What do we mean?

Computational approaches to natural language semantics

Ann Copestake

Natural Language and Information Processing Group
Computer Laboratory
University of Cambridge

May 2008

Outline.

Language and language processing

Compositional semantics

Lexical semantics

Scientific text processing

Natural and non-natural languages

Current research in language processing related to semantics, mostly NLIP group, with flashbacks to Karen's work.

Outline.

Language and language processing
Compositional semantics
- Language as an interface to a microworld
- Broad coverage compositional semantics
- Question answering

Lexical semantics
- Clustering
- Compound nouns
- Ontology extraction

Scientific text processing
- Flyslip
- Hedge terms and citations
- Chemistry Information Extraction

Natural and non-natural languages

Language and language processing

Why is automatic language processing difficult?

Similar strings mean different things:

1. How fast is the TZ? (*fast* CPU speed)
2. How fast will my TZ arrive? (*fast* delivery time)

local ambiguity/vagueness

Different strings mean the same thing:

1. How fast will my TZ arrive? (*my* ordered by me)
2. Please tell me when I can expect the TZ I ordered.

synonymy/near synonymy

Language and language processing

Why is automatic language processing difficult?

Similar strings mean different things:

1. How fast is the TZ? (*fast* CPU speed)
2. How fast will my TZ arrive? (*fast* delivery time)

local ambiguity/vagueness

Different strings mean the same thing:

1. How fast will my TZ arrive? (*my* ordered by me)
2. Please tell me when I can expect the TZ I ordered.

synonymy/near synonymy

Language and language processing

Why is automatic language processing difficult?

Similar strings mean different things:

1. How fast is the TZ? (*fast* CPU speed)
2. How fast will my TZ arrive? (*fast* delivery time)

local ambiguity/vagueness

Different strings mean the same thing:

1. How fast will my TZ arrive? (*my* ordered by me)
2. Please tell me when I can expect the TZ I ordered.

synonymy/near synonymy

Language and language processing

Why is automatic language processing difficult?

Similar strings mean different things:

1. How fast is the TZ? (*fast* CPU speed)
2. How fast will my TZ arrive? (*fast* delivery time)

local ambiguity/vagueness

Different strings mean the same thing:

1. How fast will my TZ arrive? (*my* ordered by me)
2. Please tell me when I can expect the TZ I ordered.

synonymy/near synonymy

Language and language processing

Why is automatic language processing difficult?

Similar strings mean different things:

1. How fast is the TZ? (*fast* CPU speed)
2. How fast will my TZ arrive? (*fast* delivery time)

local ambiguity/vagueness

Different strings mean the same thing:

1. How fast will my TZ arrive? (*my* ordered by me)
2. Please tell me when I can expect the TZ I ordered.

synonymy/near synonymy

Language and language processing

Why is automatic language processing difficult?

Similar strings mean different things:

1. How fast is the TZ? (*fast* CPU speed)
2. How fast will my TZ arrive? (*fast* delivery time)

local ambiguity/vagueness

Different strings mean the same thing:

1. How fast will my TZ arrive? (*my* ordered by me)
2. Please tell me when I can expect the TZ I ordered.

synonymy/near synonymy

Language and language processing

Why is automatic language processing difficult?

Similar strings mean different things:

1. How fast is the TZ? (*fast* CPU speed)
2. How fast will my TZ arrive? (*fast* delivery time)

local ambiguity/vagueness

Different strings mean the same thing:

1. How fast will my TZ arrive? (*my* ordered by me)
2. Please tell me when I can expect the TZ I ordered.

synonymy/near synonymy

Language and language processing

Why is automatic language processing difficult?

Similar strings mean different things:

1. How fast is the TZ? (*fast* CPU speed)
2. How fast will my TZ arrive? (*fast* delivery time)

local ambiguity/vagueness

Different strings mean the same thing:

1. How fast will my TZ arrive? (*my* ordered by me)
2. Please tell me when I can expect the TZ I ordered.

synonymy/near synonymy

Language and language processing

Why is automatic language processing difficult?

Similar strings mean different things:

1. How fast is the TZ? (*fast* CPU speed)
2. How fast will my TZ arrive? (*fast* delivery time)

local ambiguity/vagueness

Different strings mean the same thing:

1. How fast will my TZ arrive? (*my* ordered by me)
2. Please tell me when I can expect the TZ I ordered.

synonymy/near synonymy

Language and language processing

So, natural languages are a bad thing, to be replaced wherever possible by precise, well-specified formal languages?

Natural language properties essential to communication:

- incredibly flexible; learnable while compact
- emergent, evolving systems

Ambiguity/synonymy properties are inherent to flexibility and learnability. (Spärck Jones, 1964, p126–136: 'Model 4 languages')

Language can be indefinitely precise:

- ambiguity is largely local (at least for humans)
- natural languages accommodate (semi-)formal additions

Language and language processing

So, natural languages are a bad thing, to be replaced wherever possible by precise, well-specified formal languages?

Natural language properties essential to communication:

- incredibly flexible; learnable while compact
- emergent, evolving systems

Ambiguity/synonymy properties are inherent to flexibility and learnability. (Spärck Jones, 1964, p126–136: ‘Model 4 languages’)

Language can be indefinitely precise:

- ambiguity is largely local (at least for humans)
- natural languages accommodate (semi-)formal additions

Language and language processing

So, natural languages are a bad thing, to be replaced wherever possible by precise, well-specified formal languages?

Natural language properties essential to communication:

- incredibly flexible; learnable while compact
- emergent, evolving systems

Ambiguity/synonymy properties are inherent to flexibility and learnability. (Spärck Jones, 1964, p126–136: 'Model 4 languages')

Language can be indefinitely precise:

- ambiguity is largely local (at least for humans)
- natural languages accommodate (semi-)formal additions

Outline.

Language and language processing

Compositional semantics

- Language as an interface to a microworld
- Broad coverage compositional semantics
- Question answering

Lexical semantics

- Clustering
- Compound nouns
- Ontology extraction

Scientific text processing

- Flyslip
- Hedge terms and citations
- Chemistry Information Extraction

Natural and non-natural languages

Natural language interfaces to databases (e.g., Copestake and Spärck Jones, 1989)

OWNER	OOid OSurnam OInits
OWNERSHIP	**OWOid OWPid**
PARCEL	PPid **PBid** PStrnum PStrnam PLuc
	PPark PDwell PFl PCityv PSqft
BLOCK	BBid **BWid**
WARD	WWid

- Who owns a house in a street with parcels in Block 3/2?
- Which owners are in Market Place?
  i.e., Which owners own properties which are in Market Place? metonymy

Approach: analyse to produce semantic representation, map to domain semantics, map to SQL.

