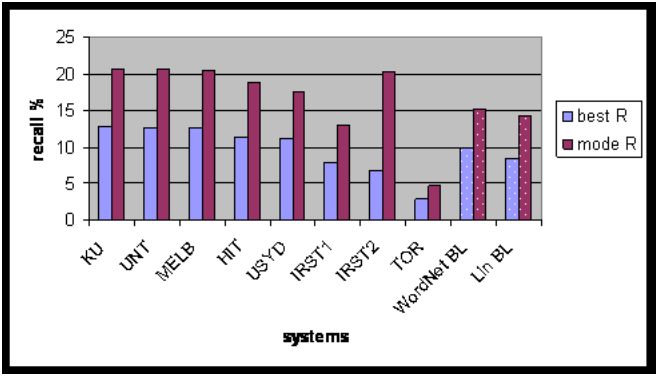
10 Systems (8 teams): inventories

Systems	WordNet	Macquire	Roget	Other
MELB	$\surd$			
HIT	$\surd$			
UNT	$\surd$			Encarta
IRST1	$\surd$			OAWT
IRST2	$\surd$			OAWT
KU			$\surd$	
SWAG1			$\surd$	
SWAG2			$\surd$	
USYD	$\surd$	$\surd$		Web 1T corpus
TOR		$\surd$		

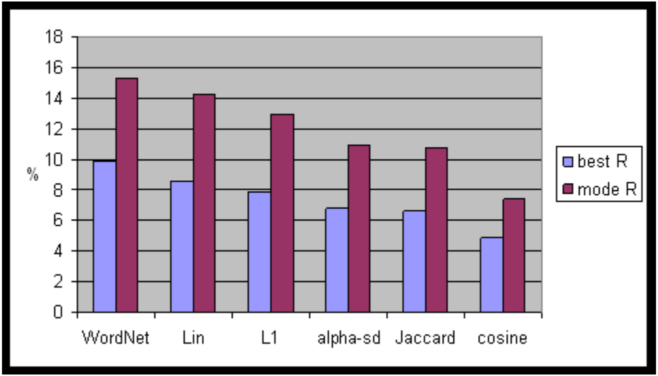
10 Systems: approaches

Systems	google	Web 1T	BNC	Sense tags	other
MELB	n-gram			SemCor	
HIT	n-gram				
UNT	n-gram	n-gram	morph	SemCor	TE+Wiki+GA
IRST1			LSA		
IRST2		n-gram			
KU		n-gram			
SWAG1		n-gram			
SWAG2		n-gram	freq vectors		
USYD		*pMI*			
TOR			*pMI*+freq		

best results



best Baseline Results



best recall results by PoS

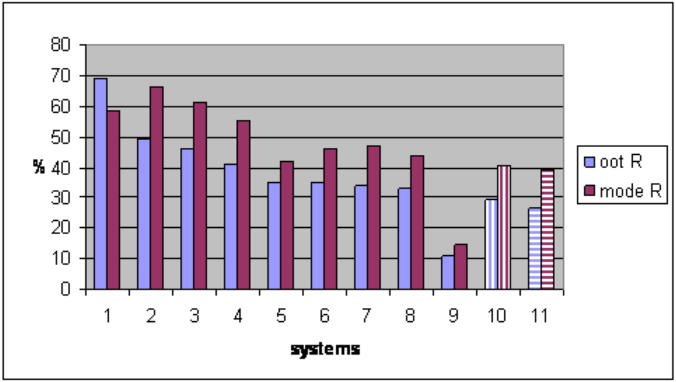
systems	all	nouns	verbs	adjectives	adverbs
KU	**12.90**	12.14	**10.68**	**13.92**	15.85
UNT	12.77	**12.26**	7.90	12.25	21.63
MELB	12.68	9.41	9.01	12.94	**23.09**
HIT	11.35	11.91	6.47	9.54	20.43
USYD	10.88	11.01	8.31	9.60	16.46
IRST1	8.06	8.29	6.20	7.81	10.81
IRST2	6.94	5.77	4.65	6.89	12.33
TOR	2.98	2.79	0.99	4.04	4.59
WordNet bl	9.95	8.14	7.16	6.99	21.69
Lin bl	8.53	**12.52**	5.16	7.97	7.76

$$best\ \ upper\ \ bound = \frac{\sum_{i \in I} \frac{freq_{most\ \ freq\ \ substitute_i}}{|H_i|}}{|I|} \times 100 = 0.4576$$

best baseline results by PoS

systems	all	nouns	verbs	adjectives	adverbs
WordNet	9.95	8.14	**7.16**	6.99	**21.69**
lin	8.53	**12.52**	5.16	7.97	7.76
l1	7.82	10.22	6.14	7.32	7.13
lee	6.74	9.39	2.99	**8.50**	5.15
jaccard	6.60	8.86	4.37	5.96	7.15
cos	4.89	6.79	1.99	5.14	5.62
Roget	4.65	1.99	5.47	4.85	7.51

