Exploiting Distributional Similarity for Lexical Acquisition

Diana McCarthy

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

Dialog May 2011

Outline

Lexical Acquisition From Corpora

- statistical (e.g. collocations)
- statistical with processed data, PoS or parsed (e.g. verbal subcategorisation, collocations)
- statistical with manually constructed resource (e.g. selectional preferences)
- parallel corpora (e.g. word senses, multiwords)
- distributional similarity (e.g. semantic classes)

Lexical Acquisition From Corpora

- statistical (e.g. collocations)
- statistical with processed data, PoS or parsed (e.g. verbal subcategorisation, collocations)
- statistical with manually constructed resource (e.g. selectional preferences)
- parallel corpora (e.g. word senses, multiwords)
- distributional similarity (e.g. semantic classes)

Lexical Information

- morphology phonology
- word senses and associated information
- semantic class and semantic relationships
- subcategorisation frames
- selectional preferences
- diathesis and semantic roles
- multiwords, compositionality
- sentiment

Lexical Information

- morphology phonology
- word senses and associated information
- semantic class and semantic relationships
- subcategorisation frames
- selectional preferences
- diathesis and semantic roles
- multiwords, compositionality
- sentiment

Distributional Similarity

plant:

1. *see if you can make the plant grow to its full and healthy height*
2. *A hydro power plant can be operated using either a diverted water stream system*
3. *Job profile of a water/ wastewater treatment plant worker*
4. *We know from a very early age that plants obtain water through their roots*

Distributional Similarity

plant:

1. *see if you can make the plant grow to its full and healthy height*
2. *A hydro power plant can be operated using either a diverted water stream system*
3. *Job profile of a water/ wastewater treatment plant worker*
4. *We know from a very early age that plants obtain water through their roots*

Distributional Similarity

plant:

1. *see if you can make the plant grow to its full and healthy height*
2. *A hydro power plant can be operated using either a diverted water stream system*
3. *Job profile of a water/ wastewater treatment plant worker*
4. *We know from a very early age that plants obtain water through their roots*

water	grow	root	job	hydro	power ...
3	1	1	1	1	1

Proximity Relations

context	frequency		
	plant	*tree*	*factory*
worker	55	8	45
healthy	32	21	3
water	34	18	10
root	8	6	0
operate	4	1	23
power	3	1	49

Proximity Relations

context	frequency		
	plant	*tree*	*factory*
worker	55	8	45
healthy	32	21	3
water	34	18	10
root	8	6	0
operate	4	1	23
power	3	1	49

Distributional Thesaurus (Neighbour) Output:
Word: <closest word> <score> <2nd closest > <score>...
plant: *flower* 0.16 *tree* 0.13 *factory* 0.12 ...scores e.g. [Lin, 1998]

Dependency Relations

context		frequency		
		plant	*tree*	*factory*
grow	verb object	52	60	10
weed	verb object	31	23	2
water	verb object	23	15	4
dead	adj modifier	10	12	0
operate	verb subject	16	2	22
demolish	verb object	11	5	15

Distributional Thesaurus (Neighbour) Output:

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

plant: *tree* 0.17 *flower* 0.16 *factory* 0.15 *bush* 0.13

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 ?

WSD Approaches

- knowledge-based e.g. using semantic relations (WordNet) or overlap of dictionary definitions
- supervised using manually labelled data and machine learning
- unsupervised using corpus based/ distributional methods. Either
  - induce senses (fully unsupervised)
  - or associate distributional information with entries in given sense inventory NB association uses knowledge

The First Sense Heuristic

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

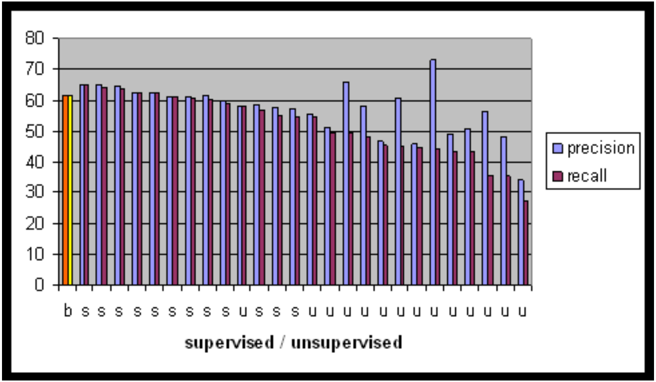
1. (63) plant, works, industrial plant – (buildings for carrying on industrial labor; "they built a large plant to manufacture automobiles")
2. (37) plant, flora, plant life – ((botany) a living organism lacking the power of locomotion)
3. plant – (an actor situated in the audience whose acting is rehearsed 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

