

Learning Lexical Semantic Representations

Timothy Baldwin



THE UNIVERSITY OF
MELBOURNE

Introduction

- Lexical semantic representations are all well and good,
BUT:
how can we produce them automatically?

Basic Approaches to Learning Lexical Representations

- **Lexico-syntactic patterns:** identify the lexico-syntactic patterns associated with a given phenomenon, and look for corpus occurrences thereof
- ★ **assumption:** lexico-syntax is a strong predictor of semantics
- ★ **presupposes:** corpus data and (optionally) pre-processing capabilities

- **Lexical similarity:** identify “near-matches” (synonyms, near-synonyms, ...) to the given lexical item, and “inherit” their lexical semantic properties
- ★ **assumption:** near-matches to a given lexical item have the same lexical semantic properties
- ★ **presupposes:**
 1. training data
 2. some method for calculating lexical similarity

- **Resource mining:** mine pre-existing language resource(s) for relevant information
- ★ **assumption:** it is possible to map the representations in the lexical resource(s) onto those required for the task of interest
- ★ **presupposes:** (lexical semantic) language resource(s)

Learning Tasks

- Learning tasks we will look at are:
 - ★ countability learning
 - ★ lexical relation discovery
 - ★ detection of alternations
 - ★ modelling of semantic compositionality
 - ★ compound noun interpretation

- Other learning tasks we could have looked at include:
 - ★ word sense discrimination (Schütze, 1998)
 - ★ word sense disambiguation (Agirre and Edmonds, 2006)
 - ★ semantic role labelling (Carreras and Màrques, 2004)
 - ★ ontology learning (Buitelaar et al., 2005)
 - ★ learning selectional preferences (Resnik, 1993)
 - ⋮

Countability Learning

Countability Classes

- **countable:** *book, button, person (one book, two books)*
- **uncountable:** *equipment, gold, wood (*one equipment, much equipment, *two equipments)*
- **plural only:** *clothes, manners, outskirts (*one clothes, clothes horse)*
- **bipartite:** *glasses, scissors, trousers (*one scissors, scissor kick, pair of scissors)*

Difficulties in learning countability

- Multi-classification (15 possible classes to consider)
- Boundary between motivated countabilities and conversions (e.g. *chicken* vs. *elephant* vs. *dog*)
- Sense and frequency effects (e.g. *information*)
- Difficulties caused by MWEs (e.g. *cat's cradle*)

Learning countability

- **Observation 1:** countability is to some degree deterministic given the semantics of a word:

dog, pooch, canine, mongrel, ...

gold, silver, copper, bronze, ...

BUT *suitcases vs. luggage, leaves vs. foliage, etc.*

Methodology

- Take an existing ontology and determine the default countability for each synset (semantic class)
- Test how reliably defaults predict the countability of members of each synset
- Base experimentation on **WordNet 1.5**
- See also Bond and Vatikiotis-Bateson, 2002; O'Hara et al., 2003

Results

<i>Class</i>	<i>Feature type</i>	<i>Precision</i>	<i>Recall</i>	$F_{\beta=1}$
Countable	Baseline	.750	1.000	.857
	WordNet	.803	.805	.804
Uncountable	Baseline	.462	1.000	.632
	WordNet	.740	.420	.536

Learning Countability

- **Observation 2:** countability is to some degree deterministic given the form/morphology of a word:

shareware, ovenware, vapourware, cloudware, ...

information, vegetation, materialisation, ...

BUT *information vs. nation, etc.*

Methodology

- Model word “morphology” according to:
 - ★ character N-grams (1–6)
 $smile \rightarrow s, m, \dots, sm, \dots, smi, \dots, smil, \dots$
 - ★ syllable N-grams (list + `Lingua::EN::Hyphenate`)
 $hyphenation \rightarrow hyp\text{-}en\text{-}a\text{-}tion$
 - ★ derivational morphology (catvar)
 $smile \rightarrow V^{\pm\phi} smile$
 $smile \rightarrow N^{[+r]} smiler$
 $smile \rightarrow Adj^{[-e+ing]} smiling$

Results

<i>Class</i>	<i>Feature type</i>	<i>Precision</i>	<i>Recall</i>	$F_{\beta=1}$
Countable	Baseline	.750	1.000	.857
	Char n-gram	.826	.965	.891
	Syllable	.826	.976	.895
	Derivation	.807	.984	.887
	All	.827	.968	.892
Uncountable	Baseline	.462	1.000	.632
	Char n-gram	.778	.563	.653
	Syllable	.896	.446	.595
	Derivation	.898	.381	.535
	All	.768	.558	.647

Learning Countability

- **Observation 3:** the countability properties of a noun type are reflected in its corpus token occurrences:

*... Cezanne snarling like a **dog** and then ...*

*... doing an impression of a rabid **dog**.*

*... with a pack of **dogs** running beside them.*

*Amnesty International has received **information** ...*

*Recent **information** from former detainees ...*

*... researchers often uncover **information** ...*

Corpus-based Countability Classification

- Identify lexical and/or constructional features associated with each countability class
- Determine the relative corpus occurrence of the features for each noun
- Use the noun feature vectors to classify the noun as a member of each of the countability classes, training from gold-standard countability data