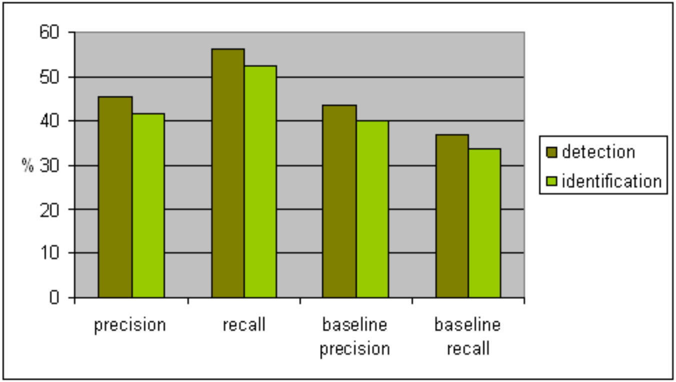
oot results



oot recall: NB duplicate issue!

systems	all	nouns	verbs	adjectives	adverbs	lwD
IRST2	68.90	**57.66**	**46.49**	**68.90**	**120.66**	1232
USYD	34.96	33.14	41.10	29.96	36.71	443
TOR	11.19	9.94	6.12	10.21	22.28	371
UNT	49.19	**48.07**	**44.24**	47.80	**60.54**	0
KU	46.15	40.84	39.78	**51.07**	56.72	0
IRST1	41.20	38.48	32.18	43.12	56.07	0
SWAG2	34.66	22.63	31.56	42.19	47.46	0
HIT	33.88	32.13	29.25	29.22	50.89	0
SWAG1	32.83	27.95	28.75	42.19	32.33	0

MWE results


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

Analysis: Finding the candidates

$$Type\ U = \frac{\sum_{wp \in WP} |GU_{wp} \cap SU_{wp}|}{|WP|}$$

where GU is union of substitute types from annotators for all 10 sentences for word and pos (wp)
SU is union of substitute types from the system for all 10 sentences for word and pos (wp)

Analysis: Finding the candidates

Systems	Type U	#subs	$TypeU_{uniq}$
KU	**2.88**	6.30	**0.58**
USYD	2.58	**7.51**	0.54
IRST2	2.57	5.50	0.29
MELB	1.91	3.77	0.27
HIT	1.87	4.29	0.18
IRST1	1.65	4.22	0.35
UNT	1.60	2.90	0.30
TOR	0.70	3.66	0.14

Analysis: Disambiguating the candidates

Mode Precision where mode found in SU

- **All** All systems found the mode within their SU_{wp} (NB there were only 17 such items)
- **Sys** The given system found the mode within its SU_{wp}

That is, precision is calculated as:

$$All\ precision = \sum_{bg_i \in All} \frac{1\ if\ \ bg_i = m_i}{|All|} \tag{1}$$

and

$$Sys\ precision = \sum_{bg_i \in Sys} \frac{1\ if\ \ bg_i = m_i}{|Sys|} \tag{2}$$

Analysis: Disambiguating the candidates

Systems	*SU* of All	*SU* of this System
HIT	**58.82**	52.53
UNT	52.94	**59.67**
KU	52.94	42.31
MELB	47.06	53.71
USYD	47.06	37.77
IRST2	41.18	44.57
IRST1	35.29	43.82
TOR	23.53	37.91

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 (1342) from original annotators and systems
- good, reasonable, bad

The Post-Hoc Annotation Interface

LexSub POST-HOC

An interface for Lexical Substitution

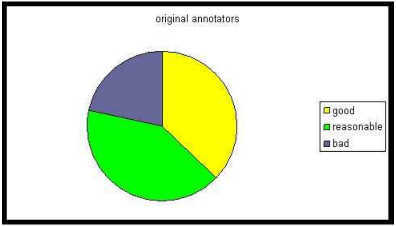
Please rate the quality of the candidate substitutes for the word in bold in the sentence below:

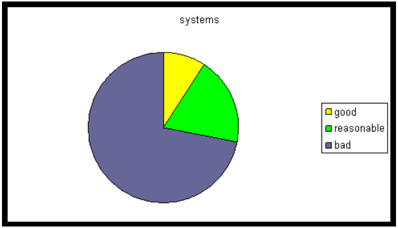
Sentence #675:
Other costs (**match** day , ground and administration) were down by 12 % on 2001/02 levels .

candidate substitutes:

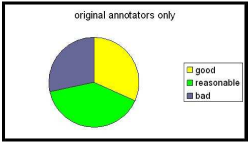
fire	bad
event	bad
equal	bad
couple	bad
tournament	bad
family	bad
contest	bad
game	bad
test	bad

