Word Sense Disambiguation

Diana McCarthy

Lexical Computing Ltd.

University of Melbourne, July 2011

Outline

Word Sense Disambiguation

Getting computers to find the correct meaning of a word in context

e.g.

What sort of plants thrive in chalky soil?

Word Sense Disambiguation

Getting computers to find the correct meaning of a word in context e.g.

What sort of plants thrive in chalky soil?

plant#n#1 ?

plant#n#2 ?

Outline

WSD Approaches

- supervised (hand labelled data)
- knowledge-based (dictionaries, thesauruses)
- unsupervised
  - induce senses (fully unsupervised) similarity of input vector to previous clusters (LSA)
  - or associate distributional information with entries in given sense inventory NB association uses knowledge

Knowledge-Based WSD

Using information from manually created lexical resources

- dictionary definitions [Lesk, 1986]
- semantic relations [Navigli and Velardi, 2005] conceptual density [Agirre and Rigau, 1996] graphical methods [Sinha and Mihalcea, 2007],
- wikipedia (with WordNet) [Ponzetto and Navigli, 2010]

Lesk

definitions e.g.

pine	1.	evergreen tree with needle-shaped leaves
cone	1.	solid body which narrows to point
	2.	fruit of certain evergreen trees (fir, pine)

The pine bore cones that seemed to bend... w_1 = pine w_2 = cone

1. for each sense i of w_1
2. for each sense j of w_2
3. $argmax(overlap(i,j))$ where overlap is number of words in definitions of both i and j

Dante experiments

Initial experiments using:

- collocates : match with context e.g *You can get a wireless mouse if you . . .*
- SCF match with context: particularly promising for verbs *the gun fired*
- definitions : overlap with definitions of words in context
- domain: overlap with domain of words in context

Collocates e.g. *mouse*

mouse: (PoS: n)
meaning: a small long-tailed rodent
domain: zoo
example: *The* **mouse** *was dead in his cage the following day*
...
SCF: N_PREMOD
COLLOC: droppings nest hole cage
example: *Look for signs of* **mouse droppings** *etc.*
example: **Mouse** cages are available in various stages, sizes and designs.
...
SCF: N_MOD
COLLOC: laboratory house wood field harvest pet
example: *A set of 50 laboratory* **mice** *were examined at monthly intervals for 2 years from birth*
...

Collocates e.g. *mouse*

mouse: (PoS: n)
meaning: a computer input device controlled with one hand which moves the cursor on the computer screen
domain: IT
example: If your mouse runs off the mat edge, lift the **mouse** *up, move it back to the mat middle, and put it down*
...
COLLOC: optical, wireless cordless
...

SCF e.g. *fire to discharge a weapon*

```
Frame1
------
    '_O' (intransitive)
Frame2
------
    'NP'
     collocations: 'shot', 'round', 'gun', 'weapon', 'rifl
        'rocket', 'missile', 'shell', 'arrow'
Frame3
------
    'PP_X'
Frame4
------
    'NP PP_X'
```

Sample senses containing domain information

```
SenseID          Domain
-------          ------
mouse#1          ['zool']
mouse#3          ['IT']
soap#1           ['cosm']
soap#2           ['TV-rad']
soar#1           ['mus']
soar#2           ['bird']
```

Definitions e.g. investigation

```
sense               Sense Definition
-----               ---------------------
investigation#1     'a formal enquiry'
investigation#2     'research or detailed study

Sense                List of salient words in definition
----                -----------------------------------
investigation#1      ['formal', 'enquiry']
investigation#2      ['research', 'detailed', 'study']
```

Supervised WSD

- most prolific approach due to higher precision
- requires hand-labelled data, and lots of it
- typically lexical sample otherwise data is insufficient ([Ng, 1997] uses 100 minimum)
- typically determining optimum features best done on a word by word basis [Véronique et al., 2002]
- hard to be sure of any approach being globally best because of interaction of parameters
- binary vs n-ary models

Representation of example by features

- local features (with position) capture collocations and limited syntactic information:
  - PoS tags
  - lemmas
  - word forms
- topical features, wider windows or lexical info in extended context, capture semantic domain
- dependencies at a sentence level, better argument head relations

Algorithms

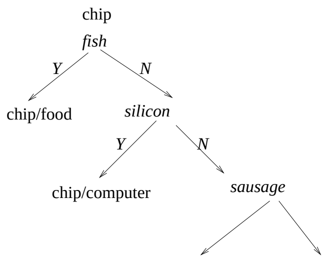
- decision list [Yarowsky, 1994]
  - {feature, value, class}
  - training data used to determine importance of rules (e.g. log likelihood)

$$\log\left(\frac{p(sense_a|collocation_i)}{p(sense_b|collocation_i)}\right)$$

  - rules ordered
  - first matching is applied

fish	7.2	**food**
silicon	5.2	**computer**
sausage	4.3	**food**

Algorithms



- decision trees e.g. C4.5
  - recursive partitioning
  - features have too many values
  - computationally expensive, not reliable
  - terminals with few examples

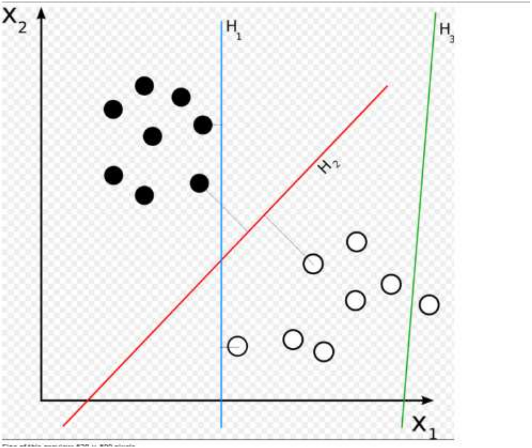
Algorithms contd. . .

- probabilistic
  - Naive Bayes
  - maximum entropy
- similarity
  - vector space mode; prototypes
  - kNN (memory, instance, exemplar based, case-based)

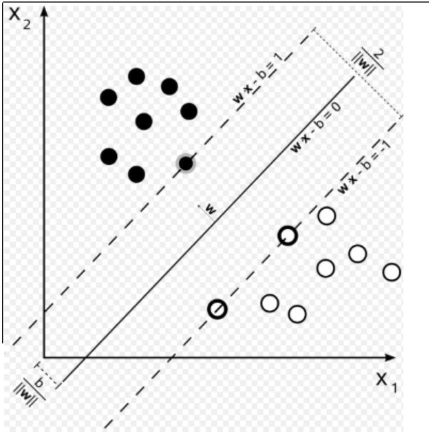
Algorithms contd. . .