Feature Value Extraction

- POS tagging and templates
 - ★ extract features with regexp-base templates
- Full text chunking
 - ★ conservative inter-chunk attachment disambiguation
- Robust parsing (RASP)
- Combine by averaging over feature values from three systems

Example Feature Clusters

Head noun number:^{1D} target noun number as head of NP (e.g. *a shaggy dog* = SINGULAR)

Subject–verb agreement:^{2D} target noun number as subject vs. verb number agreement (e.g. *the dog barks* = $\langle \underline{\text{SINGULAR}}, \underline{\text{SINGULAR}} \rangle$)

Coordinate noun number:^{2D} target noun number vs. the number of the head nouns of conjuncts (e.g. *dogs and mud* = $\langle \underline{\text{PLURAL}}, \underline{\text{SINGULAR}} \rangle$)

Occurrence in PPs:^{2D} the presence or absence of a

determiner (\pm DET) in singular head complement of PP (e.g. *per dog* = $\langle \textit{per}, \textit{-DET} \rangle$).

Pronoun co-occurrence:^{2D} what pronouns occur in the same sentence as singular and plural instances (e.g. *The dog ate its dinner* = $\langle \textit{its}, \textit{SINGULAR} \rangle$). Approximation of pronoun co-indexation.

Singular determiners:^{1D} singular-selecting determiners (e.g. *a dog* = *a*). Two types: countable (e.g. *another, each*), uncountable (e.g. *much, little*).

Feature Values

$$1D \quad \text{corpfreq}(f_s, w) = \frac{\text{freq}(f_s|w)}{\text{freq}(*)} \quad (1)$$

$$\text{wordfreq}(f_s, w) = \frac{\text{freq}(f_s|w)}{\text{freq}(w)} \quad (2)$$

$$\text{featfreq}(f_s, w) = \frac{\text{freq}(f_s|w)}{\sum_i \text{freq}(f_i|w)} \quad (3)$$

$$2D \quad \text{featdimfreq}_1(f_{s,t}, w) = \frac{\text{freq}(f_{s,t}|w)}{\sum_i \text{freq}(f_{i,t}|w)} \quad (4)$$

$$\text{featdimfreq}_2(f_{s,t}, w) = \frac{\text{freq}(f_{s,t}|w)}{\sum_j \text{freq}(f_{s,j}|w)} \quad (5)$$

Results

<i>Class</i>	<i>Feature type</i>	<i>Precision</i>	<i>Recall</i>	$F_{\beta=1}$
Countable	Baseline	.750	1.000	.857
	POS	.927	.947	.937
	Chunk	.924	.944	.934
	RASP	.760	.776	.768
	All	.930	.957	.943
Uncountable	Baseline	.462	1.000	.632
	POS	.962	.662	.785
	Chunk	.960	.628	.759
	RASP	.513	.338	.407
	All	.981	.667	.794

Summary

- Three orthogonal approaches proposed for learning countabilities (lexical semantic, morphological, corpus-based)
- Corpus-based methods > morphological > lexical semantic
- A simple POS tagger gets us a surprising distance with corpus-based methods

Other Approaches

- Crosslingual countability learning (English→Dutch):
 - ★ learn for English and translate into Dutch
 - ★ learn via lexical similarity (as defined by EuroWordNet)



Lexical Relation Discovery

Lexical Relations

- Synonymy
- Hyponymy/hypernymy (sub/superordinate)
- Antonymy (opposite)
- Meronymy/holonymy (part-whole)
- Troponymy/hypernymy (cf. *walk* vs. *lollop*)
- ⋮

Standard Approach to Lexical Relation Discovery

- Lexico-syntactic patterns and corpus data, e.g. synonymy:

N_1 (*such as|including*) (N_2, N_3, \dots (*and|or*)) N_n
 N_1, N_2, \dots *and other* N_n

Joys of Lexical Relation Discovery

- Certain constructions indicative of lexical relations, but:
 - ★ what is the full extent of such constructions?
 - ★ how to use statistics of occurrence in such constructions?

Case Study: Qualia Structure

- Axiomatic description of noun semantics, made up of:
 - ★ **formal role:** the conceptual superclass of the noun
 $\text{formal}(\textit{book}) = \textit{publication}$
 - ★ **constitutive role:** internal constitution of entity
 $\text{constitutive}(\textit{book}) = \textit{page}, \textit{cover}, \textit{spine}, \dots$
 - ★ **telic role:** the typical function of the entity
 $\text{telic}(\textit{book}) = \textit{read}, \dots$
 - ★ **agentive role:** the origin of the entity
 $\text{agentive}(\textit{book}) = \textit{write}, \dots$

Template Set for Telic Role

Template

N (*BE*|\(\phi\)) (*worth*|\(*deserving*|\(*meriting*\)) (V[+ing]|V[+nom])