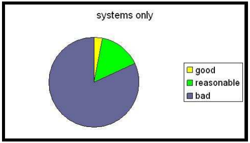
Post-Hoc Verdicts

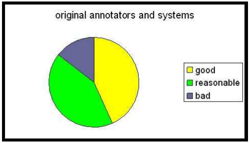




Post-Hoc Verdicts (separating substitutes)







Post-Hoc Analysis

- 52 examples where only humans provided the substitute and the post hoc annotators categorised this as 'bad'
- however many seem reasonable, for example
  Appointed by the CDFA, public members are chosen for their usefulness in helping the commodity **board** *carry out its purpose and to represent the public interest.*
  The annotation judged as "bad" was *management* which seemed reasonable to us.
- easier to make categorial judgment (bad, reasonable, good) compared to finding a substitute

Post-Hoc Analysis

of the 52 'bad' annotations only provided by humans:

- $\frac{50}{52}$ provided by only one annotator of the five
- $\frac{2}{52}$ substitutes provided by only two of the original annotators
- $\frac{38}{52}$ one of the three post hoc annotators was of a different opinion: (outlier gave 31 "reasonable" and 7 "good")
- $\frac{14}{52}$ all annotators disliked, however all of these cases only of original annotators provided this substitute

LEXSUB Conclusions

- lexical substitution task successful
  - no training data and no fixed inventory
  - 8 teams 10 systems
- participants used a range of man-made inventories
- most systems use web data for disambiguation
- system using explicit WSD module did best at 'disambiguation'
- lots of scope for unsupervised systems
- human substitutes are preferred by post-hoc annotators
- only a small percentage of system responses were good or reasonable and not found by original annotators

Post LEXSUB agenda

- look at word meaning overlap using synonym overlaps [Erk et al., 2009, McCarthy, 2011]
- examine if lexicographer decisions correlate with substitutions
- try contextual disambiguation with distributional inventories
- analyse multiword data [McCarthy, 2008]

Outline

Cross-Lingual Lexical Substitution (CLLS)

with Rada Mihalcea and Ravi Sinha, University of Noth Texas

- Find Spanish alternatives for an English target word in context
- Allows us to examine subtle relationships between usages
- Full fledged machine translation not required, just the target words

For twelve hours Livewire will be broadcasting live from the blue <u>bar</u> of Union House at UEA in an attempt to raise as much money as possible for a very worthy cause. [bar, cantina, taberna, caf].

Motivation

- Assist human translators
    - provide several translations the human could choose from
- Assist language learners
    - provide interpretation of difficult English words in their native language
- Help cross-lingual information retrieval
- Help automatic machine translation

Annotation

- Four native speakers of Spanish from Mexico, with high-level of proficiency in English
- Annotators were allowed to use any resource they wanted to, and provide as many substitutes as they could think of
- Similar to Lexical Substitution, except that the annotations are not synonyms but translations
- The annotators indicate whether the target word is part of a multiword and what that multiword is to clearly demarcate what the substitute is replacing

Annotation

- All words and contexts drawn from the English Lexical Substitution
- 30 words in the development set
- 100 words in the test set
- Each word had 10 contexts
- No limit to number of translations allowed

Interface

Start page/ Lexicon | Logout
LEXICON : BRIGHT.A
Page 1 of 10 (Has been Annotated!)
Next >>
The actual field is not much different than that of a 40mm , only it is smaller and quite a bit noticeably brighter , which is probably the main benefit .
ANNOTATION
Possible Translations (comma separated)
brillante, mejor,
Nil
Name
Used a dictionary
Target is part of the following phrase / multiword / idiom:
Comment
Save

Inter-Tagger Agreement (pairwise agreement)

- without mode 0.2777
- 0.2775 for English Lexical Substitution
- very comparable

$$pa = \sum_{i \in I} \frac{\sum_{\{h_i, h_i'\} \in P_i} \frac{h_i \cap h_i'}{h_i \cup h_i'}}{|P_i| \cdot |I|} \times 100 = 27.7\ (31.13)$$

$$pa_m = \frac{\sum_{i \in I_m} \sum_{h_i\ :\ h \in H} \frac{1\ \text{if}\ m_i \in h_i}{|h_i|}}{|HI_m|} \times 100 = 50.7\ (64.7)$$

Comparison with Cross-Lingual Word Sense Disambiguation (CLWSD)

- CLWSD is word sense disambiguation with sense inventory provided by humans using a parallel data resource
- CLLS does not assume clustering
- CLLS does not partition into senses
- usages share meaning yet not have identical translations
- CLWSD does though a translation can theoretically occur in more than one cluster, not yet seen how much this occurs

Translations and Senses

Senses might have some translations in common, but not all [Resnik and Yarowsky, 2000] (Table 4)
first two senses from WordNet for the noun *interest*:

WordNet sense	Spanish Translation
monetary e.g. on loan	*interés, rédito*
stake/share	*interés,participación*