- rule combination, (ensemble methods) e.g. majority voting, Adaboost combines weaker classifiers
- linear (binary) classifier:
  - hyperplane in n-dimensional space, weight vector
  - learn non-linear transformation to higher dimensional space via kernel function (boundaries may be easier to spot in high dimensional space)
  - SVM good example (and very good results), better with less training data compared to adaboost, which is better with more

SVMs



SVMs



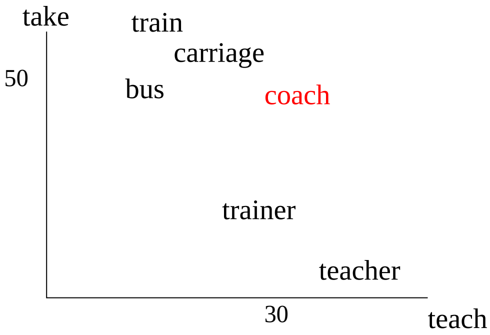
“Unsupervised”

- NB that many systems described as unsupervised are indeed knowledge based
- some ([McCarthy et al., 2004]) use info from the inventory for mapping the corpus data to the gold standard
- others use some level of explicit knowledge [Yarowsky, 1995]
- many many systems calling themselves unsupervised use hand tagged data (SemCor)

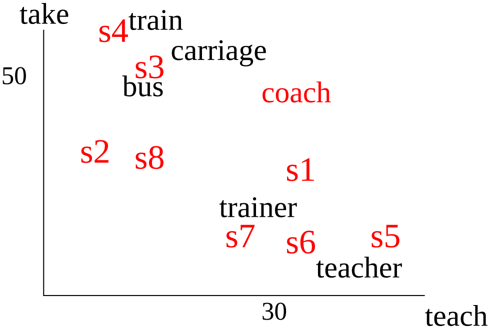
Unsupervised [Schütze, 1992, Schütze, 1998]

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

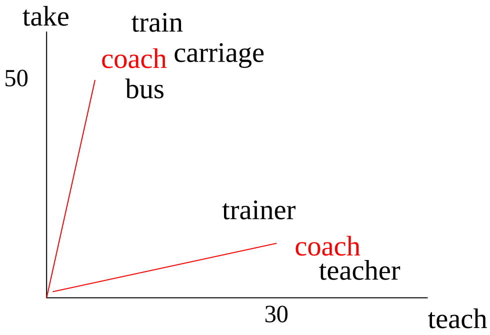
Vector Based Approaches



Vector Based Approaches



Vector Based Approaches



Similarity between two words: cosine

$$sim(a, b) = \frac{a.b}{|a||b|} = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} b_i^2} \sqrt{\sum_{i=1}^{n} b_i^2}}$$

Context Group Discrimination [Schütze, 1998, Schütze, 1992]

- SVD to reduce dimensionality
- Agglomorative clustering as seeds for EM (Buckshot)
- clusters 2 vs 10 (predetermined)
- evaluation of separating senses
- evaluation of disambiguation: pseudo-disambiguation, Information retrieval
- information retrieval (filtering matches) 7.4% percent better than word-based (combined 14.4%)

Bootstrapping

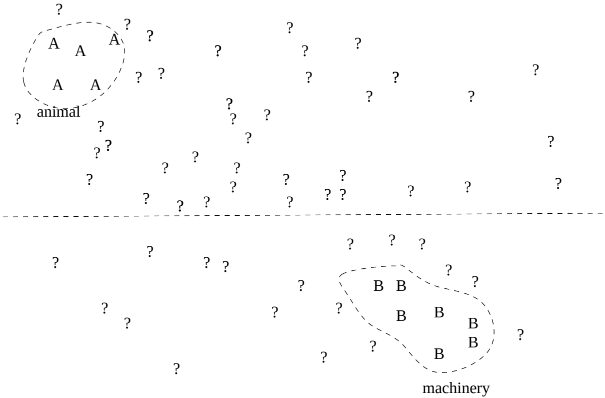
- self-training [Yarowsky, 1995]
  - seed data
  - iterate
- co-training
  - iterate between two classifiers
  - different views on the data

Unsupervised word sense disambiguation rivaling supervised methods

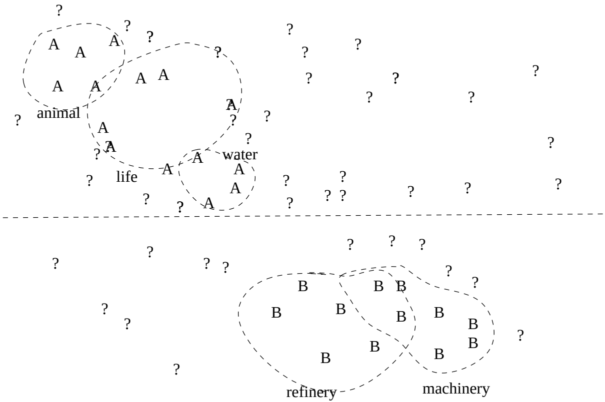
[Yarowsky, 1995]

- start with seeds e.g. *plant* (**animal** vs **machinery**)
- tag the data using these seeds (1% each) (rest is residual)
- train supervised classifier (decision list)
- apply, with threshold on probability and add new examples to seed sets
- optionally apply one sense per discourse hypothesis (extend seeds, or change classification, or remove to residual)
- stop when residual is stable

Yarowsky Algorithm



Yarowsky Algorithm



Yarowsky Algorithm

- seeds from experts or
- can escape from initial misclassifications, but to help:
  - increase context window after intermediate convergence
  - randomly change threshold

Yarowsky 1995 Results

Word	senses	SSize	%FSH	Sup	Y2w	Ydic	colls
plant	living/factory	7538	53.1	97.7	97.1	97.3	97.6
space	volume/outer	5745	50.7	93.9	89.1	92.3	93.5
tank	vehicle/container	11420	58.2	97.1	94.2	94.6	95.8
motion	legal/physical	11968	57.5	98.0	93.5	97.4	97.4
bass	fish/music	1859	56.1	97.8	96.6	97.2	97.7
palm	tree/hand	1572	74.9	96.5	93.9	94.7	95.8
poach	steal/boil	585	84.6	97.1	96.6	97.2	97.7
axes	grid/tools	1344	71.8	95.5	94.0	94.3	94.7
duty	tax/obligation	1280	50.0	93.7	90.4	92.1	93.2
drug	medicine/narcotic	1380	50.0	93.0	90.4	91.4	92.6
sake	benefit/drink	407	82.8	96.3	59.6	95.8	96.1
crane	bird/machine	2145	78.0	96.6	92.3	93.6	94.2
AVG		3936	63.9	96.1	90.6	94.8	95.5

Semi-automatic Dictionary Drafting: SADD [Kilgarriff and Rychlý, 2010]

- Yarowsky like algorithm
- senses as clusters of instances
- one sense per collocate
- clusters of collocates

Demo (or pictures)

Sketch Engine: Clusters of Collocates

Demo (or pictures)