N *BE* *worthy of* V[+nom]

N (*deserves*|\(*merits*\)) V[+nom]

Adverb-V[+en] N

Adverb V[+en] N

N BE Adverb-V[ed]

V[+ing] Noun

N *to* V

Example Output (Telic)

Template		MaxEnt		Gold-standard	
publish	0.157	dedicate	1.084	write	10.0
write	0.102	publish	0.898	publish	8.0
read	0.019	compile	0.651	compile	8.0
call	0.015	dispose	0.605	print	7.5
dedicate	0.011	write	0.438	make	7.5
print	0.008	browse	0.408	start	7.0
keep	0.007	borrow	0.399	design	7.0
compile	0.006	print	0.386	translate	6.0

Extensions

- Use of web data (optionally with web data)
- Combined with filtering, such as:
 - ★ syntactic pre-/post-processing
 - ★ graph-based analysis
 - ★ statistical filtering
- Distributional similarity (synonymy)

Summary

- Lexical relation detection relatively well understood, with (specialised) methods proposed for a wide range of relation types
- General strategy: lexico-syntactic patterns (++)

Detection of Alternations

Diathesis Alternations

- A **diathesis alternation** is a regular variation in the argument structure of a verb
- Examples:
 - ★ Causative/inchoative alternation:
Kim broke the window \leftrightarrow *The window broke*
 - ★ Middle construction alternation:
Kim cut the bread \leftrightarrow *The bread cut easily*
- Applications in verb clustering

Detection of Alternations

- Alternations can be learnt through:
 - ★ subcategorisation frame acquisition
 - ★ (and optionally) selectional preference learning

Subcategorisation Frames

- A subcategorisation (subcat) frame is a statement of what types of arguments a verb ... takes as objects, infinitives, *that*-clauses, participial clauses and subcategorised PPs:

John wants Mary to be happy

John hopes that Mary is happy

**John wants that Mary is happy*

**John hopes Mary to be happy*

Subcat Acquisition a lá Brent, 1993

1. Identify verb tokens
2. For each verb type, use high-precision lexico-syntactic patterns to identify evidence for 6 different subcat frames
3. Use a statistical filter to remove noise in the extracted subcat data

Lexico-syntactic patterns

- Based on closed-class words (pronouns, determiners, complementisers, auxiliaries, punctuation)
- NPs captured in the form of pronouns or sequences of capitalised words
- VPs based on auxiliaries and the verbs learned in step 1

Statistical filtering (1)

- Assumption that the probability of false evidence for a given subcat frame S (e.g. transitive) occurring is equal for all verbs incompatible with S (e.g. *snore*, *put*, *say*, ...)
- NOTE: probability of false evidence (π_{-S}) constant for a given S but varies across different subcat frames
- Null hypothesis: the verb does not belong to subcat class S , i.e. it is $-S$

Statistical filtering (2)

- **Binomial test:** the probability of an event with probability p occurring exactly m out of n times is given by

$$P(m, n, p) = \frac{n!}{m!(n-m)!} p^m (1-p)^{n-m}$$

- The probability of the event occurring m or more times out of n is given by

$$P(m+, n, p) = \sum_{i=m}^n P(i, n, p)$$

$\frac{m}{n}$	$P(m, n, p = 0.1)$	$P(m+, n, p = 0.1)$
$\frac{0}{10}$	0.349	1.000
$\frac{1}{10}$	0.387	0.651
$\frac{2}{10}$	0.194	0.264
$\frac{3}{10}$	0.057	0.070
$\frac{4}{10}$	0.011	0.013
$\frac{5}{10}$	0.001	0.002
$\frac{6}{10}$	0.000	0.000
$\frac{7}{10}$	0.000	0.000
$\frac{8}{10}$	0.000	0.000
$\frac{9}{10}$	0.000	0.000
$\frac{10}{10}$	0.000	0.000

Statistical filtering (3)

- Given n and p ($= \pi_{-S}$), we can apply a threshold θ to determine m such that verbs which occur with subcat frame S at least m times can be classified as $+S$ with $(1 - \theta)$ confidence
- In practice we don't know π_{-S} for each subcat frame S
SOLUTION: set θ and n , and estimate p based on the histogram distribution around each m ; select the p which best fits the binomial distribution

Shortcomings of the Brent approach

- Assumption of π_{-S} being equal for all verbs given a class S shown to be flawed due to verb detection method
- Applicability of method to low-frequency words
- Scalability of method to other subcat frames

An update on more recent research

- Greater coverage of subcat frames (up to 160)
- Simple frequency shown to be at least as effective as binomial test at filtering out noise
- Verb sense shown to interface closely with subcategorisation properties
- AND YET the Brent method still has remarkable currency to this day

Open questions

- How to deal with low-frequency occurrences of subcat frames
- How well do the proposed methods port to other word classes (adjectives, nouns, ...) and languages
- Challenges for subcat acquisition in pro-drop languages (e.g. Japanese)