Natural language interfaces to databases (e.g., Copestake and Spärck Jones, 1989)

OWNER	OOid OSurnam OInits
OWNERSHIP	**OWOid OWPid**
PARCEL	PPid **PBid** PStrnum PStrnam PLuc PPark PDwell PFl PCityv PSqft
BLOCK	BBid **BWid**
WARD	WWid

- Who owns a house in a street with parcels in Block 3/2?
- Which owners are in Market Place?
  i.e., Which owners own properties which are in Market Place? metonymy

Approach: analyse to produce semantic representation, map to domain semantics, map to SQL.

Natural language interfaces to databases (e.g., Copestake and Spärck Jones, 1989)

OWNER	OOid OSurnam OInits
OWNERSHIP	**OWOid OWPid**
PARCEL	PPid **PBid** PStrnum PStrnam PLuc
	PPark PDwell PFI PCityv PSqft
BLOCK	BBid **BWid**
WARD	WWid

- Who owns a house in a street with parcels in Block 3/2?
- Which owners are in Market Place?
  i.e., Which owners own properties which are in Market Place? metonymy

Approach: analyse to produce semantic representation, map to domain semantics, map to SQL.

Natural language interfaces to databases (e.g., Copestake and Spärck Jones, 1989)

OWNER	OOid OSurnam OInits
OWNERSHIP	**OWOid OWPid**
PARCEL	PPid **PBid** PStrnum PStrnam PLuc
	PPark PDwell PFI PCityv PSqft
BLOCK	BBid **BWid**
WARD	WWid

- Who owns a house in a street with parcels in Block 3/2?
- Which owners are in Market Place?
  i.e., Which owners own properties which are in Market Place? metonymy

Approach: analyse to produce semantic representation, map to domain semantics, map to SQL.

Limited domain vs broad coverage language processing

- Until late 1980s: limited domain, often detailed semantics. Systems as agents.
- 1990–2005: broad coverage, information management. Systems as aids to humans.
  - Spoken dialogue systems: limited domain-dependent grammars.
  - Broad coverage text processing: shallow analysis.

  Limited compositional semantics.
- 2005–: question answering (aka ‘semantic search’), robust inference.

Technical progress on broad-coverage compositional semantics

- Better parsing (e.g., PARC/Powerset, DELPH-IN, CCG):
  - Deep parsers incorporating statistical ranking
  - Faster deep parsers
  - More robustness
- Better representations:
  - Language-friendly logical representations (event variables, generalised quantifiers)
  - Underspecification (Alshawi and Crouch (1992): Quasi-logical form (QLF). Copestake, Flickinger, Sag, Pollard (2005): MRS)
  - Semantics from shallower parsers (RMRS)
- Semantics as automatic markup on natural language, not replacement.

Technical progress on broad-coverage compositional semantics

- Better parsing (e.g., PARC/Powerset, DELPH-IN, CCG):
  - Deep parsers incorporating statistical ranking
  - Faster deep parsers
  - More robustness
- Better representations:
  - Language-friendly logical representations (event variables, generalised quantifiers)
  - Underspecification (Alshawi and Crouch (1992): Quasi-logical form (QLF). Copestake, Flickinger, Sag, Pollard (2005): MRS)
  - Semantics from shallower parsers (RMRS)
- Semantics as automatic markup on natural language, not replacement.

Logical representations: first order predicate calculus

Every cat chased some dog

$$\forall x[\text{cat}'(x) \implies \exists y[\text{dog}'(y) \wedge \text{chase}'(x, y)]]$$
$$\exists y[\text{dog}'(y) \wedge \forall x[\text{cat}'(x) \implies \text{chase}'(x, y)]]$$

Cannot decide between scope on the basis of syntax.

Thus requires full parse and scope disambiguation to produce a valid logical representation.

Underspecification allows useful semantic representation even when this is impossible.

Underspecification and Sudoku solving

			7					8
		9					2	
	5			3			9	
8					2			
		6				7		
			4					1
	3			9			6	
	2					4		
7					1			

Solving.

			7					8
		9					2	
	5			3			9	
8					2			
		6				7		
			4					1
	3			9			6	
	2					4		
7					1			

Possibility 1.

			7					8
		9					2	7
	5			3			9	
8					2			
		6				7		
			4					1
	3			9			6	
	2					4		
7					1			

Possibility 2.

			7					8
		9					2	
	5			3			9	7
8					2			
		6				7		
			4					1
	3			9			6	
	2					4		
7					1			

Underspecification.

			7					8
		9					2	7
	5			3			9	7
8					2			
		6				7		
			4					1
	3			9			6	
	2					4		
7					1			

Inference on underspecified form.

			7					8
		9					2	7
	5			3			9	7
8					2			
		6				7		
			4					1
	3			9			6	
	2					4		
7					1			

Inference on underspecified form.

			7					8
		9					2	7
	5			3			9	7
8					2			
		6				7		
			4					1
	3			9			6	
	2					4	7	
7					1			

Semantics via incremental annotation (RMRS)

Most cats noisily chased a large dog

most_DAT cat_NN2 noisily_RR chase_VVD a_AT1 large_JJ dog_NN1

a1:l1:most_q(x1)
a2:l2:cat_n(x2)
a3:l3:noisy(e3)
a4:l4:chase(e4)
a5:l5:a(x5)
a6:l6:large(e6)
a7:l7:dog(x7)

Semantics via incremental annotation (RMRS)

Most cats noisily chased a large dog

most_DAT cat_NN2 noisily_RR chase_VVD a_AT1 large_JJ dog_NN1

a1:l1:most_q(x1)	x1=x2
a2:l2:cat_n(x2)	
a3:l3:noisy(e3)	
a4:l4:chase(e4)	
a5:l5:a(x5)	x5=x7
a6:l6:large(e6)	a6:ARG1(x7) l6=l7
a7:l7:dog(x7)	

Semantics via incremental annotation (RMRS)

Most cats noisily chased a large dog

most_DAT cat_NN2 noisily_RR chase_VVD a_AT1 large_JJ dog_NN1

a1:l1:most_q(x1)	x1=x2
a2:l2:cat_n(x2)	
a3:l3:noisy(e3)	l3=l4 e3=e4
a4:l4:chase(e4)	a4:ARG1(x1) a4:ARG2(x5)
a5:l5:a(x5)	x5=x7
a6:l6:large(e6)	a6:ARG1(x7) l6=l7
a7:l7:dog(x7)	