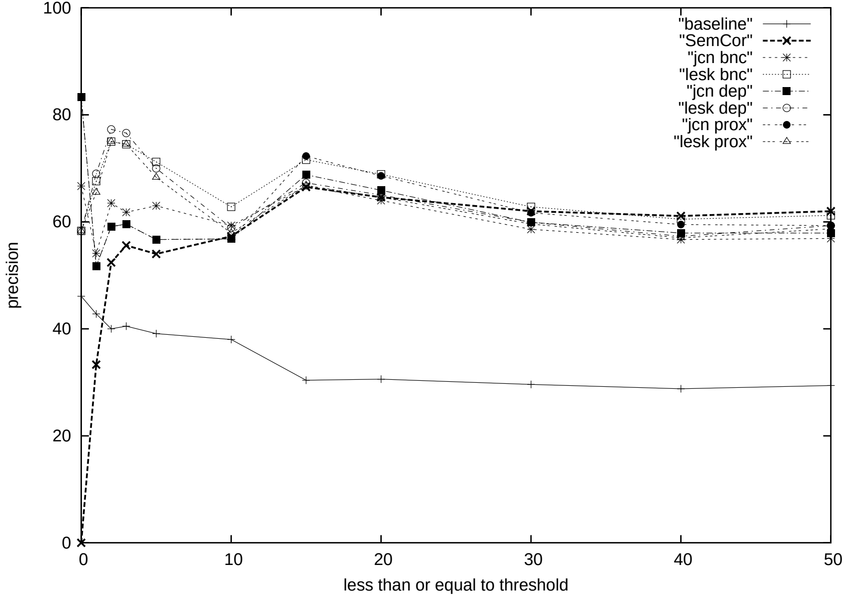
Distributional thesaurus:

- 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 newsire thesaurus: proximity and dependency

SENSEVAL-2 WSD Precision with SemCor Frequency



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.*

Further and Ongoing Work

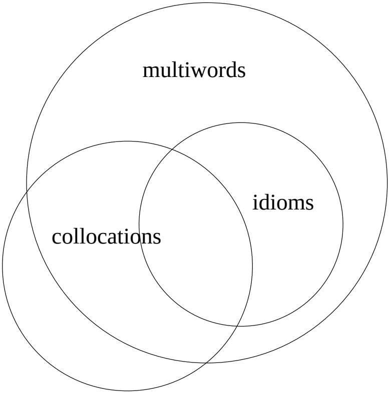
- automatic text categorisation [Koeling et al., 2007]
- detecting the skew (entropy) to increase performance
- combining first sense heuristic with local evidence
    - unsupervised: using collocates of neighbours [Koeling and McCarthy, 2008]
    - graphical methods [Reddy et al., 2010]
    - weighing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

Outline

Multiword Expression NLP Publications

- A Pain in the Neck for NLP [Sag et al., 2002]
- workshops:
  - Collocations/ Multiwords (ACL) 2001, 2003, 2004, 2007, 2009
  - Collocations (Vienna) 2002
  - Collocations and Idioms (Berlin) 2003, 2006,
  - Multiwords (LREC) 2008
  - Multiwords Coling 2010
  - Multiwords ACL 2011
- Multiword Special Issues:
  - Having a crack at a hard nut [Villavicencio et al., 2005]
  - Hard going or plain sailing? [Rayson et al., 2009]

Terminology: Multiwords, Idioms and Collocations



Multiword Expression: A Working Definition

A multiword expression is a combination of two or more words whose semantic, syntactic etc... properties cannot fully be predicted from those of its components, and which therefore has to be listed in a lexicon.
[Boleda and Evert, ESLLI 2009]