Work on Levin-style Verb Classification

- Use alternations and general verbal features to classify verbs according to Levin (1993) classes
- Dodge the issue of alternation detection or subcat acquisition by relying on features which capture alternation effects only indirectly
- Supplement alternation-based features with various weak lexical semantic indicators

Summary

- Work on alternation detection based on lexico-syntactic patterns + statistical filtering (again)
- Mixed results to date, despite complexity of methods tested over the task

Modelling of Semantic Compositionality

Semantic Compositionality

- Compositionality = *degree to which the semantics of the parts of an MWE contribute towards those of the whole*
- Why care? To work out what MWEs are lexically semantically-marked that we need to have an independent account of
- Domain considerations: *monosodium glutamate* in chemistry vs. health domains

Approaches to Evaluation

- **Dictionary based:** binary evaluation, based on prediction that non-compositional MWEs will be lexically listed
- **Similarity based:** relative similarity of the parts to the whole (e.g. relative to WordNet)

$$\text{sim}(\textit{pig metal}, \textit{metal}) \gg \text{sim}(\textit{pig metal}, \textit{pig})$$

- **Entailment based:** binary evaluation, based on whether the whole “entails” the parts or not

Susan finished up her paper \models *Susan finished her paper*

- **Ranking based:** describe MWE compositionality by way of continuous/discrete scale of compositionality

$\text{comp}(\textit{put up}) \geq \text{comp}(\textit{eat up}) \geq \text{comp}(\textit{gun down})$

...

Exercise: Rate the Compositionality

<i>VPC</i>	<i>Compositionality</i>			
	<i>Dic</i>	<i>Ent(V)</i>	<i>Ent(P)</i>	<i>Rank</i>
<i>get downTRANS</i>	1			
<i>piss offTRANS</i>	0			
<i>pay offTRANS</i>	1			
<i>lift outTRANS</i>	0			
<i>roll backTRANS</i>	0			
<i>dig upTRANS</i>	1			
<i>lie downINTRANS</i>	1			
<i>wear onINTRANS</i>	1			
<i>chicken outINTRANS</i>	1			
<i>hand outTRANS</i>	1			

Compositionality via Substitution

- Use substitution as a test of compositionality:

red tape → *yellow* *tape*, *red* *cassette*

economic impact → *political* *impact*, *economic*
effect

System Resources

- POS-conditioned thesaurus (nouns, verbs, adjectives/adverbs)
 - ★ derived from dependency data (Minipar):
- Collocation data
 - ★ dependency tuples (H,R,M) with high log-likelihood ratio (H = head, R = relation, M = modifier)

Mutual Information and Compositionality

- Scaling up to 3 events A , B and C , where B and C are conditionally independent given A :

$$MI(A, B, C) = \log_2 \frac{P(A, B, C)}{P(B|A)P(C|A)P(A)}$$

$$MI(H, R, M) = \log_2 \frac{\begin{array}{|c|c|c|} \hline H & R & M \\ \hline * & * & * \\ \hline \end{array}}{\begin{array}{|c|c|c|} \hline H & R & * \\ \hline * & R & * \\ \hline \end{array} \begin{array}{|c|c|c|} \hline * & R & M \\ \hline * & R & * \\ \hline \end{array} \begin{array}{|c|c|c|} \hline * & R & * \\ \hline * & * & * \\ \hline \end{array}}$$

$$= \log_2 \frac{\begin{array}{|c|c|c|} \hline H & R & M \\ \hline * & R & * \\ \hline \end{array} \begin{array}{|c|c|c|} \hline * & R & * \\ \hline * & R & M \\ \hline \end{array}}{\begin{array}{|c|c|c|} \hline H & R & * \\ \hline * & R & M \\ \hline \end{array}}$$

Definition of Compositionality

- A phrase α is non-compositional iff there is no β s.t.:
 - (a) β can be produced by substitution of the components of α for any of 10 most-similar words, and
 - (b) there is an overlap between the 95% confidence interval of the MI values of α and β
- 10 most-similar words tested for each of H and M (R fixed)

Example 1: *spill (one's) guts*

- $(spill, V:comp1:N, gut)$:
 - ★ *spill*: *leak, pour, spew, ..., spray*
 - ★ *gut*: *intestine, instinct, foresight, ..., charisma*
- Check for each of $(leak, V:comp1:N, gut)$, $(spill, V:comp1:N, intestine)$, ... in the collocation database
- None found, so *spill (one's) guts* is non-compositional

Example 2: *red tape*

- $(\textit{tape}, \text{N:adj:N}, \textit{red})$:
 - ★ *tape*: *videotape, cassette, videocassette, ..., audio*
 - ★ *red*: *yellow, purple, pink, ..., shade*
- Find $(\textit{tape}, \text{N:adj:N}, \textit{yellow})$, $(\textit{tape}, \text{N:adj:N}, \textit{orange})$, $(\textit{tape}, \text{N:adj:N}, \textit{black})$ in the collocation database but with very different MI values
- *red tape* is non-compositional