Semantics via incremental annotation (RMRS)

Most cats noisily chased a large dog

most_DAT cat_NN2 noisily_RR chase_VVD a_AT1 large_JJ dog_NN1

a1:l1:most_q(x1)	x1=x2 a1:RSTR(h1) $h1 =_q l2$
a2:l2:cat_n(x2)	
a3:l3:noisy(e3)	l3=l4 e3=e4
a4:l4:chase(e4)	a4:ARG1(x1) a4:ARG2(x5)
a5:l5:a(x5)	x5=x7 a5:RSTR(h5) $h5 =_q l6$
a6:l6:large(e6)	a6:ARG1(x7) l6=l7
a7:l7:dog(x7)	

Semantics via incremental annotation (RMRS)

Most cats noisily chased a large dog

most_DAT cat_NN2 noisily_RR chase_VVD a_AT1 large_JJ dog_NN1

a1:l1:most_q(x1)	x1=x2 a1:RSTR(h1) $h1 =_q l2$	a1:BODY(l5)
a2:l2:cat_n(x2)		
a3:l3:noisy(e3)	l3=l4 e3=e4	
a4:l4:chase(e4)	a4:ARG1(x1) a4:ARG2(x5)	
a5:l5:a(x5)	x5=x7 a5:RSTR(h5) $h5 =_q l6$	a1:BODY(l3)
a6:l6:large(e6)	a6:ARG1(x7) l6=l7	
a7:l7:dog(x7)		

Semantics via incremental annotation (RMRS)

Most cats noisily chased a large dog

most_DAT cat_NN2 noisily_RR chase_VVD a_AT1 large_JJ dog_NN1

a1:l1:most_q(x1)	x1=x2 a1:RSTR(h1) $h1 =_q l2$	a1:BODY(l3)
a2:l2:cat_n(x2)		
a3:l3:noisy(e3)	l3=l4 e3=e4	
a4:l4:chase(e4)	a4:ARG1(x1) a4:ARG2(x5)	
a5:l5:a(x5)	x5=x7 a5:RSTR(h5) $h5 =_q l6$	a1:BODY(l1)
a6:l6:large(e6)	a6:ARG1(x7) l6=l7	
a7:l7:dog(x7)		

A real example

Very few of the Chinese construction companies consulted were even remotely interested in entering into such an arrangement with a local partner.

A real example

Very few of the Chinese construction companies consulted were even remotely interested in entering into such an arrangement with a local partner.

modified quantifier

A real example

Very few of the Chinese construction companies consulted were even remotely interested in entering into such an arrangement with a local partner.

partitive

A real example

Very few of the Chinese construction companies consulted were even remotely interested in entering into such an arrangement with a local partner.

compound nominal

A real example

Very few of the Chinese construction companies consulted were even remotely interested in entering into such an arrangement with a local partner.

reduced relative

A real example

Very few of the Chinese construction companies consulted were even remotely interested in entering into such an arrangement with a local partner.

modified modifier

A real example

Very few of the Chinese construction companies consulted were even remotely interested in entering into such an arrangement with a local partner.

predeterminer

Question Answering by semantic pattern matching

What eats jellyfish?

Match robust semantics of question with semantics of possible answer:

[?x, a:eat(e), a:ARG1(x), a:ARG2(y), jellyfish(y)] (simplified)

Matches on *turtles eat jellyfish*, *jellyfish are eaten by turtles*

[turtle(x), a:eat(e), a:ARG1(x), a:ARG2(y), jellyfish(y)]

But won't match on *jellyfish eat fish*

[jellyfish(x), a:eat(e), a:ARG1(x), a:ARG2(y), fish(y)]

Jellyfish eaters: pattern matching and inference

Turtles eat jellyfish and they have special hooks in their throats to help them swallow these slimy animals.

Semantic pattern matches
Inference: $P \wedge Q$ entails P

Jellyfish eaters: pattern matching and inference

Sea turtles, ocean sunfish (Mola mola) and blue rockfish all are able to eat large jellyfish, seemingly without being affected by the nematocysts.

Semantic pattern matching: contexts have to be specified to block.
Inference: axioms have to be specified to license.

Negative context may exist in another document, especially in scientific text.

Jellyfish eaters: pattern matching and inference

Sea turtles, ocean sunfish (Mola mola) and blue rockfish all are able to eat large jellyfish, seemingly without being affected by the nematocysts.

Semantic pattern matching: contexts have to be specified to block.
Inference: axioms have to be specified to license.

Negative context may exist in another document, especially in scientific text.

Compositional semantics: summary

- Broad coverage grammars for English and other languages exist which can provide quite detailed compositional semantic representations.
- Logics are relatively ‘language friendly’ and support underspecification.
- Compositional semantics seen as annotation of text rather than replacement.
- Robust inference and semantic pattern matching (NB ongoing work by Bergmair)

Karen on compositional semantics

Spärck Jones, 1985

More recent developments in the theory of grammar, for example Generalized Phrase Structure Grammar (Gazdar et al, 1985) are much more hospitable to exploitation for automatic language processing, though as far as the semantic content necessary for effective language processing goes, one view is that they are essentially still empty vessels, awaiting the water of life in an account of word meanings.

‘They all had a use once’

'They all had a use once'




Outline.

Language and language processing

Compositional semantics
- Language as an interface to a microworld
- Broad coverage compositional semantics
- Question answering

Lexical semantics
- Clustering
- Compound nouns
- Ontology extraction

Scientific text processing
- Flyslip
- Hedge terms and citations
- Chemistry Information Extraction

Natural and non-natural languages

Lexical semantics in language applications

- The Information Retrieval approach: no explicit semantic representation.
- Domain-specific semantics: e.g., interfaces to databases.
- Hand code: e.g., WordNet, specialist terminology resources/ontologies.
- Supervised and unsupervised machine learning.

You shall know a word by the company it keeps! (Firth, 1957)

Words represented as vectors of features:

	feature_1	feature_2	...	feature_n
word_1	$f_{1,1}$	$f_{2,1}$		$f_{n,1}$
word_2	$f_{1,2}$	$f_{2,2}$		$f_{n,2}$
...				
word_m	$f_{1,m}$	$f_{2,m}$		$f_{n,m}$

Features: co-occur with word_n in some window, co-occur with word_n as a syntactic dependent, occur in paragraph_n, occur in document_n ...