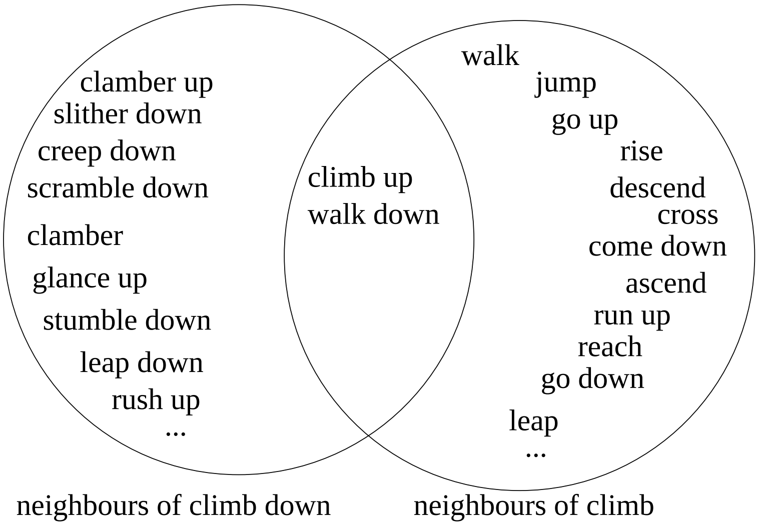
Approaches for Detecting MWEs

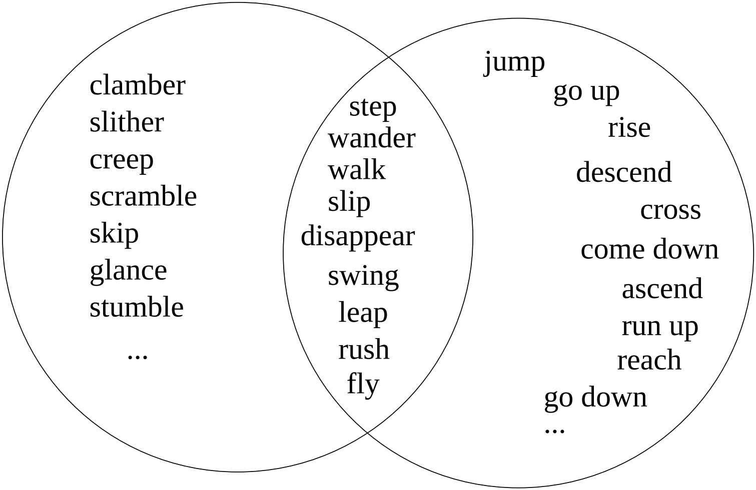
- statistical: e.g. pointwise mutual information $PMI = log \frac{p(chew, fat)}{p(chew)p(fat)}$
- translations in parallel text: *chew the fat* $\leftrightarrow$ *conversar*
- dictionaries: listings, semantic codes and relationships
- lexical variation *couch potato* *sofa potato, couch onion*
- syntactic variation: *take heart*
- distributional similarity:

Distributional Similarity Thesaurus for Compositionality Detection: phrasal verbs

e.g. *blow up* vs *eat up* [McCarthy et al., 2003]

- intuition: the more compositional the phrasal, the closer will be the neighbours of the phrasal and the corresponding constituent verb
- also, the more likely that the verb will appear as a neighbour of the phrasal
- some measures control for particle


clamber up
slither down
creep down
scramble down
clamber
glance up
stumble down
leap down
rush up
...
climb up
walk down
walk
jump
go up
rise
descend
cross
come down
ascend
run up
reach
go down
leap
...
neighbours of climb down
neighbours of climb



neighbours of climb down
with phrasals as verb constituent

neighbours of climb

Results: correlated against human judgments (0-10)

Correlation with Measures Using the Thesaurus		
measure	correlation statistic	p under H_o
overlap	$\rho = 0.166$	0.04
overlapS	$\rho = 0.303$	<0.0007
sameparticle	$\rho = 0.414$	<0.00003
sameparticle-simplex	$\rho = 0.490$	<0.00003
Correlation with Man-made Resources		
WordNet	Mann W	0.008
ANLT Phrasals	Mann W	0.012
ANLT Prepositionals	Mann W	0.334
Correlation with Statistics (used for multiword extraction)		
χ^2	$\rho =$ -0.213	0.0139
LLR	$\rho =$ -0.168	0.0392
MI	$\rho =$ -0.248	0.0047
phrasal Freq	$\rho =$ -0.096	0.156
simplex Freq	$\rho = 0.092$	0.169