MI Confidence Interval: the Z-test

- Possible to calculate the “true” MI of (H,R,M) according to the Z-test:

$$\bar{p} \pm z_N \sqrt{\frac{\bar{p}(1 - \bar{p})}{n}} = \frac{k}{n} \pm z_N \sqrt{\frac{\frac{k}{n}(1 - \frac{k}{n})}{n}} \approx \frac{k \pm z_N \sqrt{k}}{n}$$

where \bar{p} is the MLE of p , n is $|* * *|$, k is $|H R M|$, and z_N is a constant determined by the confidence level N , e.g. $z_{0.95} = 1.96$

Applying the Z-test

- Determine the “fit” between two MI values by calculating the Z-score interval for the putative non-compositional MWE and determining whether the MI of the second falls into that interval

Reflections

- Is substitution really a good test for non-compositionality?
 - ★ institutionalised phrases: *frying pan, salt and pepper, many thanks*
 - ★ productive MWEs: *call/phone/ring up*
- Look to alternative methods

Compositionality via Distributional Similarity

- Define similarity in terms of distributional similarity, i.e. assume that if an MWE is compositional, it will occur in the same lexical context as its parts

Verb-particle Constructions (VPCs)

- VPC = A verb plus one or more obligatory (prepositional) particles

Peter put the picture up

Susan finished up her paper

Philip gunned down the intruder

Barbara and Simon made out

- **Assumption:** VPCs are not always fully compositional or fully non-compositional, but rather populate a continuum between the two extremes

Distributional Similarity-based Methods

- **pair-wise**: the distributional similarity between the VPC and verb OR particle
- **overlap**: relative overlap between the top N neighbours of the VPC and its simplex verb
- **sameparticle**: the number of VPCs which select for the same particle as the given VPC amongst the top N neighbours of that VPC
- **sameparticle** – **simplex**: the value for **sameparticle**

minus the number of top N neighbours of the simplex verb which select for that same particle

- **simplexasneighbour:** does the simplex verb occur in the top 50 neighbours of the VPC?
- **rankofsimplex:** what is the rank of the simplex verb in the neighbours of the VPC?
- **overlapS:** the overlap of neighbours in the top N neighbours of the VPC and simplex verb, where VPC neighbours are converted to simplex verbs in the VPC case

Statistical Methods

- χ^2
- Log-likelihood ratio
- (Point-wise) mutual information
- Simple frequency of the VPC
- Simple frequency of the simplex verb

Resource-based Method

- Binary test for the occurrence of the VPC in:
 - ★ WordNet
 - ★ Alvey Tools (ANLT) VPC data
 - ★ Alvey Tools (ANLT) prepositional verb data

Summary of Results

- Promising results observed for detecting compositionality/ decomposability, but less so for determining the semantic contribution of individual words in an overall MWE
- What about MWEs where the simplex words don't occur with that same POS (e.g. *chicken out*)
- Effects of polysemy (e.g. *run down*, *run over*)

Compound Noun Interpretation

Compound Nominals and Compound Nominalisations

- **Compound nominal:** \bar{N} made up of two or more nouns, e.g.:

telephone box/booth, river bed, radar footprint, chest X-ray

- **Compound nominalisation:** subclass of compound nominals in which the head noun is deverbal, e.g.:

machine performance, museum construction

Interpreting Compound Nominalisations

- **Task:** binary classification of nominalisations as having a SUBJ or OBJ interpretation (ignore nominalisations such as *soccer competition* — i.e. constrain the space in such a way that interpretation is a well-defined task)
- **Assumption:** $P(rel|n_1, n_2) \approx P(rel|v_{n_2}, n_1)$
- **Problem:** getting accurate estimates of $P(rel|v_{n_2}, n_1)$

Basic Model

$$RA(rel, n_1, n_2) = \log_2 \frac{P(\text{OBJ}|n_1, n_2)}{P(\text{SUBJ}|n_1, n_2)}$$
$$P(rel|n_1, n_2) \approx \frac{f(v_{n_2}, rel, n_1)}{\sum_i f(v_{n_2}, rel_i, n_1)}$$

Observation

- Of 796 items in gold-standard nominalisation set, 47% not attested in BNC in either a verb-object or verb-subject relation
- How to get accurate estimates of $f(v_{n_2}, rel, n_1)$?
- **Answer:** smoothing based on the frequencies of observed verb-argument pairs

Smoothing

1. **Discounting:** redistribute probability from observed events to unobserved events
2. **Class-based smoothing:** word-to-class distributional similarity
3. **Distance-weighted averaging:** word-to-word distributional similarity

Discounting

- Katz's backing-off:

$$P(rel|n_1, n_2) = \begin{cases} \alpha \frac{f(v_{n_2}, rel, n_1)}{\sum_i f(v_{n_2}, rel_i, n_1)} & \text{if } f(v_{n_2}, rel, n_1) > 0 \\ \beta \frac{f(rel, n_1)}{f(n_1)} & \text{if } f(rel, n_1) > 0 \\ (1 - \alpha - \beta) \frac{f(rel)}{\sum_i f(rel_i)} & \text{otherwise} \end{cases}$$