Sketch Engine: Clusters of Collocates

object	58612	4.0
disorder	2377	9.01
meat 1792	3808	8.84
bean 101 beef 294 carrot 117 chicken 191 egg 359 lamb 67 mushroom 68 pork 115 potato 105 rice 150 turkey 73 vegetable 376		
meal 1795	9396	8.33
breakfast 895 dinner 630 food 5021 lunch 1055		
diet 1388	2400	8.19
habit 1012		
fruit 1290	2226	8.0
apple 342 banana 159 cereal 70 grain 90 nut 96 seed 179		
bread 692	3283	7.67
biscuit 166 cake 403 cheese 206 chocolate 421 cream 157 pasta 62 pie 251 salad 105 sandwich 372		

SADD initialisation

Home | **Concordance** | Word List | **Word Sketch** | Thesaurus | Sketch-Diff

charge

205 **object:** fee, sum
Label: money

174 **pp_with-p:** murder
Label: crime

138 **pp_with-p:** offence, crime
Label: crime

91 **modifier:** yesterday
Label:

73 **modifier:** highly
Label:

59 **and/or:** arrest
Label: crime

44 **modifier:** positively
Label: electric

40 **pp_to-p:** profit
Label: money

SADD annotating

Home | **Concordance** | Word List | **Word Sketch** | Thesaurus | Sketch-Diff | **View**

Annotating: **charge-v**

Not assigned 6045 P / N

modifier: negatively, electrically, emotionally, *pp_with-p:* conspiracy, assault, o

word sketch

---

crime 416 P / N

pp_with-p: murder, offence, theft, *and/or:* arrest, *subject:* magistrate, *modifie*

word sketch

---

electric 44 P / N

modifier: positively, *object:* nucleus, proton, atom, word sketch

---

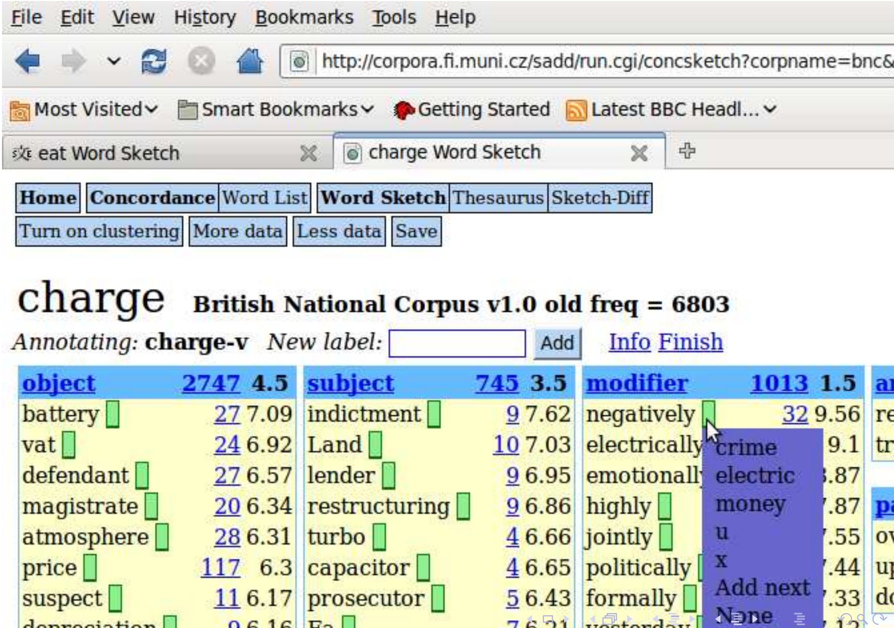
money 298 P / N

object: fee, sum, call, *subject:* Girobank, *pp_to-p:* profit, student, *pp_at-p:* rate

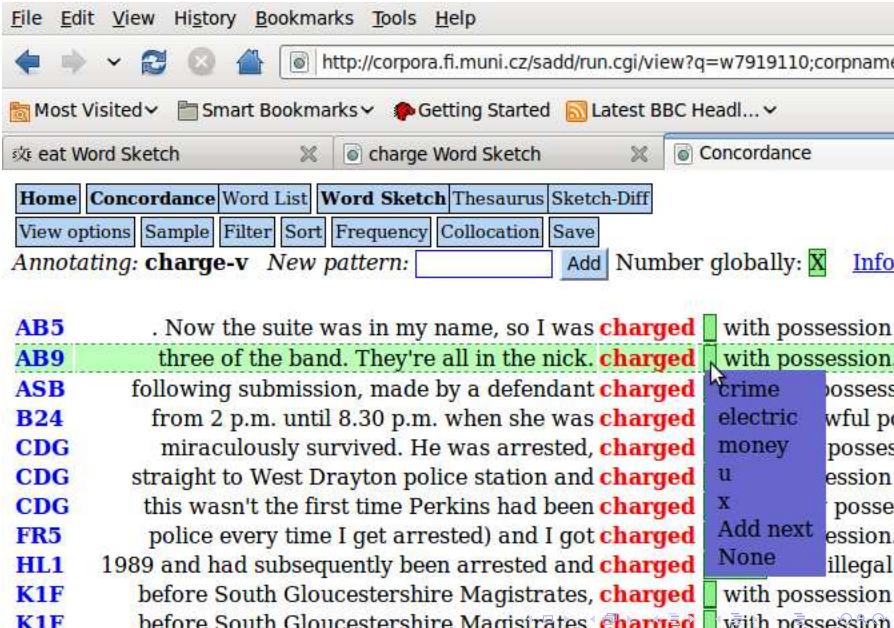
---

Minscore: 0.5 Mindiff: 0.8 Automatically annotate

SADD annotating word sketch



SADD annotating at the concordance



Cross Lingual Word Sense Disambiguation

[Resnik and Yarowsky, 2000, Lefever and Hoste, 2010, Diab and Resnik, 2002, Chan and Ng, 2005]

1. *bank* $\leftrightarrow$ *dijk* or *oever* (Dutch)
   giving fish to people living on the bank of the river
2. *bank* $\leftrightarrow$ *bank* or *kredietinstelling* (Dutch)
   The bank of Scotland ...

Cross Lingual Word Sense Disambiguation

Language	Sense label
	The bank of Scotland
Dutch	oever/dijk
French	rives/rivage/bord/bords
German	Ufer
Italian	riva
Spanish	orilla
	The bank of Scotland
Dutch	bank/kredietinstelling
French	banque/établissement de crèdit
German	Bank/Kreditinstitut
Italian	banca
Spanish	banco

Cross Lingual Word Sense Disambiguation

- unsupervised BUT corpus based, relies on aligned corpora
- uses word alignment tools (GIZA++) to provide inventory and training data
- best way to go if you have cross lingual application and know your source and target
- if the goal is translation into several languages eventually every distinction that can be made will be made [Palmer et al., 2007]

Outline

Evaluation

- in vitro (stand alone) in vivo (within an application)
- prior to senseval
  - small samples of words [Leacock et al., 1993, Yarowsky, 1995] [Yarowsky, 1995]
  - or different subsets [Wilks and Stevenson, 1998]

but what about:

but what about:

- the inventory?
- all words vs lexical selection?
- lexical selection?
- data selection?
- amount of context?
- training data vs testing data?
- scoring?