Translations from one annotator for the adverb *severely*

1. Perhaps the effect of West Nile Virus is sufficient to extinguish endemic birds already **severely** stressed by habitat losses. {*fuertemente, severamente, duramente, exageradamente*}
2. She looked as **severely** as she could muster at Draco. {*rigurosamente, seriamente*}
3. A day before he was due to return to the United States Patton was **severely** injured in a road accident. {*seriamente, duramente, severamente*}
4. Use market tools to address environmental issues , such as eliminating subsidies for industries that **severely** harm the environment, like coal. {*peligrosamente, seriamente, severamente*}
5. This picture was **severely** damaged in the flood of 1913 and has rarely been seen until now. {*altamente, seriamente, exageradamente*}

Translations from the annotators for some of the sentences for the adjective *straight*

1. There is one question that demands an answer - a **straight** answer - from those who would seek to lead this nation and its people. {*directo 3;concreto 1;espontaneo 1;verdadero 1;exacto 1;inmediato 1;sin tapujos 1;preciso 1;real 1*}
2. This strong youth culture rapidly influenced other musical styles with its phrasing and break beats and gave birth to many contrasting styles including pop , funk , dance , techno , acid jazz , indie rock etc. A **straight** rap record is still hard-core and only relevant for a specific group and market , it does not have a commercial appeal. {*puro 3;directo 2;unico 1;simple 1;derecho 1;basico 1;sencillo 1*}
3. What is sure , but I don't believe anyone needs this warning , is that is most important to do things **straight**, fair and honest, and never think you can outsmart Scientology on your own. { *derecho 2;directo 1;recto 1;correcto 1;al punto 1;legal 1;al grano 1;claro 1;sencillo 1*}

Systems

- Nine teams / fifteen systems
- Resources used: bilingual dictionaries, parallel corpora (Europarl, or custom Wikipedia-built corpora), monolingual corpora (Web1T, newswire collections), translation systems (Moses, GIZA, Google)
- Some systems attempted the selection on the English side, some on the Spanish side

Baselines

- Generated from an online English-Spanish dictionary and the Spanish Wikipedia
  - First baseline: dictionary-based
    - DICT
    - Translations collected from the dictionary in the order returned by the online query page
  - Second baseline: dictionary and corpus-based
    - DICTCORP
    - Translations from the dictionary were ranked based on their frequencies in Wikipedia

Systems

System	resources	resource type	best rank	oot rank
WLVusp	Europarl; WordReference	parallel corpora; dictionary	4	6
USPwlv	Europarl	dictionary built from parallel corpora	2	8
SWAT-E	English and Spanish n-grams; Roget; NLTK's Lancaster stemmer; Google and SpanishDict dictionaries	dictionaries; n-grams	5	1
SWAT-S	Google and Yahoo translation; Spanish n-grams; Roget; TreeTagger; Google and Yahoo dictionaries	dictionaries; translation systems; n-grams	10	2

Systems . . .

System	resources	resource type	best rank	oot rank
ColEur	GIZA++; TreeTagger; SemCor; Europarl; WordNet	parallel corpora; lexicon; alignment tool	11	10
ColSlm	GIZA++; TreeTagger; SemCor; own created parallel corpus; WordNet	parallel corpora; lexicon; alignment tool	3	9
UBA-W	DBPedia; Google Dictionary; Babylon Dictionary; SpanishDict; Lucene; DBpedia extended abstracts for English and Spanish	dictionary; parallel corpora	8	5
UBA-T	Google Dictionary; Babylon Dictionary; SpanishDict; META; FreeLing	dictionary; translation tool	1	7

Systems . . .

UvT-v	Europarl; GIZA++; FreeLing	parallel corpora; alignment tool	6	3
UvT-g	Europarl; GIZA++; FreeLing	parallel corpora; alignment tool	9	4
FCC-LS	Europarl; GIZA++; WordNet	parallel corpora; alignment tool	N/A	13
CU-SMT	Europarl	parallel corpora	7	N/A
TYO	WordNet; Penn Treebank; BLIP; FreeDict; Google Dictionary; Spanish word frequency list	dictionary (lexicon); corpus	14	11
IRST-1	Moses; EuroParl; WordReference; TreeTagger; LSA built on Spanish Google News	parallel corpora; alignment tool; dictionary; LSA	12	12
IRSTbs	Moses; EuroParl	parallel corpora	13	14