First computational application: Spärck Jones (1964)

Words co-occurring with words

	arts	boil	data	function	large	sugar	summarized	water
apricot	0	1	0	0	1	1	0	1
pineapple	0	1	0	0	1	1	0	1
digital	1	0	1	1	0	0	1	0
information	1	0	1	1	0	0	1	0

(from Jurafsky and Martin, 2008)

apricot: { boil, large, sugar, water }
pineapple: { boil, large, sugar, water }
digital: { arts, data, function, summarized }
information: { arts, data, function, summarized }

Clustering: group together words with ‘similar’ vectors.

Words co-occurring with words

	arts	boil	data	function	large	sugar	summarized	water
apricot	0	1	0	0	1	1	0	1
pineapple	0	1	0	0	1	1	0	1
digital	1	0	1	1	0	0	1	0
information	1	0	1	1	0	0	1	0

(from Jurafsky and Martin, 2008)

apricot: { boil, large, sugar, water }
pineapple: { boil, large, sugar, water }
digital: { arts, data, function, summarized }
information: { arts, data, function, summarized }

Clustering: group together words with ‘similar’ vectors.

Words co-occurring with words

	arts	boil	data	function	large	sugar	summarized	water
apricot	0	1	0	0	1	1	0	1
pineapple	0	1	0	0	1	1	0	1
digital	1	0	1	1	0	0	1	0
information	1	0	1	1	0	0	1	0

(from Jurafsky and Martin, 2008)

apricot: { boil, large, sugar, water }
pineapple: { boil, large, sugar, water }
digital: { arts, data, function, summarized }
information: { arts, data, function, summarized }

Clustering: group together words with 'similar' vectors.

Early clustering: Spärck Jones (1967)

Harper (1965): cooccurrence data for 40 nouns from 120,000 words of Russian scientific text: adjective dependents, noun dependents, noun governors.
Harper clustered by:

$$\frac{|V_1 \cap V_2|}{F_1 F_2}$$

where V_1, V_2 are cooccurring sets, F_1, F_2 are the frequencies of the nouns in the corpus.
Spärck Jones (1967): Harper's similarity coefficient is 'of doubtful propriety'. Instead clustered ('clumped') by Jaccard:

$$\frac{|V_1 \cap V_2|}{|V_1 \cup V_2|}$$

Early clustering: Spärck Jones (1967)

Harper (1965): cooccurrence data for 40 nouns from 120,000 words of Russian scientific text: adjective dependents, noun dependents, noun governors.
Harper clustered by:

$$\frac{|V_1 \cap V_2|}{F_1 F_2}$$

where V_1, V_2 are cooccurring sets, F_1, F_2 are the frequencies of the nouns in the corpus.
Spärck Jones (1967): Harper's similarity coefficient is 'of doubtful propriety'. Instead clustered ('clumped') by Jaccard:

$$\frac{|V_1 \cap V_2|}{|V_1 \cup V_2|}$$

Spärck Jones (1967)

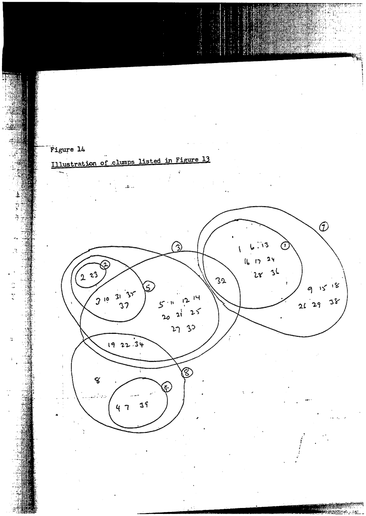
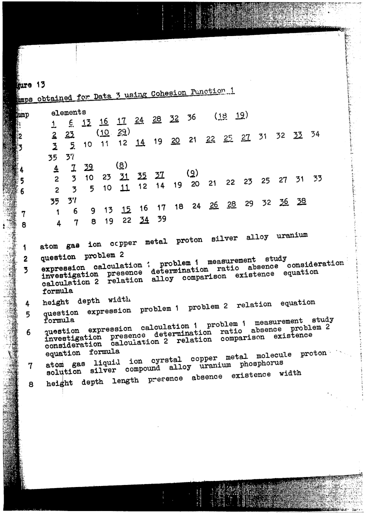


Figure 14

Illustration of clumps listed in Figure 13



gure 13

umps obtained for Data 3 using Cohesion Function 1

ump	elements
1	1 6 13 16 17 24 28 32 36 (18 19)
2	2 23 (10 29)
3	3 5 10 11 12 14 19 20 21 22 25 27 31 32 33 34 35 37
4	4 7 39 (8)
5	2 3 10 23 31 35 37 (9)
6	2 3 5 10 11 12 14 19 20 21 22 23 25 27 31 33 35 37
7	1 6 9 13 15 16 17 18 24 26 28 29 32 36 38
8	4 7 8 19 22 34 39

1 atom gas ion ccpper metal proton silver alloy uranium

2 question problem 2

3 expression calculation 1 problem 1 measurement study consideration investigation presence determination ratio absence existence equation calculation 2 relation alloy comparison formula

4 height depth width

5 question expression problem 1 problem 2 relation equation formula

6 question expression calculation 1 problem 1 measurement study investigation presence determination ratio absence problem 2 consideration calculation 2 relation comparison existence equation formula

7 atom gas liquid ion cyrstal copper metal molecule proton solution silver compound alloy uranium phosphorus

8 height depth length presence absence existence width

IR (Robertson and Spärck Jones, 1976, 1994)

Term Frequency:

```
TF(i,j) = number of terms t(i) in document d(j)
```

Collection Frequency Weight (inverse document frequency):

```
CFW(i) = log N - log n
where n is the number of documents t(i) occurs in,
N is the total number of documents
```

Document length:

```
NDL = number of terms in d(j) / average number terms
```

Combined weight:

```
CW(i,j) = [CFW(i)*TF(i,j)*(K+1)] / [K*NDL(j)+TF(i,j)]
```

Verbs in biomedical text (Korhonen et al, 2006)

Gold standard clusters:

1 Have an effect on activity (BIO/29)	4 Experimental Procedures (BIO/30)
1.1 Activate / Inactivate	**4.1 Prepare**
1.1.1 Change activity: *activate, inhibit*	4.1.1 Wash: *wash, rinse*
1.1.2 Suppress: *suppress, repress*	4.1.2 Mix: *mix*
1.1.3 Stimulate: *stimulate*	4.1.3 Label: *stain, immunoblot*
1.1.4 Inactivate: *delay, diminish*	4.1.4 Incubate: *preincubate, incubate*
1.2 Affect	4.1.5 Elute: *elute*
1.2.1 Modulate: *stabilize, modulate*	**4.2 Precipitate:** *coprecipitate coimmunoprecipitate*
1.2.2 Regulate: *control, support*	**4.3 Solubilize:** *solubilize,lyse*
1.3 Increase / decrease: *increase,* decrease	**4.4 Dissolve:** *homogenize, dissolve*
1.4 Modify: *modify, catalyze*	**4.5 Place:** *load, mount*

Verb clustering using a range of features derived via robust parsing (Briscoe and Carroll, 2002).