Baselines

- Random: fairest baseline for unsupervised system

$$\sum_{i \in instances} \frac{1}{senses(i)}$$

- First sense
- Most frequent sense
- Upper bound (pairwise inter-tagger agreement)

Pseudo-Words

- merge two words to create an artificial test set *banana-shell*
- which word is correct for the context
- similar to word1 word2 confounder test sets for structural / collocational disambiguation e.g. PP attachment
- issues (see for example [Stokoe, 2005])
  - frequency of words
  - frequency bounds (between 500 and 1000 [Schütze, 1998])
  - ambiguity of words (word pairs [Schütze, 1998])
  - closeness in meaning of words and therefore ease of disambiguation

Senseval

- first senseval organised in 1998 at Hertmonceaux, UK
- arising from discussions preceding year: SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What and How?
- level playing field, same time constraints
- same words, same test instances, same measures
- sampling and inventory?
- English, Italian and French lexical samples (25 systems)
- English Inventory: Hector (OUP and DEC project) with WordNet mapping
- scoring allowed a degree of confidence

Senseval-2

- 2001 Toulouse France
- all words as well as lexical sample
- 12 different languages (93 systems)
- Basque, Chinese, Czech, Danish, Dutch, English, Estonian, Italian, Japanese, Korean, Spanish, Swedish.
- Japanese translation task as well as lexical sample
- coarse grained mapping for English Lexical Sample Senseval-2

Senseval-3

- 2004 Barcelona
- 14 tasks (160 systems)
- eight languages WSD (all words and lexical sample - ceiling at 73%)
- SRL
- WSD for SCF acquisition
- gloss disambiguation
- logic forms (transform english sentences to first order logic notation) some students like to study in the mornings. $\rightarrow student : n_{(x1)} like : v_{(e4,x1,e5)} to(e4, e5) study : v_{(e5,x1,x2)} in(e5, x2) morning : n_{(x2)}$.

SemEval

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

- Workshop at ACL 2007 Prague, Czech Republic
- 18 tasks including:

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

SemEval-2

see http://semeval2.fbk.eu/semeval2.php

- Workshop at ACL 2010, Uppsala Sweden
- 18 tasks including:

- Cross-lingual WSD
- Co-reference resolution
- VP ellipsis - detection and resolution
- Automatic Keyphrase Extraction from Scientific Articles
- Argument selection and coercion
- Event Detection in Chinese News Sentences
- Parser Training and Evaluation using Textual Entailment
- Tempeval-2

Plans afoot for SemEval-3: Why engage?

- what you can gain from participating?
- don't worry about being bottom
- not necessarily good to focus on coming top
- don't forget the science!!!
- what you can gain from co-organising?
- wonderful opportunity to explore new ideas
- use for learning and experience
- be careful of fools gold

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

Outline

The First Sense Heuristic

Simple but powerful. For example WordNet (v3.0) noun *plant*:

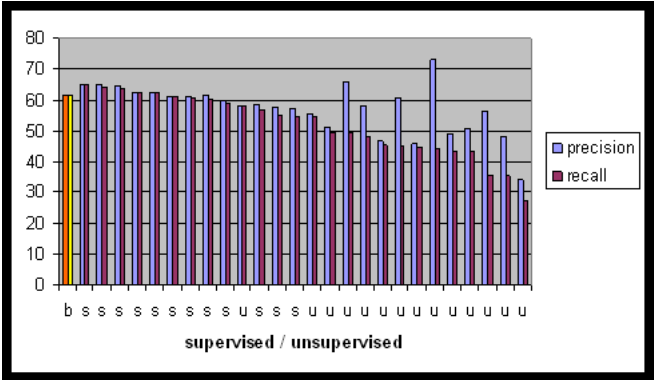
1. (63) plant, works, industrial plant – (buildings for carrying on industrial labor; "they built a large plant to manufacture automobiles")
2. (37) plant, flora, plant life – ((botany) a living organism lacking the power of locomotion)
3. plant – (an actor situated in the audience whose acting is rehearsed but seems spontaneous to the audience)
4. plant – (something planted secretly for discovery by another; "the police used a plant to trick the thieves"; "he claimed that the evidence against him was a plant")

The First Sense Heuristic

- obtained from manually labelled data or lexicographer intuition
- many WSD systems use (even those that profess to be unsupervised)
- systems use it when there is no evidence from the context (more often than you would expect)
- BUT there is a shortage of hand-tagged text
- AND the first sense of a word changes with domain

WSD Lessons

best systems performing just better than first sense heuristic over all words e.g. English all words SENSEVAL-3



First Sense Heuristic from SemCor is not always reliable e.g. *pipe* (noun)

1. (6) pipe, tobacco pipe – (a tube with a small bowl at one end; used for smoking tobacco)
2. (4) pipe, pipage, piping – (a long tube made of metal or plastic that is used to carry water or oil or gas etc.)
3. pipe, tube – (a hollow cylindrical shape)
4. pipe – (a tubular wind instrument)
5. organ pipe, pipe, pipework – (the flues and stops on a pipe organ)

First Sense Heuristic from SemCor is not always reliable e.g. *pipe* (noun)

1. (6) pipe, tobacco pipe – (a tube with a small bowl at one end; used for smoking tobacco)
2. (4) pipe, pipage, piping – (a long tube made of metal or plastic that is used to carry water or oil or gas etc.)
3. pipe, tube – (a hollow cylindrical shape)
4. pipe – (a tubular wind instrument)
5. organ pipe, pipe, pipework – (the flues and stops on a pipe organ)

Distributional neighbours of *pipe* from the British National Corpus (BNC) : tube (0.139) cable (0.137) wire (0.131) tank (0.131) hole (0.120) cylinder (0.116) . . .

Method [McCarthy et al., 2004]

Distributional neighbours of *pipe* from BNC:
tube (0.139) cable (0.137) wire (0.131) tank (0.131) hole (0.120) cylinder (0.116) . . .

Method [McCarthy et al., 2004]

Distributional neighbours of *pipe* from BNC:
tube (0.139) cable (0.137) wire (0.131) tank (0.131) hole (0.120) cylinder (0.116) . . .