best results

Systems	*R*	*P*	*Mode R*	*Mode P*
UBA-T	27.15	27.15	57.20	57.20
USPWLV	26.81	26.81	58.85	58.85
ColSlm	25.99	27.59	56.24	59.16
WLVUSP	25.27	25.27	52.81	52.81
SWAT-E	21.46	21.46	43.21	43.21
UvT-v	21.09	21.09	43.76	43.76
CU-SMT	20.56	21.62	44.58	45.01
UBA-W	19.68	19.68	39.09	39.09
UvT-g	19.59	19.59	41.02	41.02
SWAT-S	18.87	18.87	36.63	36.63
ColEur	18.15	19.47	37.72	40.03
IRST-1	15.38	22.16	33.47	45.95
IRSTbs	13.21	22.51	28.26	45.27
TYO	8.39	8.62	14.95	15.31
DICT	24.34	24.34	50.34	50.34
DICTCORP	15.09	15.09	29.22	29.22

oot results

Systems	R	P	$Mode\ R$	$Mode\ P$	dups
SWAT-E	174.59	174.59	66.94	66.94	968
SWAT-S	97.98	97.98	79.01	79.01	872
UvT-v	58.91	58.91	62.96	62.96	345
UvT-g	55.29	55.29	73.94	73.94	146
UBA-W	52.75	52.75	83.54	83.54	-
WLVUSP	48.48	48.48	77.91	77.91	64
UBA-T	47.99	47.99	81.07	81.07	-
USPWLV	47.60	47.60	79.84	79.84	30
ColSlm	43.91	46.61	65.98	69.41	509
ColEur	41.72	44.77	67.35	71.47	125
TYO	34.54	35.46	58.02	59.16	-
IRST-1	31.48	33.14	55.42	58.30	-
FCC-LS	23.90	23.90	31.96	31.96	308
IRSTbs	8.33	29.74	19.89	64.44	-
DICT	44.04	44.04	73.53	73.53	30
DICTCORP	42.65	42.65	71.60	71.60	-

Upper Bounds

$$best\ upper\ bound = \frac{\sum_{i \in I} \frac{freq_{most\ freq\ substitute_i}}{|T_i|}}{|I|} \times 100 = 40.57$$

405.78 is **oot** upper bound

Minor Issues with Encodings

- Some participants did not clean their files of incoherent character encodings our result files indicated 4-5 different character encodings
- Some of these encodings included diacritics and malformed characters, despite instructions: no diacritics
- We performed some basic cleaning cleaned out diacritics but left the malformed characters since they would have taken a significant amount of manual effort
- These malformed characters caused some systems to lose some points

Scores

- remember, score for each item depends on consensus from annotators. This allows items with greater consensus to have more weight.
- An item with perfect consensus will have an upper bound of 1
- Allowing duplicates means that the out-of-ten precision and recall scores can exceed a value of 100
- Duplicates do not influence the mode scores
- The column Dups shows the number of items for which at least one duplicate was provided
- Most systems did not provide duplicates

System performance (normalised) by PoS

- analyse **best** results by PoS
- normalise scores by upper bound for each item (to make comparisons across PoS possible)
- macro precision (number correct / attempted) and recall (number correct / total)

best results: Nouns

sys	att	recall	precision
UBA-T	300	67	67
ColSlm	298	55	56
SWAT-S	300	54	54
WLVusp	300	54	54
uspWLV	300	52	52
CU-SMT	294	51	52
DICT	300	50	50
SWAT-E	300	49	49
UvT-v	300	47	47
DICTCORP	300	42	42
UvT-g	300	42	42
UBA-W	300	41	41
IRST-1	246	36	43
ColEur	298	33	34
IRSTbs	229	33	43
TYO	290	15	15

best results: Verbs

sys	att	recall	precision
uspWLV	310	61	61
ColSlm	301	55	57
UBA-T	310	54	54
WLVusp	310	50	50
SWAT-E	310	48	48
DICT	310	46	46
UvT-v	310	42	42
ColEur	301	40	42
DICTCORP	310	40	40
UBA-W	310	40	40
UvT-g	310	40	40
CU-SMT	292	36	38
SWAT-S	310	36	36
IRST-1	179	21	36
IRSTbs	153	16	33
TYO	307	12	12

best results: Adjectives

sys	att	recall	precision
uspWLV	280	80	80
WLVusp	280	76	76
UBA-T	280	74	74
ColSlm	264	73	77
DICT	280	72	72
UBA-W	280	66	66
SWAT-E	280	59	59
UvT-v	280	59	59
UvT-g	280	58	58
ColEur	254	55	61
CU-SMT	269	51	53
IRST-1	196	48	69
SWAT-S	280	48	48
IRSTbs	165	40	68
DICTCORP	280	39	39
TYO	278	26	26

best results: Adverbs

sys	att	recall	precision
DICT	110	54	54
uspWLV	110	54	54
WLVusp	110	52	52
ColSlm	79	47	66
SWAT-E	110	37	37
UBA-T	110	36	36
UvT-v	110	34	34
CU-SMT	96	32	37
TYO	99	32	35
UvT-g	110	32	32
ColEur	79	29	40
IRST-1	73	28	42
SWAT-S	110	27	27
UBA-W	110	23	23
IRSTbs	40	22	62
DICTCORP	110	12	12

Results

- Results are higher than those for English Lexical Substitution
- Are translations easier than paraphrases?
- Is it because there are parallel corpora available for different languages but not for paraphrases?