Distributional differences (Copestake, 2005)

Magnitude adjectives and non-physical-solid nouns.
Distributional data from the British National Corpus (100 million words)

	importance	success	majority	number	proportion	quality	role	problem	part	winds	support	rain
great	310	360	382	172	9	11	3	44	71	0	22	0
large	1	1	112	1790	404	0	13	10	533	0	1	0
high	8	0	0	92	501	799	1	0	3	90	2	0
major	62	60	0	0	7	0	272	356	408	1	8	0
big	0	40	5	11	1	0	3	79	79	3	1	1
strong	0	0	2	0	0	1	8	0	3	132	147	0
heavy	0	0	1	0	0	1	0	0	1	2	4	198

Andersen: evidence from error corpus that language learners overuse *big*.

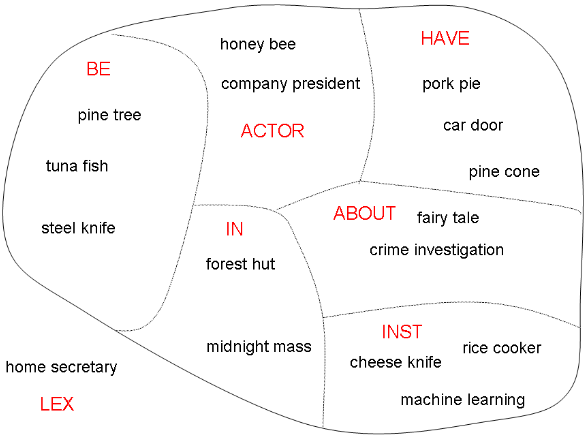
Compound noun relations

- *cheese knife*: knife for cutting cheese
- *steel knife*: knife made of steel
- *kitchen knife*: knife characteristically used in the kitchen

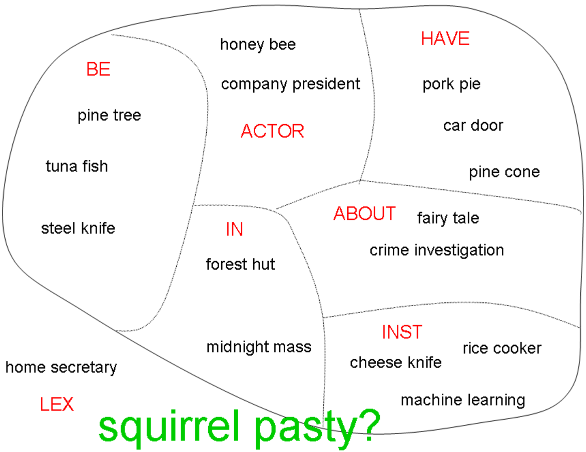
(Spärck Jones (1983) on compound nouns: implications for overall processing architecture.)

- Syntactic parsers can't distinguish: N1(x), N2(y), compound(x,y)
- One approach: human annotation of compounds, machine learning of unseen examples.

Compound noun relation learning (Ó Séaghdha, 2007)



Compound noun relation learning (Ó Séaghdha)



Compound noun relation learning (Ó Séaghdha)

- Treat compounds as single words: doesn't work!
- Constituent similarity: compounds x1 x2 and y1 y2, compare x1 vs y1 and x2 vs y2.
  squirrel vs *pork*, *pasty* vs *pie*
- Relational similarity: **sentences** with x1 and x2 vs sentences with y1 and y2.
  squirrel is very tasty, especially in a pasty vs *pies are filled with tasty pork*

Human annotation

- Preliminary to supervised machine learning, evaluation of unsupervised techniques.
- Methodology: define categories, develop guidelines, multiple annotators, measure annotator agreement, refine categories and guidelines . . .
- Agreement of 70% quite usual in semantic annotation.
- What's going on?
  Sometimes, local effects: *sponsorship cash*. Cash gained through sponsorship (INST) or sponsorship in form of cash (BE)?

Human annotation

- Preliminary to supervised machine learning, evaluation of unsupervised techniques.
- Methodology: define categories, develop guidelines, multiple annotators, measure annotator agreement, refine categories and guidelines . . .
- Agreement of 70% quite usual in semantic annotation.
- What's going on?
  Sometimes, local effects: *sponsorship cash*. Cash gained through sponsorship (INST) or sponsorship in form of cash (BE)?

Ontology extraction

Extraction of hyponymies

A beetle is an insect

A tibia is a bone

Ontology extraction (Herbelot, 2007, 2008)

- Improving recall by extracting complex examples with robust semantic patterns:
  Opah (also known colloquially as moonfish, sunfish, kingfish, and Jerusalem haddock) are large, colourful, deep-bodied pelagic Lampriform fish comprising the small family Lampridae (also spelt Lamprididae).
- Learning difference between generic and individual uses:
  - A whale is a mammal.
  - A whale escaped from a zoo yesterday.

Computational lexical semantics

- Karen was a pioneer of many of the basic methods.
- Research really took off in the 1990s with the availability of corpora (and disk space).
- Many linguistic phenomena involved: generics, compounds, polysemy, metonymy.
- Semantic annotation requires considerable thought about phenomenon and experimentation to be successful: even then, quite low agreement.
- Unsupervised methods, such as clustering, are very attractive, but evaluation can be a problem (especially soft clustering).

Outline.