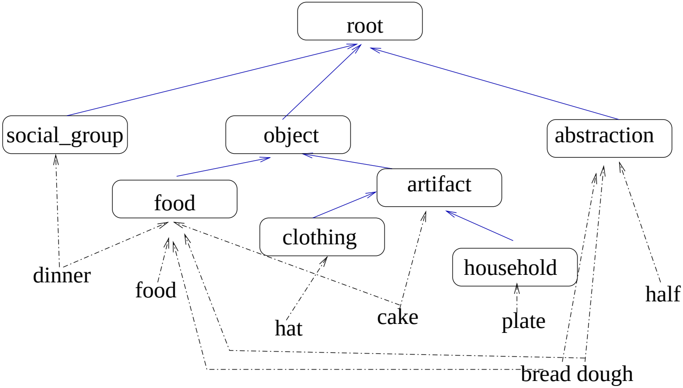
Selectional Preferences for Compositionality: verb-object

[McCarthy et al., 2007] e.g. *shoot the breeze* vs *shoot the gun*

- measure likelihood of verb object combinations
- does the verb have a preference for this sort of object?
- compare WordNet and distributional similarity preference models

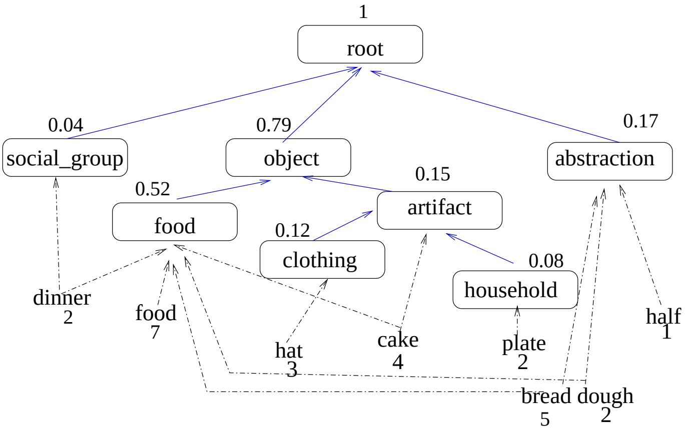
WordNet based models: example *eat*

food 7, bread 5, cake 4, hat 3, dinner 2, dough 2, plate 2, half 1
[Resnik, 1993]



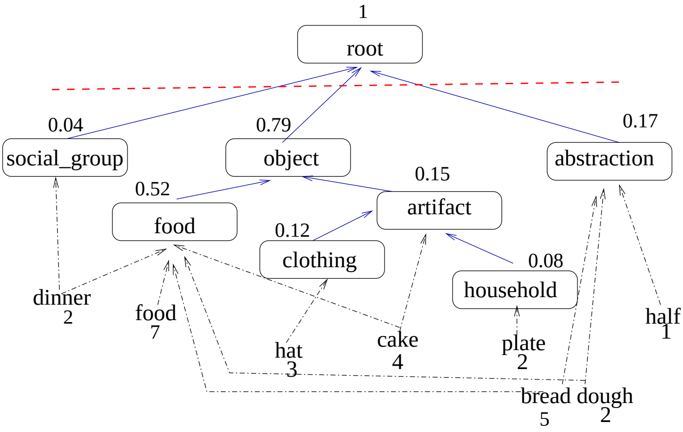
WordNet based models: example *eat*

food 7, bread 5, cake 4, hat 3, dinner 2, dough 2, plate 2, half 1 [Resnik, 1993]