- Estimate α and β by Good-Turing estimation

Class-based Smoothing

- Map observed verb-argument tuples onto the WordNet/Roget classes of the noun, distributing equally across all synsets the noun is categorised as belonging to
- Calculate $f(v_{n_2}, rel, n_1)$ by averaging across the classes that n_1 occurs in
- Closed world assumption for nouns

Distance-weighted Averaging

- Use **confusion probability** or **Jensen-Shannon divergence** to estimate the distributional similarity between v_{n_2} and each verb w_1' , and estimate $f(v_{n_2}, rel, n_1)$ according to:

$$f_s(v_{n_2}, rel, n_1) = \sum_{w_1'} \text{sim}(v_{n_2}, w_1') f(w_1', rel, n_1)$$

Results

- Confusion probability and WordNet-based smoothing tend to do the best overall
- Good results for system classification, combined with context modelling in the form of the right word context of the compound nominal (85% test accuracy)

Reflections

- Interesting task-oriented smoothing experiment
- What to do with non-SUBJ/OBJ nominalisations?
- What to do with prepositional verbs, verb particles?
- Influence of pragmatics on interpretation

Interpreting Compound Nouns (1): Rosario and Hearst, 2001

- **Task:** interpretation of (2-word) compound nominals within the biomedical domain
- **Method:** use lexical or conceptual knowledge about the component nouns to interpret the whole (context-independent)
- **Resource:** MeSH (biomedical thesaurus)

Semantic Roles

- Compound nominals interpreted via 18 (out of 38) relations:
 - ★ more specific than case roles, and less specific than IE template fillers
 - ★ customised to the biomedical domain (e.g. *polio survivors* → PERSON-AFFLICTED)
 - ★ thresholded for frequency
 - ★ overlapping (multiclass classification possible: *cell growth* → ACTIVITY + CHANGE)

Method

- **Class-based model:** describe NN according to the concatenation of the MeSH representations of N_1 and N_2 (up to level N)
- **Lexical model:** describe NN by its component words (*closed-word assumption*)
- **Learner:** neural network (feed-forward network with one hidden layer)

Results

- Over closed data, the lexical and class-based models perform equivalently ($\approx 60\%$)
- Over open data, the class-based model performs better (unsurprisingly)
- Suggestion that N_2 has a stronger impact on the interpretation than N_1

Reflections

- Question of interpretation system sidestepped to some degree by picking a technical domain
- Multiclassification awkward effect, which raises questions about the appropriateness of the interpretation system
- Possibility for a hybrid approach combining the class-based and lexical models?
- No systematic treatment of lexicalised nominals

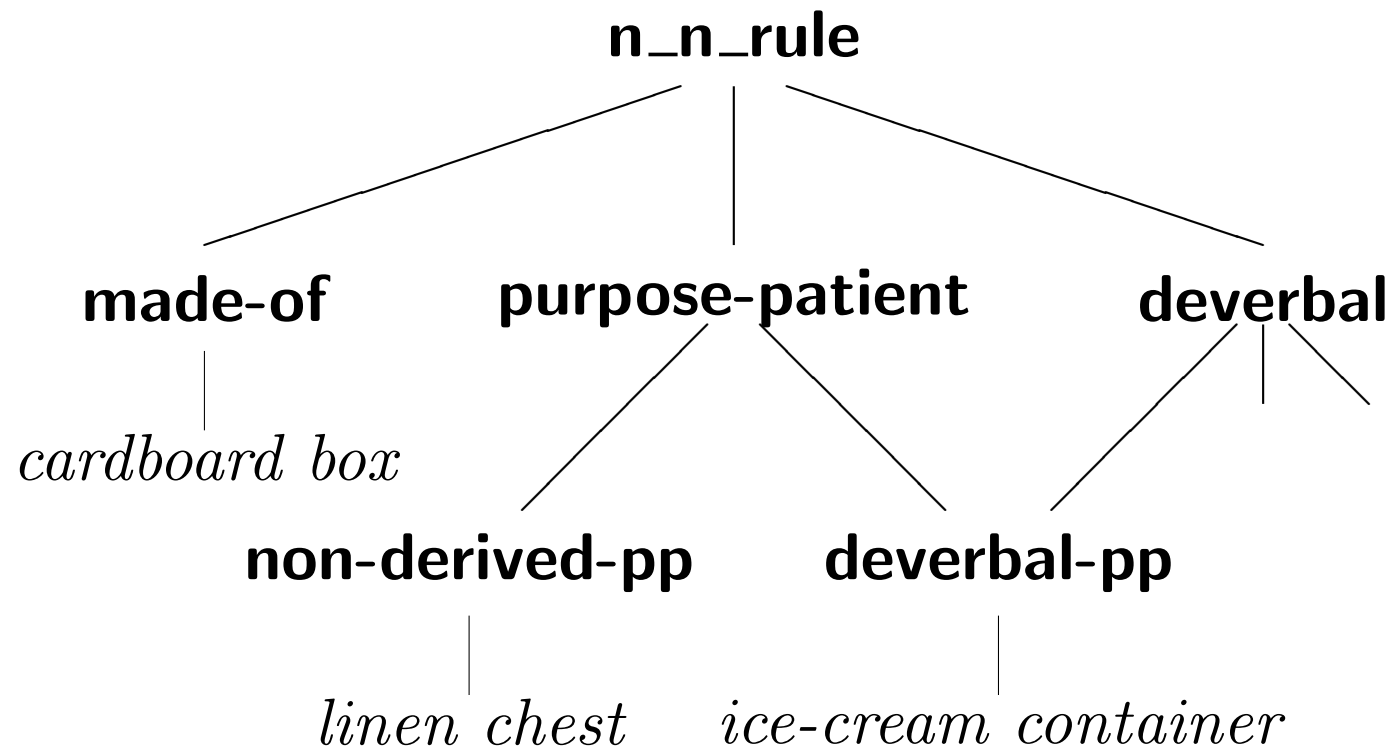
Interpreting Compound Nouns (2): Copestake and Lascarides, 1997