- Use number and score (*ds*) of distributional neighbours pertaining to each sense
- Tie distributional neighbours to senses (*ss*). We use WordNet Similarity, 2 useful measures:
  - lesk [Lesk, 1986]: definition overlap,
  - jcn [Jiang and Conrath, 1997]: uses frequency counts from corpus and hypernym hierarchy

Our Sense Ranking Score

$$Prevalence\ Score(w, s_i) = \sum_{n_j \in N_w} ds(w, n_j) \times \frac{ss(s_i, n_j)}{\sum_{s_{i'} \in senses(w)} ss(s_{i'}, n_j)}$$

plant: senses	Neighbours tree 0.17	flower 0.16	factory 0.14 ...
flora	$0.17 \times \frac{ss(flora, tree)}{\sum ss(*, tree)}$	$0.16 \times \frac{ss(flora, flower)}{\sum ss(*, flower)}$	$0.14 \times \frac{ss(flora, factory)}{\sum ss(*, factory)}$
works	$0.17 \times \frac{ss(works, tree)}{\sum ss(*, tree)}$	$0.16 \times \frac{ss(works, flower)}{\sum ss(*, flower)}$	$0.14 \times \frac{ss(factory, works)}{\sum ss(*, factory)}$

Experimental Set Up

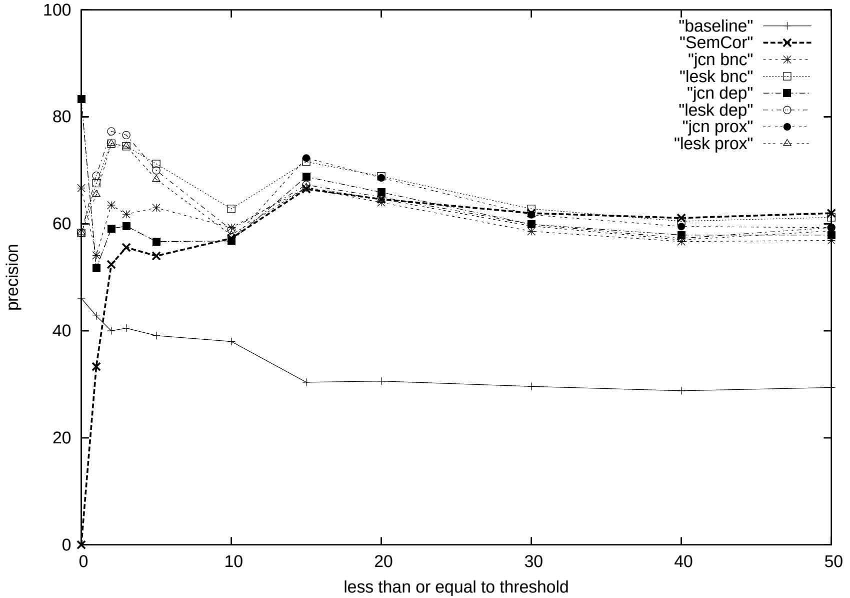
Distributional thesaurus:

- BNC [Leech, 1992]
- RASP parser [Briscoe and Carroll, 2002]

PoS	Grammatical contexts
noun	verb in object or subject relation, adj or noun modifier
verb	noun as object or subject
adjective	modified noun, modifying adverb
adverb	modified adj or verb

- Lin's newswire thesaurus: proximity and dependency

SENSEVAL-2 WSD Precision with SemCor Frequency



Automatic Detection of Entropy

with Peng Jin

$$Prevalence\ Score(ws_i)$$

$$= \sum_{n_j \in N_w} ds(w, n_j) \times \frac{wnss(ws_i, n_j)}{\sum_{ws_{i'} \in senses(w)} wnss(ws_{i'}, n_j)}$$

Automatic Detection of Entropy

with Peng Jin

$$Prevalence\ Score(ws_i)$$

$$= \sum_{n_j \in N_w} ds(w, n_j) \times \frac{wnss(ws_i, n_j)}{\sum_{ws_{i'} \in senses(w)} wnss(ws_{i'}, n_j)} \times \frac{1}{rank_{n_j}}$$

Automatic Detection of Entropy

with Peng Jin

$$Prevalence\ Score(ws_i)$$

$$= \sum_{n_j \in N_w} ds(w, n_j) \times \frac{wnss(ws_i, n_j)}{\sum_{ws_{i'} \in senses(w)} wnss(ws_{i'}, n_j)} \times \frac{1}{rank_{n_j}}$$

$$\hat{p}(ws_i) = \frac{prevalence\ score(ws_i)}{\sum_{ws_j \in w} prevalence\ score(ws_j)}$$

Automatic Detection of Entropy

with Peng Jin

$$Prevalence\ Score(ws_i)$$

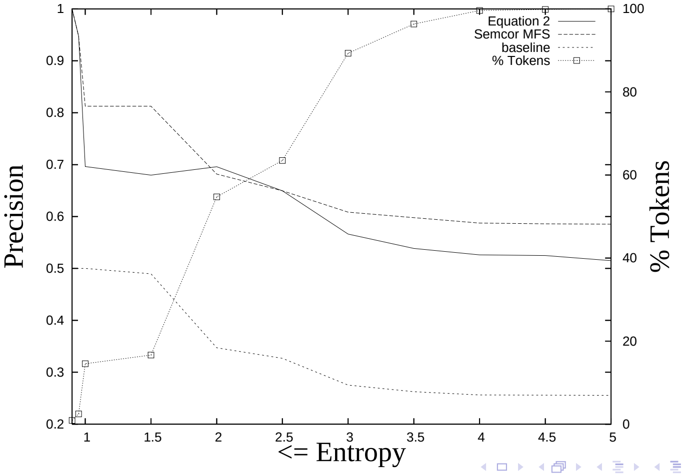
$$= \sum_{n_j \in N_w} ds(w, n_j) \times \frac{wnss(ws_i, n_j)}{\sum_{ws_{i'} \in senses(w)} wnss(ws_{i'}, n_j)} \times \frac{1}{rank_{n_j}}$$

$$\hat{p}(ws_i) = \frac{prevalence\ score(ws_i)}{\sum_{ws_j \in w} prevalence\ score(ws_j)}$$

$$H(senses(w)) = - \sum_{ws_i \in senses(w)} p(ws_i) log(p(ws_i))$$

Automatic Entropy Detection and the First Sense Heuristic

with Peng Jin



Distributional Neighbours of *tie* (noun)

- BNC:
  links (0.165) shirt (0.162) scarf (0.152) jacket (0.142) bond (0.130) match (0.128) trousers (0.126) link (0.125) collar (0.125) dress (0.121)
- Reuters Finance:
  relation (0.329) links (0.247) relationship (0.232) cooperation (0.228) contact (0.142) partnership (0.141) trade (0.137) role (0.133) integration (0.133) finances (0.132)
- Reuters Sport:
  qualifier (0.191) match (0.174) clash (0.150) round (0.135) semifinal (0.132) series (0.129) fixture (0.125) matchup (0.120) encounter (0.120) win (0.116)

Reuters Domain Specific Corpora

40 words (100 sentences each) [Koeling et al., 2005]

- finance and sport codes[Magnini and Cavaglià, 2000]: *club, manager, record, right, bill, check, competition, conversion, crew, delivery, division, fishing, reserve, return, score, receiver, running*
- finance salience: *package, chip, bond, market, strike, bank, share, target*
- sports salience: *fan, star, transfer, striker, goal, title, tie, coach*
- equal salience: *will, phase, half, top, performance, level, country*