Language and language processing

Compositional semantics
- Language as an interface to a microworld
- Broad coverage compositional semantics
- Question answering

Lexical semantics
- Clustering
- Compound nouns
- Ontology extraction

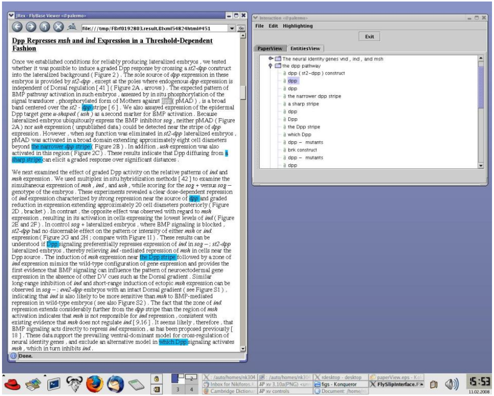
Scientific text processing
- Flyslip
- Hedge terms and citations
- Chemistry Information Extraction

Natural and non-natural languages

FlySlip: aiding manual curation

- FlyBase: database for Drosophila genetics, manually constructed from literature.
- FlySlip: using NLP to improve the process: NLIP group and Dept of Genetics (Karamanis, Seal, Lewin, McQuilton, Vlachos, Gasperin, Drysdale, Briscoe)

FlySlip: PaperBrowser



- Entity view: anaphorically-linked gene references highlighted (focus determined by curator).
- Base NPs identified: more useful than just gene names.

Hedge terms: Medlock and Briscoe (2007)

Hedge: *a word or phrase used to allow for additional possibilities or to avoid over-precise commitment.* (OED)

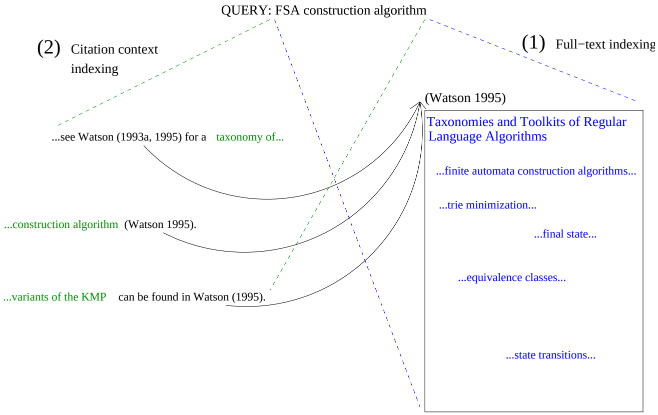
Hedge classification is the task of identifying and labeling the use of speculative language in written text.

Speculative: This unusual substrate specificity may explain why Dronc is resistant to inhibition by the pan-caspase inhibitor.

Non-speculative: These results demonstrate that ADGF-A overexpression can partially rescue the effects of constitutively active Toll signaling in larvae

Weakly-supervised machine learning technique.

Citations in IR: Ritchie (2008)



SciBorg: extracting the science from scientific publications

- Use RMRS language as semantic annotation on chemistry papers (standoff annotation on SciXML).
- Support ontology extraction, discourse markup and information extraction.
- NLIP group, Chemistry dept, CeSC (Copestake, Teufel, Murray-Rust, Parker, Corbett, Rupp, Siddharthan, Waldron) with IUCr, Nature, Royal Society of Chemistry (Batchelor).

SciBorg: information extraction

Paper 1: The synthesis of 2,8-dimethyl-6H,12H-5,11 methanodibenzo[b,f][1,5]diazocine (Troger's base) from p-toluidine and of two Troger's base analogs from other anilines

Paper 2: ... Tröger's base (TB) ... The TBs are usually prepared from para-substituted anilines

Eventually, robust inference: e.g., search for papers describing Tröger's base syntheses which don't involve anilines?

OSCAR: chemistry terms (Corbett, Murray-Rust)

Nitrone dipolar cycloaddition routes to piperidines and indolizidines. Part 9. Formal synthesis of (−)-pinidine and total synthesis of (−)-histrionicotoxin, (+)-histrio

Results and discussion

Model cyclisation studies

We first examined the model cyclisation of the non-terminal alkyne, hept-5-ynylhydroxylamine **7**, prepared by sodium cyanoborohydride reduction of the corresponding oxime **6**. Formation of the nitrone **8** occurred in 94% overall yield after the reaction mixture had been heated in refluxing toluene for 2 hours (Scheme 2). This is consistent with our general observation that hydroxylamine-alkyne cyclisations onto terminal and silyl-substituted acetylenes are much faster than cyclisations onto other non-terminal alkynes.[19,22] This observation is analogous to those of Ciganek[31] and Black[32] in the Cope–House cyclisation[33] of alkenyl hydroxylamines.

An enantioselective synthesis of HTX **1** would require the (*S*)-hydroxylamino-alkyne derivative (*e.g.* **40**) from which all other stereocentres could then be induced diastereoselectively. Whilst a number of methods for the enantioselective synthesis of hydroxylamines exist (*e.g.* oxidation of amines,[34] nucleophilic displacement of triflates,[35**] addition of organometallics to nitrones[36–39] and oximes[40]) it was decided to mimic the enolate hydroxylamination protocol of Oppolzer,[41] but using an Evans oxazolidinone auxiliary. The terminally silylated heptynoic acid **12** was prepared in 4 steps from commercially available hex-5-yn-1-ol **9** as shown in Scheme 3, and was then coupled to the Evans benzyl oxazolidinone auxiliary[42] by a mixed anhydride method. Attempted electrophilic hydroxylamination of the sodium enolate of the *N*-acyloxazolidinone **13** using 1-chloro-1-nitrosocyclohexane followed by acid hydrolysis of the nitrone intermediate, base extraction (to release the intermediate hydroxylamine **14**) and stirring at 25 °C for 1 hour to induce the Cope–House cyclisation was unsatisfactory, giving the required nitrone **15** in poor yield, along with the by-product **16**, resulting from attack on the carbonyl of the auxiliary by the hydroxylamine **14**. Evans has noted similar side reactions with related amines,[43] and clearly the more demanding cyclisation conditions required for a non-terminal alkyne would be incompatible with the Evans auxiliary. The diastereoselectivity of the hydroxylamination reaction was assumed to follow the usual reactivity pattern of the Evans auxiliaries,[44] and was shown by ^{1}H NMR spectroscopy to be >95 : 5. Given the above mentioned problems this approach was abandoned in favour of the Oppolzer camphorsultam auxiliary.[41,45]

- Experimental data
- Ontology term
- Chemical (etc.) with structure
- Chemical (etc.), without structure
- Reaction
- Chemical adjective
- enzyme -ase word
- Chemical prefix