- **Basic method:**

1. use the grammar/lexicon to delimit the range of potential interpretations of a given NN
2. use “productivity” probabilities to rank the individual interpretations
3. use pragmatics to filter out interpretations which produce discourse incoherence within a given context

- Possible to derive non-standard interpretations for a compound nominal (e.g. *garbage man*)

Semantic Hierarchy



Estimating Productivity

- Estimate productivity based on the number of attested forms of a given schemata:

$$\text{Prod}(cmp_schema) = \frac{M + 1}{N}$$

where N is the number of pairs of senses which match *cmp_schema* and M is the number of attested forms

- Cf. substitution tests for collocations/compositionality

Applying the Productivity Estimates

- Interpretations for *cotton bag* based on analysis of fabric/container NNs in the BNC (based on WordNet):

MADE-OF $P = 0.84$

PURPOSE-PATIENT $P = 0.14$

GENERAL-NN $P = 0.02$

- Prediction that the default interpretation for *cotton bag* is MADE-OF

Interface with Pragmatics

- Model pragmatics with SDRT and world knowledge with DICE
- Use SDRT and DICE to filter out interpretations that produce discourse incoherence:
 - a. *Mary sorted her clothes into various bags made from plastic*
 - b. *She put her skirt into the cotton bag*

Reflections

- Rare instance of method which provides direct handling of the lexicon-pragmatics interface
- Implausible interpretations supported explicitly, but dispreferred
- Difficulties in collecting productivity statistics
- Question of real-world applicability of SDRT/pragmatic reasoning

Overall summary

- Learning lexical semantic representations encompasses many sub-tasks with different peculiarities/anomalies
- Lexico-syntactic patterns common theme to most approaches
- Constant challenges in differentiating between noise and positive attestations of a given phenomenon

References

- Agirre, E. and Edmonds, P., editors (2006). *Word Sense Disambiguation: Algorithms and Applications*. Springer, Dordrecht, Netherlands.
- Baldwin, T. (2005). Bootstrapping deep lexical resources: Resources for courses. In *Proc. of the ACL-SIGLEX 2005 Workshop on Deep Lexical Acquisition*, pages 67–76, Ann Arbor, USA.
- Baldwin, T. and Bond, F. (2003a). Learning the countability of English nouns from corpus data. In *Proc. of the 41st Annual Meeting of the ACL*, pages 463–70, Sapporo, Japan.
- Baldwin, T. and Bond, F. (2003b). A plethora of methods for learning English countability. In *Proc. of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP 2003)*, pages 73–80, Sapporo, Japan.
- Baldwin, T. and van der Beek, L. (2003). The ins and outs of Dutch noun countability classification. In *Proceedings of the 2003 Australasian Language Technology Workshop (ALTW2003)*, pages 33–40, Melbourne, Australia.
- Bannard, C., Baldwin, T., and Lascarides, A. (2003). A statistical approach to the semantics of verb-particles. In *Proc. of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 65–72, Sapporo, Japan.
- Bond, F. and Vatikiotis-Bateson, C. (2002). Using an ontology to determine English countability. In *Proc. of the 19th International Conference on Computational Linguistics (COLING 2002)*, Taipei, Taiwan.
- Brent, M. R. (1993). From grammar to lexicon: Unsupervised learning of lexical syntax. *Computational Linguistics*, 19(2):243–62.
- Buitelaar, P., Cimiano, P., and Magnini, B., editors (2005). *Ontology Learning from Text: Methods, Evaluation and Applications*. IOS Press, Amsterdam, The Netherlands.