Accuracy for Domain Specific Words

Train – Test	RBL	all	F&S cds	F sal	S sal	eq sal
BNC–BNC	19.8	40.7	33.3	51.5	39.7	48.0
SemCor–BNC	19.8	32.0	28.3	44.0	24.6	36.2
FINANCE–FINANCE	19.6	49.9	37.0	70.2	38.5	70.1
SemCor–FINANCE	19.6	33.9	30.3	51.1	22.9	33.5
SPORTS–SPORTS	19.4	43.7	42.6	18.1	65.7	46.9
SemCor–SPORTS	19.4	16.3	9.4	38.1	13.2	12.2

Accuracy for Domain Specific Words

Train – Test	RBL	all	F&S cds	F sal	S sal	eq sal
BNC–BNC	19.8	40.7	33.3	51.5	39.7	48.0
SemCor–BNC	19.8	32.0	28.3	44.0	24.6	36.2
FINANCE–FINANCE	19.6	49.9	37.0	70.2	38.5	70.1
SemCor–FINANCE	19.6	33.9	30.3	51.1	22.9	33.5
SPORTS–SPORTS	19.4	43.7	42.6	18.1	65.7	46.9
SemCor–SPORTS	19.4	16.3	9.4	38.1	13.2	12.2

Application to Japanese

Ryu Iida [Iida et al., 2008]

- Japanese Inventories with Gold-Standard data:
  1. EDR
  2. Iwanami Kokugo Jiten (SENSEVAL-2)
- Semantic Relations not present in all resources
- Increase coverage of LESK using distributional similarity
  - *pigeon: a fat grey and white bird with short legs.*
  - *bird: a creature that is covered with feathers and has wings and two legs.*

Adapting Lesk with Distributional Similarity

use Distributional Similarity to find the maximum similarity between each pair of words in the definitions and take the average.

$$DSlesk(s1, s2) = \frac{1}{|a \in g_1|} \sum_{a \in g_1} \max_{b \in g_2} ds(a, b)$$

Further and Ongoing Work

- automatic text categorisation [Koeling et al., 2007]
- detecting the skew (entropy) to increase performance
- combining first sense heuristic with local evidence
  - unsupervised: using collocates of neighbours [Koeling and McCarthy, 2008]
  - graphical methods [Reddy et al., 2010]
  - weighing local evidence against entropy
- representation of sense

Outline

The Sense Inventory

- raging debate since the very inception of Senseval
- how to make it fair to systems?
  - avoid bias
  - availability to all
- how to make appropriate distinctions?
  - for applications?
  - as humans do?
  - what are word senses anyway?

Granularity

- much WSD done with WordNet because:
  - it has an abundance of useful lexical information
  - it is freely available
  - it comes equipped with a large tagged gold standard corpus (SemCor)

Granularity

- much WSD done with WordNet because:
  - it has an abundance of useful lexical information
  - it is freely available
  - it comes equipped with a large tagged gold standard corpus (SemCor)
- but ...
- many believe too fine grained for WSD [Ide and Wilks, 2006] [Navigli, 2006]
  - we cannot do it and
  - why should we?
- should we settle for what annotators can agree on? (OntoNotes [Hovy et al., 2006])

n#1 54 evidence, grounds – (your basis for belief or disbelief; knowledge on which to base belief; "the evidence that smoking causes lung cancer is very compelling")

n#2 (23) evidence – (an indication that makes something evident; "his trembling was evidence of his fear")

n#3 (7) evidence – ((law) all the means by which any alleged matter of fact whose truth is investigated at judicial trial is established or disproved)

Merging Word Senses

The problem : *evidence*

v#1 (10) attest, certify, manifest, demonstrate, evidence – (provide evidence for; stand as proof of; show by one's behavior, attitude, or external attributes; "His high fever attested to his illness"; "The buildings in Rome manifest a high level of architectural sophistication"; "This decision demonstrates his sense of fairness")

v#2 (3) testify, bear witness, prove, evidence, show – (provide evidence for; "The blood test showed that he was the father"; "Her behavior testified to her incompetence")

v#3 (1) tell, evidence – (give evidence; "he was telling on all his former colleague")

Clustering WordNet Senses

- clustering senses [Navigli, 2006] knowledge-based mapping to ODE
- group verb senses using predicate argument structure [Palmer et al., 2007]
- contexts of senses from manually tagged corpora, or occurrences of monosemous relatives [Agirre and Lopez de Lacalle, 2003]
- Relating WordNet Senses (RLISTS) [McCarthy, 2006]

Relating WordNet Senses (RLISTS) [McCarthy, 2006]

- idea not to group senses but to see how close each is to another
- motivation, one sense may between others

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

RLIST

1 young person:	human offspring,	baby,	clan member
2 human offspring:	clan member,	young person,	baby
3 baby:	human offspring,	young person,	clan member
4 clan member:	human offspring,	young person,	baby

Relating Senses with Distributional Vectors (DIST)

	Nearest Neighbours		
	stool	boss	president
$\overrightarrow{V_{s1}}$	jcn(**seat**,stool)	jcn(**seat**,boss)	...
$\overrightarrow{V_{s2}}$	jcn(**professor**,stool)	jcn(**professor**,boss)	...
$\overrightarrow{V_{s3}}$	jcn(**chairperson**,stool)	jcn(**chairperson**,boss)	...
$\overrightarrow{V_{s4}}$	jcn(**electric**,stool)	jcn(**electric**,boss)	...

Relating Senses with Distributional Vectors (DIST)

	Nearest Neighbours		
	girl	son	baby
$\overrightarrow{V_{s1}}$	jcn(**youth**,girl)	jcn(**youth**,son)	...
$\overrightarrow{V_{s2}}$	jcn(**offspring**,girl)	jcn(**offspring**,son)	...
$\overrightarrow{V_{s3}}$	jcn(**immature**,girl)	jcn(**immature**,son)	...
$\overrightarrow{V_{s4}}$	jcn(**clan**,girl)	jcn(**clan**,son)	...