System Correlations (I) (Spearman's ρ)

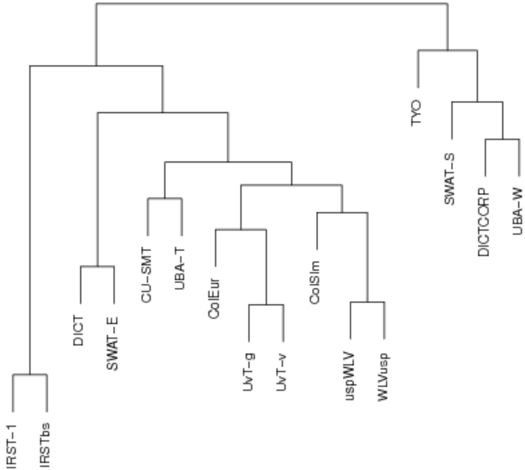
	ColEur	ColSlm	CU-SMT	DICT	DICTCORP	IRST-1	IRSTbs	SWAT-E
ColEur	1	0.4	0.39	0.29	0.28	0.43	0.41	0.3
ColSlm	0.4	1	0.36	0.48	0.25	0.34	0.27	0.45
CU-SMT	0.39	0.36	1	0.25	0.16	0.48	0.43	0.27
DICT	0.29	0.48	0.25	1	0.3	0.3	0.22	0.56
DICTCORP	0.28	0.25	0.16	0.3	1	0.12	0.13	0.3
IRST-1	0.43	0.34	0.48	0.3	0.12	1	0.88	0.32
IRSTbs	0.41	0.27	0.43	0.22	0.13	0.88	1	0.24
SWAT-E	0.3	0.45	0.27	0.56	0.3	0.32	0.24	1
SWAT-S	0.24	0.23	0.34	0.2	0.18	0.3	0.26	0.24
TYO	0.27	0.2	0.18	0.18	0.09	0.2	0.21	0.18
UBA-T	0.36	0.42	0.43	0.4	0.24	0.31	0.29	0.37
UBA-W	0.38	0.34	0.21	0.24	0.26	0.19	0.2	0.21
uspWLV	0.44	0.59	0.43	0.45	0.26	0.39	0.33	0.43
UvT-g	0.6	0.48	0.46	0.33	0.23	0.42	0.36	0.34
UvT-v	0.49	0.45	0.47	0.3	0.18	0.43	0.38	0.38
WLVusp	0.44	0.43	0.39	0.42	0.23	0.37	0.33	0.35

System Correlations (II) (Spearman's ρ)

	SWAT-S	TYO	UBA-T	UBA-W	uspWLV	UvT-g	UvT-v	WLVusp
ColEur	0.24	0.27	0.36	0.38	0.44	0.6	0.49	0.44
ColSlm	0.23	0.2	0.42	0.34	0.59	0.48	0.45	0.43
CU-SMT	0.34	0.18	0.43	0.21	0.43	0.46	0.47	0.39
DICT	0.2	0.18	0.4	0.24	0.45	0.33	0.3	0.42
DICTCORP	0.18	0.09	0.24	0.26	0.26	0.23	0.18	0.23
IRST-1	0.3	0.2	0.31	0.19	0.39	0.42	0.43	0.37
IRSTbs	0.26	0.21	0.29	0.2	0.33	0.36	0.38	0.33
SWAT-E	0.24	0.18	0.37	0.21	0.43	0.34	0.38	0.35
SWAT-S	1	0.15	0.33	0.19	0.25	0.33	0.32	0.3
TYO	0.15	1	0.1	0.06	0.18	0.21	0.21	0.17
UBA-T	0.33	0.1	1	0.35	0.42	0.42	0.44	0.39
UBA-W	0.19	0.06	0.35	1	0.36	0.29	0.27	0.35
uspWLV	0.25	0.18	0.42	0.36	1	0.54	0.53	0.67
UvT-g	0.33	0.21	0.42	0.29	0.54	1	0.66	0.5
UvT-v	0.32	0.21	0.44	0.27	0.53	0.66	1	0.49
WLVusp	0.3	0.17	0.39	0.35	0.67	0.5	0.49	1

Dendrogram from Clustering the Correlation Matrix

converted with Euclidean distance (following the method described on page 140 of [Baayen, 2008])