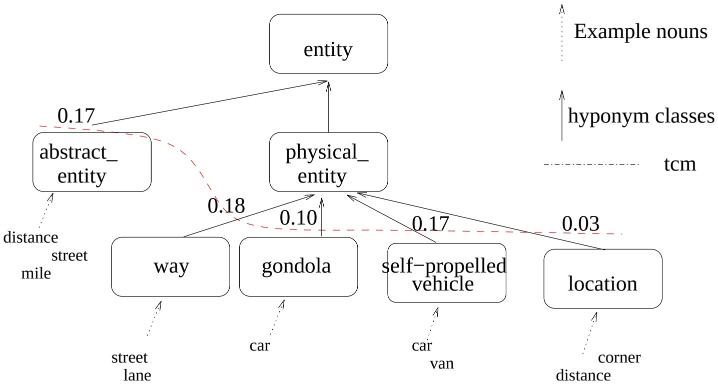


WordNet based TCMs: example *eat*

food 7, bread 5, cake 4, hat 3, dinner 2, dough 2, plate 2, half 1
[Li and Abe, 1998]



Portion of TCM for object of *park*



- Noise from *car* which occurs 174 times (out of 345).
- Contrast tokens (TCM) and type (WNPROTO) to obtain classes for representation, (tokens to estimate probability).

DSPROTOS

[McCarthy et al., 2007]

- nouns are listed in thesaurus built from parses of the BNC
  van: truck 0.230, lorry 0.229, car 0.222, vehicle 0.196, . . .
  bread: loaf 0.195, cheese 0.179, cake 0.169, potato 0.158, . . .
- each listing is considered a grouping or "class"
- classes with at least 2 types
- argument head nouns are disambiguated by whichever class has largest type ratio
- the noun frequency is used to calculate probability over the classes in the model

DSPROTO for object slot of *park*

class ($p(c)$)	disambiguated objects (freq)
van (0.86)	car (174) van (11) vehicle (8) ...
mile (0.05)	street (5) distance (4) mile (1) ...
yard (0.03)	corner (4) lane (3) door (1)
backside (0.02)	backside (2) bum (1) butt (1) ...

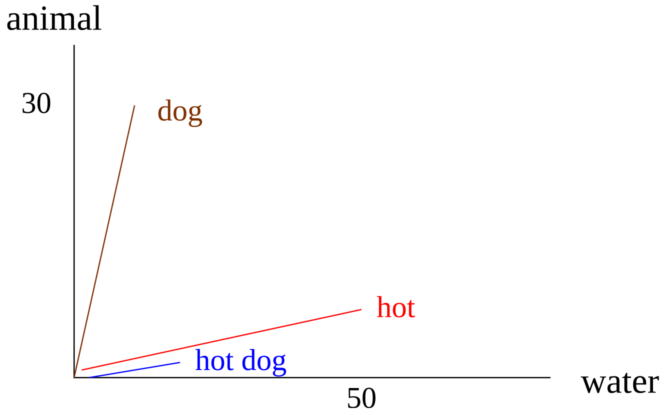
Evaluating DSPROTOS

[Venkatapathy and Joshi, 2005] data

method	ρ	$p <$ (one tailed)
selectional preferences		
TCM	0.090	0.0119
WNPROTO	0.223	0.00003
DSPROTO	0.398	0.00003
features from V&J		
frequency (f1)	0.141	0.00023
MI (f2)	0.274	0.00003
Lin [Lin, 1999] (f3)	0.139	0.00023
LSA2 (f7)	0.209	0.00003
combination		
f2,3,7	0.413	0.00003
f1,2,3,7	0.419	0.00003
DSPROTO f1,2,3,7	0.454	0.00003

Distributional Approaches: Latent Semantic Analysis

[Baldwin et al., 2003] similarity of multiword to constituent words



(related work [Katz and Giesbrecht, 2006])

Current Work: Noun-noun compounds

Siva Reddy, University of York

couch potato vs *roast potato*

- vector space models for the constituents and the compound phrase [Mitchell and Lapata, 2008]
- refined vector space models for each constituent (reduce polysemy) [Erk and Padó, 2010]
- ACL/HLT DisCo 2011: 1st with 2 metrics and 2nd place with both other metrics [Reddy et al., 2011]
- mechanical turker judgments on compositionality of whole phrase and constituent

Summary

- distributional similarity is useful for lexical acquisition because we can build models direct from data
- in this talk I described approaches to acquire
  - sense frequency information from distributional thesauruses, using distributional neighbours as clue to prevalence of different senses
  - compositionality detection of phrasal verbs using distributional thesauruses, comparing distributional neighbours of constituent and phrasal
  - compositionality detection of verb objects using distributional thesauruses to find prototypical arguments (selectional preferences)

Future and Ongoing Work

- current work with Siva Reddy on compositionality detection of noun-noun compounds [Reddy et al., 2011]
- distributional similarity for lexical paraphrases [McCarthy et al., 2010]
- filtering antonyms from distributional similarity thesauruses
- different methods of evaluating distributional similarity using paraphrases and usage similarity judgments

Credits

Thank you for your attention!