RLISTs for *child*

sense	**jcn** RLIST		
1	2 (0.11)	3 (0.096)	4 (0.095)
2	4 (0.24)	1 (0.11)	3 (0.099)
3	2 (0.099)	1 (0.096)	4 (0.089)
4	2 (0.24)	1 (0.095)	3 (0.089)
sense	DIST RLIST		
1	3 (0.88)	4 (0.50)	2 (0.48)
2	4 (0.99)	3 (0.60)	1 (0.48)
3	1 (0.88)	4 (0.60)	2 (0.60)
4	2 (0.99)	3 (0.60)	1 (0.50)

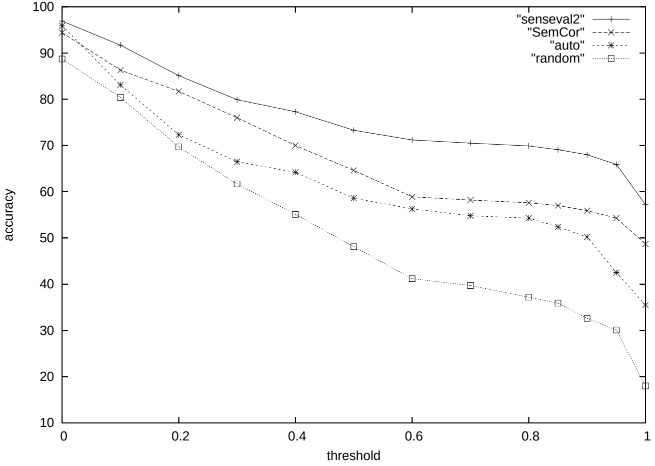
Groupings for *sense* in SENSEVAL-2 LS

sense	GSgr	RLIST
1	g1	5 (0.99) 4 (0.84) 3 (0.83) 2 (-0.22)
a general conscious awareness; "sense of security"		
2	g2	4 (-0.20) 5 (-0.22) 1 (-0.23) 3 (-0.23)
the meaning of a word or expression		
3	g1	4 (0.99) 5 (0.82) 1 (0.82) 2 (-0.23)
sensation		
4	g3	3 (0.99) 5 (0.84) 1 (0.84) 2 (-0.21)
common sense		
5	g4	1 (0.99) 4 (0.84) 3 (0.83) 2 (-0.22)
a natural appreciation or ability; "a musical sense"		

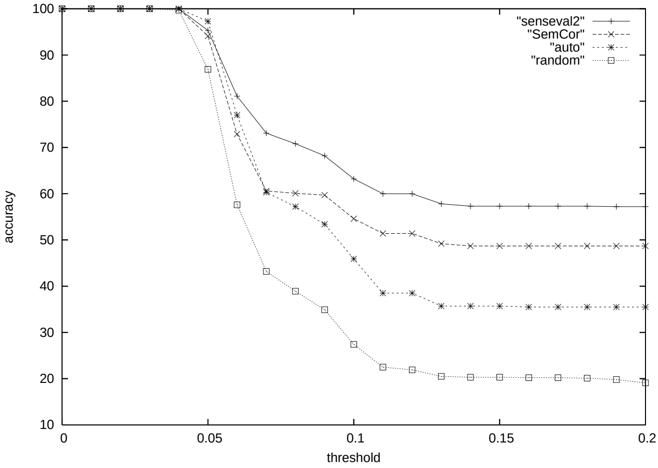
Accuracy of Coarse-grained first sense heuristic on Senseval Lexical Sample

	groupings			thresh on RLISTs			
				DIST		jcn	
	fine-grained	SE2gss	GS	0.90	0.20	0.09	0.0585
SEVAL-2 FS	55.6	65.7	87.8	68.0	85.1	68.2	84.7
SemCor FS	47.0	59.1	82.8	55.9	81.7	59.7	79.4
Auto FS	35.5	48.8	82.9	50.2	72.3	53.4	83.3
random BL	17.5	34.8	65.3	32.6	69.7	34.9	63.5

Relating Senses with DIST and the First Sense Heuristic



Relating senses with JCN and the First Sense Heuristic



Word Sense Induction (WSI)

- induce senses
- may then be applied to WSD
- all methods use corpus co-occurrence data, distributional and graphical
- evaluation still a thorny issue

Distributional Approaches

- Context group discrimination [Schütze, 1998]
- Clustering by committee [Pantel and Lin, 2002]
  - cluster neighbours using average-link clustering
  - residual words not in any committee (not close enough to centroid of formed clusters) remain for next iteration
  - intersecting features in a committee are removed from representation of remaining words so as to allow for less frequent senses

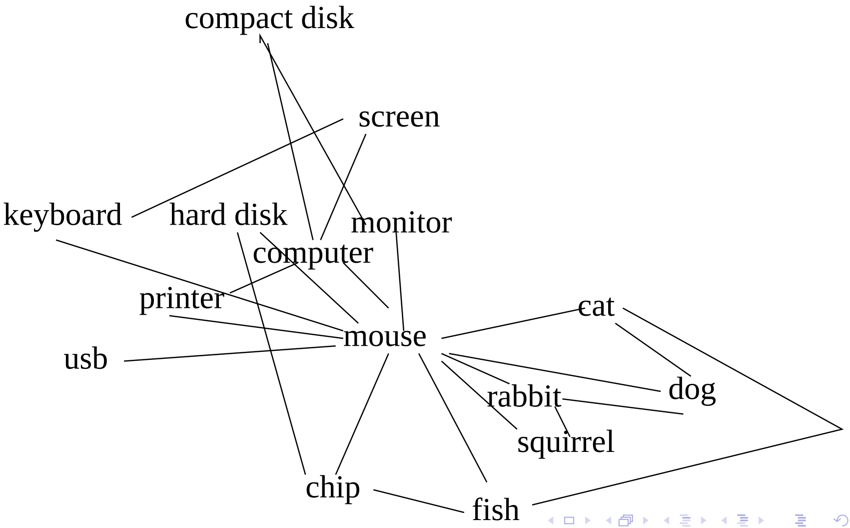
co-occurrence graph [Dorrow and Widdows, 2003]

- vertices words
- edges co-occurrences in syntactic relation of proximity (paragraph)
- create graph for word *w*
- Markov Clustering , random walks within graph will tend to stay in the same cluster rather than jump to more
- 2 steps with parameters
  - inflation (supports popular neighbours and at expense of less frequent, inflates and then rescales so entries sum to 1) and
  - expansion (expands to new node neighbours)

[Dorrow and Widdows, 2003] Algorithm

- remove links of 1, then w
- apply clustering
- remove best cluster and its features
- iterate
- merge similar clusters (using taxonomy?)
- label classes using hypernyms from WordNet

Graphical clustering



Other clustering algorithms

- PageRank [Brin and Page, 1998] used by [Agirre et al., 2006] for WSD
- chinese whispers [Biemann, 2006] (efficient, scales to large graphs useful for WSD features)
- collocations as vertices [Klapaftis and Manandhar, 2008]

Evaluation of WSI?

- against gold standard resource [Pantel and Lin, 2002]
- against gold standard annotations (clusters) e.g. OntoNotes: purity, entropy, v-measure (homogeneity and completeness)
- mapped to inventory (supervised evaluation) and then standard WSD
- separate training and test data [Manandhar et al., 2010]
- bias in evaluation depending on cluster granularity and distribution of instances in cluster [Manandhar et al., 2010]

Credits

Thank you for your attention!

Credits

Thank you for your attention!

Acknowledgments to my collaborators on these projects:
John Carroll, Ryu Iida, Rob Koeling, Peng Jin, Julie Weeds and David Weir.