Disruptive Set Analysis

due to Ravi Sinha

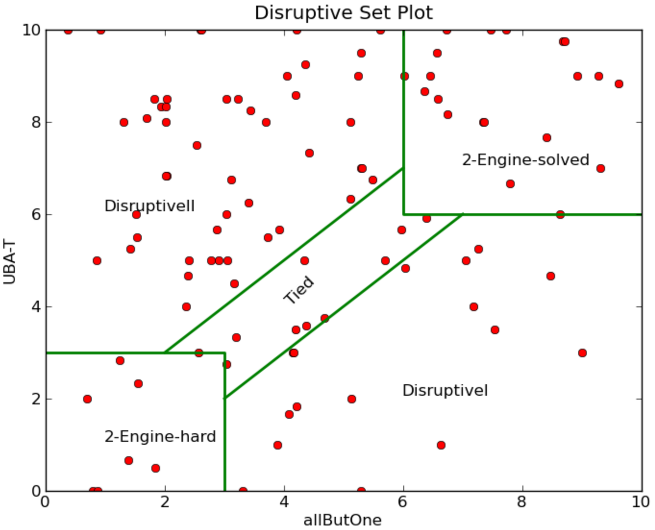
- graphical visualisation of comparison between two systems
- originally designed for search engine comparison
- threshold of *solving*
- The *disruptive set* of a system is defined as the set of queries that that particular system can solve and the other one cannot
- for each system divide instances *solved* ($> thresh1$) and *hard* ($< thresh2$)
- relevant intersections give *two-system-solved* and *two-system-hard*
- find Disruptive I and Disruptive II ...

. . . Disruptive Set Analysis

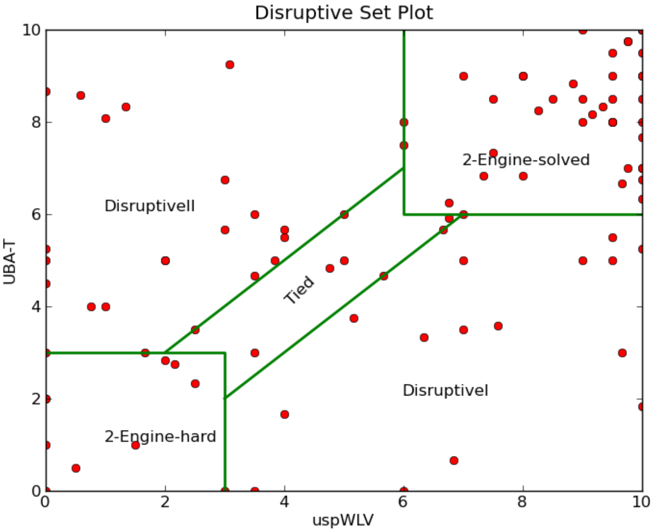
due to Ravi Sinha

- we used lemmas as datapoints (could use instances or PoS or any other grouping)
- tied region: diff in scores of the two systems < *tied threshold*
- Normalised scores 0-10
- $\delta_{solved} = 6$, $\delta_{hard} = 3$, $\delta_{tied} = 2$

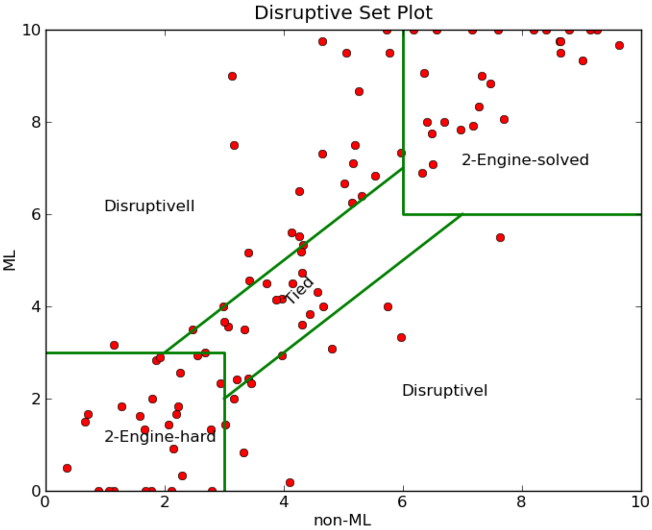
average vs UBA-T (best)