Credits

Thank you for your attention!

Acknowledgments to my collaborators on these projects:
John Carroll, Katrin Erk, Abhilash Inumella, Nick Gaylord, Spandana Gella, Ryu Iida, Bill Keller, Rob Koeling, Peng Jin, Aravind Joshi, Suresh Manadhar, Rada Mihalcea, Roberto Navigli, Mark Stevenson, Siva Reddy, Ravi Sinha, Sriram Venkatapathy, Julie Weeds and David Weir.

Baldwin, T., Bannard, C., Tanaka, T., and Widdows, D. (2003).
An empirical model of multiword expression decomposability.
In *Proceedings of the ACL Workshop on multiword expressions: analysis, acquisition and treatment*, pages 89–96.

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.

Erk, K. and Padó, S. (2010).
Exemplar-based models for word meaning in context.
In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, pages 92–97, Stroudsburg, PA, USA. Association for Computational Linguistics.

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.

Katz, G. and Giesbrecht, E. (2006).
Automatic identification of non-compositional multi-word expressions using latent semantic analysis.
In *Proceedings of the ACL Workshop on multiword expressions: identifying and exploiting Underlying Properties*, pages 12–19.

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.

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

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.

Li, H. and Abe, N. (1998).
Generalizing case frames using a thesaurus and the MDL principle.

Computational Linguistics, 24(2):217–244.

Lin, D. (1998).
Automatic retrieval and clustering of similar words.
In *Proceedings of COLING-ACL 98*, Montreal, Canada.

Lin, D. (1999).
Automatic identification of non-compositional phrases.
In *Proceedings of ACL-99*, pages 317–324, Univeristy of Maryland, College Park, Maryland.

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

McCarthy, D., Keller, B., and Carroll, J. (2003).
Detecting a continuum of compositionality in phrasal verbs.

In *Proceedings of the ACL 03 Workshop: Multiword expressions: analysis, acquisition and treatment*, pages 73–80.

McCarthy, D., Keller, B., and Navigli, R. (2010).
Getting synonym candidates from raw data in the english lexical substitution task.
In *Proceedings of the 14th Euralex International Congress*, Leeuwarden, The Netherlands.

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.

McCarthy, D., Venkatapathy, S., and Joshi, A. (2007).
Detecting compositionality of verb-object combinations using selectional preferences.

In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 369–379.

Mitchell, J. and Lapata, M. (2008).
Vector-based models of semantic composition.
In *Proceedings of ACL-08: HLT*, pages 236–244, Columbus, Ohio. Association for Computational Linguistics.

Rayson, P., Piao, S., Sharoff, S., and Stefan Evert, B. n. V. M. (2009).
Multiword expressions: hard going or plain sailing?
Language Resources and Evaluation, 44(1-2):1–5.

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.

Reddy, S., McCarthy, D., Manandhar, S., and Gella, S. (2011).
Exemplar-based word-space model for compositionality detection.
In *Proceedings of the ACL/HLT workshop: Disco Distributional Semantics and Compositionality*, Portland, USA. Association for Computational Linguistics.

Resnik, P. (1993).
Selection and Information: A Class-Based Approach to Lexical Relationships.
PhD thesis, University of Pennsylvania.

Sag, I., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002).

Multiword expressions: A pain in the neck for NLP.
In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2002)*, pages 1–15, Mexico City, Mexico.

Venkatapathy, S. and Joshi, A. K. (2005).
Measuring the relative compositionality of verb-noun (v-n) collocations by integrating features.
In *Proceedings of the joint conference on Human Language Technology and Empirical methods in Natural Language Processing*, pages 899–906, Vancouver, B.C., Canada.

Villavicencio, A., Bond, F., Korhonen, A., and McCarthy, D. (2005).
Introduction to the special issue on multiword expressions: having a crack at a hard nut.
Computer Speech and Language, 19(4):365–377.