Agirre, E. and Lopez de Lacalle, O. (2003).
Clustering wordnet word senses.
In *Recent Advances in Natural Language Processing*, pages 121–130, Borovets, Bulgaria.

Agirre, E., Martínez, D., López de Lacalle, O., and Soroa, A. (2006).
Two graph-based algorithms for state-of-the-art WSD.
In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*.

Agirre, E. and Rigau, G. (1996).
Word sense disambiguation using conceptual density.
In *Proceedings of the 16th International Conference of Computational Linguistics, COLING-96*, pages 16–22.

Biemann, C. (2006).

Chinese whispers - an efficient graph clustering algorithm and its application to natural language processing problems.
In *Proceedings of the HLT-NAACL-06 Workshop on Textgraphs-06*, New York, USA.

Brin, S. and Page, M. (1998).
Anatomy of a large-scale hypertextual web search engine.
In *Proceedings of the 7th Conference on World Wide Web*, pages 107–117, Brisbane, Australia.

Briscoe, E. and Carroll, J. (2002).
Robust accurate statistical annotation of general text.
In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC)*, pages 1499–1504, Las Palmas, Canary Islands, Spain.

Chan, Y. S. and Ng, H. T. (2005).
Scaling up word sense disambiguation via parallel texts.

In *Proceedings of the 20th National Conference on Artificial Intelligence (AAAI 2005)*, pages 1037–1042, Pittsburgh, Pennsylvania, USA.

Diab, M. and Resnik, P. (2002).
An unsupervised method for word sense tagging using parallel corpora.
In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 255–262, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Dorrow, B. and Widdows, D. (2003).
Discovering corpus-specific word senses.
In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL-2003)*, pages 79–82, Budapest, Hungary.

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.

Iida, R., McCarthy, D., and Koeling, R. (2008).
Gloss-based semantic similarity metrics for predominant sense acquisition.

In *Proceedings of the Third International Joint Conference on Natural Language Processing*, pages 561–568.

Jiang, J. and Conrath, D. (1997).
Semantic similarity based on corpus statistics and lexical taxonomy.
In *International Conference on Research in Computational Linguistics*, Taiwan.

Kilgarriff, A. and Rychlý, P. (2010).
Semi-automatic dictionary drafting download.
In de Schryver, G.-M., editor, *A Way with Words: Recent Advances in Lexical Theory and Analysis. A Festschrift for Patrick Hanks*. Menha.

Klapaftis, I. P. and Manandhar, S. (2008).
Word sense induction using graphs of collocations.

In *Proceedings of the 18th European Conference On Artificial Intelligence (ECAI-2008)*, Patras, Greece. IOS Press.

Koeling, R. and McCarthy, D. (2008).
From Predicting Predominant Senses to Local Context for Word Sense Disambiguation.
In Bos, J. and Delmonte, R., editors, *Semantics in Text Processing. STEP 2008 Conference Proceedings*, volume 1 of *Research in Computational Semantics*, pages 129–138. College Publications.

Koeling, R., McCarthy, D., and Carroll, J. (2005).
Domain-specific sense distributions and predominant sense acquisition.
In *Proceedings of the joint conference on Human Language Technology and Empirical methods in Natural Language Processing*, pages 419–426, Vancouver, B.C., Canada.

Koeling, R., McCarthy, D., and Carroll, J. (2007).
Text categorization for improved priors of word meaning.
In *Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Text Processing, (CICLing 2007)*, volume 4394 of *Lecture Notes in Computer Science*, pages 241–252, Mexico City. Springer.

Leacock, C., Towell, G., and Voorhees, E. (1993).
Corpus-based statistical sense resolution.
In *Proceedings of the ARPA Workshop on Human Language Technology*, pages 260–265. Morgan Kaufman.

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

Lefever, E. and Hoste, V. (2010).

SemEval-2007 task 3: Cross-lingual word sense disambiguation.
In *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2010)*, Uppsala, Sweden.

Lesk, M. (1986).
Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from and ice cream cone.
In *Proceedings of the ACM SIGDOC Conference*, pages 24–26, Toronto, Canada.

Magnini, B. and Cavaglià, G. (2000).
Integrating subject field codes into WordNet.
In *Proceedings of LREC-2000*, Athens, Greece.

Manandhar, S., Klapaftis, I., Dligach, D., and Pradhan, S. (2010).

Semeval-2010 task 14: Word sense induction and disambiguation.
In *Proceedings of the 5th International Workshop on Semantic Evaluation(SemEval)*, pages 63–68, Uppsala, Sweden. Association for Computational Linguistics.

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., Koeling, R., Weeds, J., and Carroll, J. (2004).
Finding predominant senses in untagged text.
In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 280–287, Barcelona, Spain.

Navigli, R. (2006).
Meaningful clustering of senses helps boost word sense disambiguation performance.
In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics joint with the 21st International Conference on Computational Linguistics (COLING-ACL 2006)*, pages 105–112, Sydney, Australia.

Navigli, R. and Velardi, P. (2005).
Structural semantic interconnections: a knowledge-based approach to word sense disambiguation.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(7):1075–1088.

Ng, H. T. (1997).
Getting serious about word sense disambiguation.

In *Proceedings of the SIGLEX Workshop on Tagging Text with Lexical Semantics: Why What and How?*, pages 1–7, Washington, DC.

Palmer, M., Dang, H. T., and Fellbaum, C. (2007).
Making fine-grained and coarse-grained sense distinctions, both manually and automatically.
Natural Language Engineering, 13(02):137–163.

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.

Ponzetto, S. P. and Navigli, R. (2010).
Knowledge-rich Word Sense Disambiguation rivaling supervised system.

In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1522–1531, Uppsala, Sweden.

Reddy, S., Inumella, A., McCarthy, D., and Stevenson, M. (2010).
Iiith: Domain specific word sense disambiguation.
In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 387–391, Uppsala, Sweden. Association for Computational Linguistics.

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.

Schütze, H. (1992).
Dimensions of meaning.

In *Supercomputing*, pages 787–796, Minneapolis.

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

Sinha, R. and Mihalcea, R. (2007).
Unsupervised graph-based word sense disambiguation using measures of word semantic similarity.
In *Proceedings of the IEEE International Conference on Semantic Computing (ICSC 2007)*, Irvine, CA.

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.

Véronique, H., Hendrickx, I., Daelemans, W., and van den Bosch, A. (2002).
Parameter optimization for machine-learning of word sense disambiguation.
Natural Language Engineering, 8(4):311–325.

Wilks, Y. and Stevenson, M. (1998).
Optimising combinations of knowledge sources for word sense disambiguation.
In *Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics*, pages 1398–1402, Montreal, Canada.

Yarowsky, D. (1994).
Decision lists for lexical ambiguity resolution: application to accent restoration in Spanish and French.

In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 88–95.

Yarowsky, D. (1995).
Unsupervised word sense disambiguation rivaling supervised methods.
In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196.