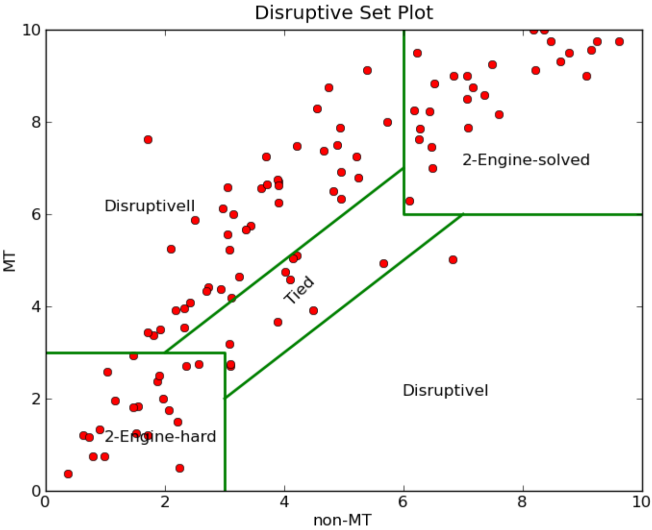
USPWLV (second) vs UBA-T



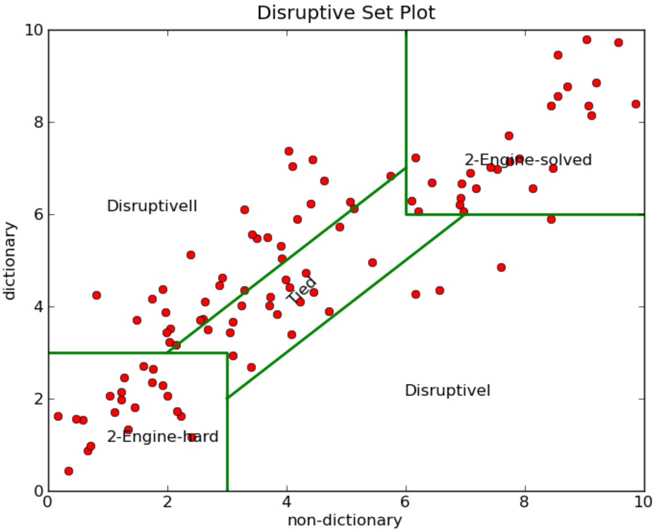
Non ML vs ML


Disruptive Set Plot
ML
non-ML
0
2
4
6
8
10
DisruptiveII
2-Engine-solved
Tied
DisruptiveI
2-Engine-hard

Non MT vs explicit MT



Non dictionary vs explicit dictionary



Lemmas solvable only by one systems

Lemma	Systems that solve those
range.n	ColSlm
closely.r	DICT
shade.n	CU-SMT
check.v	uspWLV
bug.n	DICT
ring.n	UBA-T
charge.v	UBA-T
pot.n	UBA-T
hold.v	DICTCORP

Lemmas solvable only by a few systems

Lemma	Systems that solve those
fire.v	WLVusp, UBA-T
burst.v	SWAT-E, UBA-T
return.v	UvT-v, UBA-W
figure.n	DICTCORP, ColSlm
extended.a	SWAT-S, DICTCORP, DICT
heavy.a	DICT, WLVusp, UBA-W
only.r	ColSlm, DICT, SWAT-E
way.n	UvT-g, ColEur, UBA-W
tender.a	DICT, UBA-T, UBA-W
around.r	SWAT-S, WLVusp, UBA-W
shot.n	UvT-g, uspWLV, CU-SMT
stiff.a	uspWLV, WLVusp, CU-SMT

Meta system

due to Rada Mihalcea

- ranks each translation according to credit from all systems
- outputs top (**best**) or top 10 (**oot**)

$$credit(c) = \sum_{k \in K} \frac{1}{|S_i^k|} * (c \in S_i^k ? 1 : 0)$$

where S_i^k is the set of answers submitted by system S^k for item i

- tie breaks are arbitrary

Meta system: Results

Evaluation	*R*	*P*	*Mode R*	*Mode P*
best	28.08	28.08	60.63	60.63
best system	27.15	27.15	57.20	57.20
oot	56.22	56.22	88.89	88.89

Further Work

- more analysis of MW data [McCarthy, 2008]
- more analysis to see which approach works well WHEN i.e. which are the factors of a context that can predict the right approach to use
- do translations and paraphrases cluster?
- comparison of CLLS data and Cross-Lingual Word Sense Disambiguation task dataset
- further cross lingual lexical substitution task planned for next SemEval

Credits

Thank you

Credits

Thank you

and thanks also to . . .
Collaboration with Roberto Navigli
and Rada Mihalcea, Ravi Sinha

Credits

Thank you

and thanks also to ...
Collaboration with Roberto Navigli
and Rada Mihalcea, Ravi Sinha

- LEXSUB task web site:
  http://www.dianamccarthy.co.uk/task10index.html
- CLLS web site:
  http://lit.csci.unt.edu/index.php/Semeval_2010

Baayen, H. (2008).
Analyzing Linguistic Data: A Practical Introduction to Statistics Using R.
Cambridge University Press.

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.

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.

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.

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

McCarthy, D. (2008).
Lexical substitution as a framework for multiword evaluation.

In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco.

McCarthy, D. (2011).
Measuring similarity of word meaning in context with lexical substitutes and translations.
In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing, CICLing 2011, Pt. I (Lecture Notes in Computer Science, LNTCS 6608)*. Springer.

Resnik, P. and Yarowsky, D. (2000).
Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation.
Natural Language Engineering, 5(3):113–133.

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. (1998).
Automatic word sense discrimination.
Computational Linguistics, 24(1):97–123.

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, 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.