Find: cance Next Previous Highlight all Match case

OSCAR: chemistry terms (Corbett, Murray-Rust)

Nitrone dipolar cycloaddition routes to piperidines and indolizidines. Part 9. Formal synthesis of (−)-pinidine and total synthesis of (−)-histrionicotoxin, (+)-histrio

Results and discussion

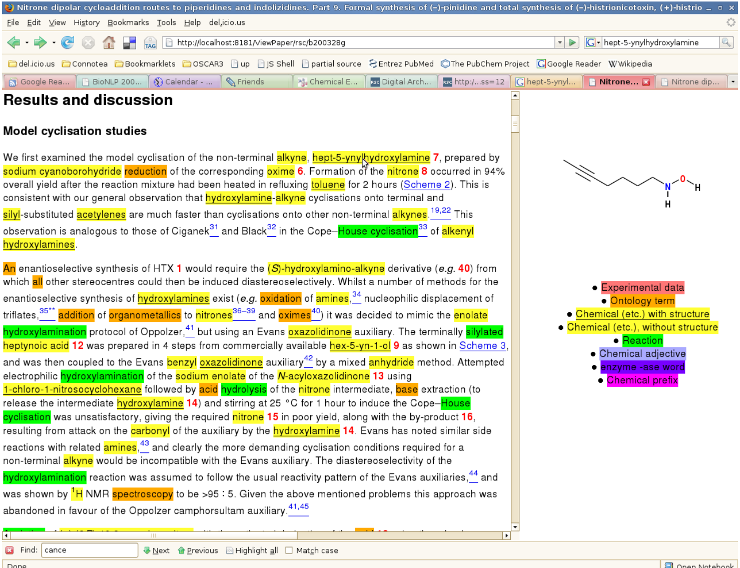
Model cyclisation studies

We first examined the model cyclisation of the non-terminal alkyne, hept-5-ynylhydroxylamine **7**, prepared by sodium cyanoborohydride reduction of the corresponding oxime **6**. Formation of the nitrone **8** occurred in 94% overall yield after the reaction mixture had been heated in refluxing toluene for 2 hours (Scheme 2). This is consistent with our general observation that hydroxylamine-alkyne cyclisations onto terminal and silyl-substituted acetylenes are much faster than cyclisations onto other non-terminal alkynes.[19,22] This observation is analogous to those of Ciganek[31] and Black[32] in the Cope–House cyclisation[33] of alkenyl hydroxylamines.

An enantioselective synthesis of HTX **1** would require the (*S*)-hydroxylamino-alkyne derivative (*e.g.* **40**) from which all other stereocentres could then be induced diastereoselectively. Whilst a number of methods for the enantioselective synthesis of hydroxylamines exist (*e.g.* oxidation of amines,[34] nucleophilic displacement of triflates,[35**] addition of organometallics to nitrones[36–39] and oximes[40]) it was decided to mimic the enolate hydroxylamination protocol of Oppolzer,[41] but using an Evans oxazolidinone auxiliary. The terminally silylated heptynoic acid **12** was prepared in 4 steps from commercially available hex-5-yn-1-ol **9** as shown in Scheme 3, and was then coupled to the Evans benzyl oxazolidinone auxiliary[42] by a mixed anhydride method. Attempted electrophilic hydroxylamination of the sodium enolate of the *N*-acyloxazolidinone **13** using 1-chloro-1-nitrosocyclohexane followed by acid hydrolysis of the nitrone intermediate, base extraction (to release the intermediate hydroxylamine **14**) and stirring at 25 °C for 1 hour to induce the Cope–House cyclisation was unsatisfactory, giving the required nitrone **15** in poor yield, along with the by-product **16**, resulting from attack on the carbonyl of the auxiliary by the hydroxylamine **14**. Evans has noted similar side reactions with related amines,[43] and clearly the more demanding cyclisation conditions required for a non-terminal alkyne would be incompatible with the Evans auxiliary. The diastereoselectivity of the hydroxylamination reaction was assumed to follow the usual reactivity pattern of the Evans auxiliaries,[44] and was shown by ^{1}H NMR spectroscopy to be >95 : 5. Given the above mentioned problems this approach was abandoned in favour of the Oppolzer camphorsultam auxiliary.[41,45]

- Experimental data
- Ontology term
- Chemical (etc.) with structure
- Chemical (etc.), without structure
- Reaction
- Chemical adjective
- enzyme -ase word
- Chemical prefix

OSCAR: chemistry terms (Corbett, Murray-Rust)



Nitrone dipolar cycloaddition routes to piperidines and indolizidines. Part 9. Formal synthesis of (−)-pinidine and total synthesis of (−)-histrionicotoxin, (+)-histrio

http://localhost:8181/ViewPaper/rsc/b200328g

Results and discussion

Model cyclisation studies

We first examined the model cyclisation of the non-terminal alkyne, hept-5-ynylhydroxylamine **7**, prepared by sodium cyanoborohydride reduction of the corresponding oxime **6**. Formation of the nitrone **8** occurred in 94% overall yield after the reaction mixture had been heated in refluxing toluene for 2 hours (Scheme 2). This is consistent with our general observation that hydroxylamine-alkyne cyclisations onto terminal and silyl-substituted acetylenes are much faster than cyclisations onto other non-terminal alkynes.[19,22] This observation is analogous to those of Ciganek[31] and Black[32] in the Cope–House cyclisation[33] of alkenyl hydroxylamines.