- Carreras, X. and Màrques, L. (2004). Introduction to the conll-2004 shared task: Semantic role labeling. In *Proc. of the 8th Conference on Natural Language Learning (CoNLL-2004)*, pages 89–97. Boston, USA.
- Carroll, J. and Fang, A. (2004). The automatic acquisition of verb subcategorisations and their impact on the performance of an HPSG parser. In *Proc. of the First International Joint Conference on Natural Language Processing (IJCNLP-04)*, pages 107–14, Sanya City, China.
- Cimiano, P. and Wenderoth, J. (2005). Automatically learning qualia structures from the Web. In *Proc. of the ACL-SIGLEX 2005 Workshop on Deep Lexical Acquisition*, pages 28–37, Ann Arbor, USA.
- Copestake, A. and Lascarides, A. (1997). Integrating symbolic and statistical representations: The lexicon pragmatics interface. In *Proc. of the 35th Annual Meeting of the ACL and 8th Conference of the EACL (ACL-EACL'97)*, pages 136–43, Madrid, Spain.
- Cruse, D. A. (1986). *Lexical Semantics*. Cambridge University Press, Cambridge, UK.
- Fellbaum, C., editor (1998). *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, USA.
- Girju, R., Badulescu, A., and Moldovan, D. (2003). Learning semantic constraints for the automatic discovery of part-whole relations. In *Proc. of the 3rd International Conference on Human Language Technology Research and 4th Annual Meeting of the NAACL (HLT-NAACL 2003)*, Edmonton, Canada.
- Hearst, M. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proc. of the 14th International Conference on Computational Linguistics (COLING '92)*, Nantes, France.
- Joanis, E. and Stevenson, S. (2003). A general feature space for automatic verb classification. In *Proc. of the 10th Conference of the EACL (EACL 2003)*, pages 163–70, Budapest, Hungary.
- Korhonen, A. (2002). *Subcategorization Acquisition*. PhD thesis, University of Cambridge.
- Korhonen, A. and Briscoe, T. (2004). Extended lexical-semantic classification of english verbs. In *Proc. of the HLT-NAACL 2004 Workshop on Computational Lexical Semantics*, Boston, USA.

- Korhonen, A., Krymolowski, Y., and Marx, Z. (2003). Clustering polysemic subcategorization frame distributions semantically. In *Proc. of the 41st Annual Meeting of the ACL*, pages 64–71, Sapporo, Japan.
- Lapata, M. (2002). The disambiguation of nominalizations. *Computational Linguistics*, 28(3):357–88.
- Levin, B. (1993). *English Verb Classes and Alterations*. University of Chicago Press, Chicago, USA.
- Lin, D. (1999). Automatic identification of non-compositional phrases. In *Proc. of the 37th Annual Meeting of the ACL*, pages 317–24, College Park, USA.
- Manning, C. D. (1993). Automatic acquisition of a large subcategorization dictionary from corpora. In *Proc. of the 31st Annual Meeting of the ACL*, pages 235–42.
- McCarthy, D. (2001). *Lexical Acquisition at the Syntax-Semantics Interface: Diathesis Alternations, Subcategorization Frames and Selectional Preferences*. PhD thesis, University of Sussex, Sussex, UK.
- McCarthy, D., Koeling, R., Weeds, J., and Carroll, J. (2004). Finding predominant senses in untagged text. In *Proc. of the 42nd Annual Meeting of the ACL*, pages 280–7, Barcelona, Spain.
- Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. (1990). Introduction to WordNet: an on-line lexical database. *International Journal of Lexicography*, 3(4):235–44.
- O’Hara, T., Salay, N., Witbrock, M., Schneider, D., Aldag, B., Bertolo, S., Panton, K., Lehmann, F., Smith, M., Baxter, D., Curtis, J., and Wagner, P. (2003). Inducing criteria for mass noun lexical mappings using the Cyc KB and its extension to WordNet. In *Proc. of the Fifth International Workshop on Computational Semantics (IWCS-5)*, Tilburg, Netherlands.
- Resnik, P. (1993). *Selection and Information: A Class-Based Approach to Lexical Relationships*. PhD thesis, University of Pennsylvania.
- Rosario, B. and Hearst, M. (2001). Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy. In *Proc. of the 6th Conference on Empirical Methods in Natural Language Processing (EMNLP 2001)*, Pittsburgh, USA.

- Schulte im Walde, S. (2003). *Experiments on the Automatic Induction of German Semantic Verb Classes*. PhD thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart. Published as AIMS Report 9(2).
- Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.
- Schwartz, L. O. (2002). Corpus-based acquisition of head noun countability features. Master's thesis, Cambridge University, Cambridge, UK.
- van der Beek, L. and Baldwin, T. (2004). Crosslingual countability classification with EuroWordNet. In *Papers from the 14th Meeting of Computational Linguistics in the Netherlands*, pages 141–55, Antwerp, Belgium. Antwerp Papers in Linguistics.
- Widdows, D. (2005). *Geometry and Meaning*. CSLI Publications, Stanford, USA.
- Yamada, I. and Baldwin, T. (2004). Automatic discovery of telic and agentive roles from corpus data. In *Proc. of the 18th Pacific Asia Conference on Language, Information and Computation (PACLIC 18)*, pages 115–26, Tokyo, Japan.