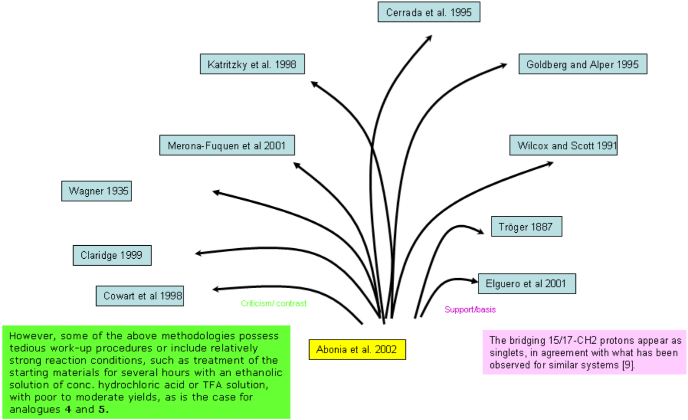
An enantioselective synthesis of HTX **1** would require the (*S*)-hydroxylamino-alkyne derivative (*e.g.* **40**) from which all other stereocentres could then be induced diastereoselectively. Whilst a number of methods for the enantioselective synthesis of hydroxylamines exist (*e.g.* oxidation of amines,[34] nucleophilic displacement of triflates,[35**] addition of organometallics to nitrones[36–39] and oximes[40]) it was decided to mimic the enolate hydroxylamination protocol of Oppolzer,[41] but using an Evans oxazolidinone auxiliary. The terminally silylated heptynoic acid **12** was prepared in 4 steps from commercially available hex-5-yn-1-ol **9** as shown in Scheme 3, and was then coupled to the Evans benzyl oxazolidinone auxiliary[42] by a mixed anhydride method. Attempted electrophilic hydroxylamination of the sodium enolate of the *N*-acyloxazolidinone **13** using 1-chloro-1-nitrosocyclohexane followed by acid hydrolysis of the nitrone intermediate, base extraction (to release the intermediate hydroxylamine **14**) and stirring at 25 °C for 1 hour to induce the Cope–House cyclisation was unsatisfactory, giving the required nitrone **15** in poor yield, along with the by-product **16**, resulting from attack on the carbonyl of the auxiliary by the hydroxylamine **14**. Evans has noted similar side reactions with related amines,[43] and clearly the more demanding cyclisation conditions required for a non-terminal alkyne would be incompatible with the Evans auxiliary. The diastereoselectivity of the hydroxylamination reaction was assumed to follow the usual reactivity pattern of the Evans auxiliaries,[44] and was shown by ^{1}H NMR spectroscopy to be >95 : 5. Given the above mentioned problems this approach was abandoned in favour of the Oppolzer camphorsultam auxiliary.[41,45]

- Experimental data
- Ontology term
- Chemical (etc.) with structure
- Chemical (etc.), without structure
- Reaction
- Chemical adjective
- enzyme -ase word
- Chemical prefix

Find: cance Next Previous Highlight all Match case

Done

Citation classification (Teufel, Siddharthan, Batchelor)



Outline.

Language and language processing

Compositional semantics

- Language as an interface to a microworld
- Broad coverage compositional semantics
- Question answering

Lexical semantics

- Clustering
- Compound nouns
- Ontology extraction

Scientific text processing

- Flyslip
- Hedge terms and citations
- Chemistry Information Extraction

Natural and non-natural languages

Semantic web, scientific text and language processing

- Description logics, OWL etc.
- Ontologies/terminology resources.
- Chemistry Markup Language (CML: Murray-Rust).
- Availability of texts in XML for language processing.
- Publishing as mixture of texts and structured output (e.g., spectra).

Semantic web publishing

- Claim: Language processing will soon just be needed for legacy texts. All new scientific publication will use semantic markup.
- Scientific publishing is not simply about facts slotting into an agreed framework.
- Counter-claim 1: where we understand what's going on in scientific text, we can learn to annotate it automatically. But most aspects cannot currently be formalised.
- Counter-claim 2: we need language processing experiments and methodology to work out how to do semantic markup.

Semantic web publishing

- Claim: Language processing will soon just be needed for legacy texts. All new scientific publication will use semantic markup.
- Scientific publishing is not simply about facts slotting into an agreed framework.
- Counter-claim 1: where we understand what's going on in scientific text, we can learn to annotate it automatically. But most aspects cannot currently be formalised.
- Counter-claim 2: we need language processing experiments and methodology to work out how to do semantic markup.

Semantic web publishing

- Claim: Language processing will soon just be needed for legacy texts. All new scientific publication will use semantic markup.
- Scientific publishing is not simply about facts slotting into an agreed framework.
- Counter-claim 1: where we understand what's going on in scientific text, we can learn to annotate it automatically. But most aspects cannot currently be formalised.
- Counter-claim 2: we need language processing experiments and methodology to work out how to do semantic markup.

Semantic web publishing

- Claim: Language processing will soon just be needed for legacy texts. All new scientific publication will use semantic markup.
- Scientific publishing is not simply about facts slotting into an agreed framework.
- Counter-claim 1: where we understand what's going on in scientific text, we can learn to annotate it automatically. But most aspects cannot currently be formalised.
- Counter-claim 2: we need language processing experiments and methodology to work out how to do semantic markup.

Information Layer and scientific publishing

- ‘Information Layer’ (Spärck Jones 2007): connection via words may be good enough for many computing system tasks.
- Semantic publishing best seen as an addition to natural language, not a replacement. One objective should be to make scientific publications more accessible to humans.
- Natural language is flexible and adaptable: can this be emulated in formal languages?

Maths texts and natural languages (Ganesalingam)

Then $\underline{V = U \cap H \text{ for some } U \text{ in } \mathcal{T}}$, by definition of $\mathcal{T}_H$, and $U \cap H = i^{-1}(U)$, so $g^{-1}(V) = g^{-1}(i^{-1}(U)) = (i \circ g)^{-1}(U)$.

Sutherland, W. A., Introduction to Metric and Topological Spaces, OUP 1975, p. 52.

Analogous to 'donkey sentence' in linguistics.

Every farmer who owns a donkey beats it.

$$\forall x[farmer(x) \wedge \exists y[donkey(y) \wedge own(x,y)]] \implies beat(x,y)]$$

Maths texts and natural languages (Ganesalingam)

Then $\underline{V = U \cap H \text{ for some } U \text{ in } \mathcal{T}}$, by definition of $\mathcal{T}_H$, and $U \cap H = i^{-1}(U)$, so $g^{-1}(V) = g^{-1}(i^{-1}(U)) = (i \circ g)^{-1}(U)$.

Sutherland, W. A., Introduction to Metric and Topological Spaces, OUP 1975, p. 52.

Analogous to 'donkey sentence' in linguistics.

Every farmer who owns a donkey beats it.

$$\forall x[\mathit{farmer}(x) \wedge \exists y[\mathit{donkey}(y) \wedge \mathit{own}(x, y)]] \implies \mathit{beat}(x, y)]$$

Concluding comments

- Computational semantics: enrich texts to make aspects of meaning more accessible to subsequent processing.
- Underspecifiable, ‘surfacy’ representations of compositional semantics: logically defined, but robustness, reasonable processing speed.
- Lexical semantics by distributional methods can (partially) model ambiguity/synonymy behaviour (though evaluation still a problem).
- Practical applications to scientific text processing.
- Karen’s ‘Information Layer’ challenges us to take natural language’s properties seriously.