Nominalisations and Compound Nominals

Timothy Baldwin and Dominic Widdows
INTRODUCTION
Compound Nominals and Nominalisations

- **Compound nominal**: N’ made up of two or more nouns, e.g.:
  
  `telephone box/booth, river bed, radar footprint, chest X-ray`

- **Nominalisation**: subclass of compound nominals in which the head noun is deverbal, e.g.:
  
  `machine performance, museum construction, family worker, student education, farm agreement`
Compound Nominals and NLP

• Compound nominals generally processed in three steps:

1. **Identification** of compound nominals in some corpus
   
   *A film interpretation of the book which satirises black assimilation into white society.*

2. **Syntactic analysis** of the structure
   
   *engine oil filter* → [[engine oil] filter]

3. **Interpretation** of the semantics
   
   *film interpretation* → OBJ

• We will focus exclusively on interpretation (Step 3)
Interpretation

- Compound nominals are largely unrestricted semantically:
  - diesel truck/oil/tanker, phone book, cloud bus, apple juice seat

- Nominalisations tend to occur with subject or object interpretation:
  - machine performance, museum construction, student education BUT also soccer competition
Properties of Compound Nominals

- Highly productive ($\approx 300K$ NN types in BNC)
- Very frequent ($> 1M$ NN tokens in BNC)
- Very skewed in frequency ($\approx 60\%$ of NN types in BNC occur once)
- Interpretation implicit (in English) and highly variable

cf. Italian: coltello da pane “bread knife”, porta a vetri “glass door”, succo di limone “lemon juice”
Open Questions

• Is there a definitive categorical system of compound nominal interpretation types? (splitters and lumpers)

• Can any one system work for all domains and compound nominal types?

• What systems of interpretation work in different domains?

• To what degree is interpretation required? (e.g. machine translation)
The Disambiguation of Nominalisations

Maria Lapata

Computational Linguistics 28(3)
Basic Outline

- **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:** \( P(\text{rel}|n_1, n_2) \approx P(\text{rel}|v_{n_2}, n_1) \)

- **Problem:** getting accurate estimates of \( P(\text{rel}|v_{n_2}, n_1) \)
Basic Model

\[ RA(\text{rel}, n_1, n_2) \approx \log_2 \frac{P(\text{OBJ}|n_1, n_2)}{P(\text{SUBJ}|n_1, n_2)} \]

\[ P(\text{rel}|n_1, n_2) \approx \frac{f(v_{n_2}, \text{rel}, n_1)}{\sum_i f(v_{n_2}, \text{rel}_i, n_1)} \]
**Resources**

- Derive frequency estimates from the BNC
- Estimate $f(v_{n_2}, rel, n_1)$ from output of dependency parser (Cass)
- Determine base verb form of nominalisation based on NOMLEX and CELEX
- Hand-annotate/filter 796 nominalisations extracted from BNC
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:

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

- Estimate \(\alpha\) and \(\beta\) 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:

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

\[ P_C(w_1 | w_1') = \sum_{rel, w_2} P(w_1 | rel, w_2) P(rel, w_2 | w_1') \]

\[ = \sum_{rel, w_2} \frac{f(w_1, rel, w_2)}{f(rel, w_2)} \frac{f(w_1', rel, w_2)}{f(w_1')} \]
Jensen-Shannon Divergence

\[
J(w_1, w_1') = \frac{1}{2} \left[ D\left( m(w_1) \| n(w_1, w_1') \right) + D\left( m(w_2) \| n(w_1, w_1') \right) \right]
\]

\[
W_J(w_1, w_1') = 10^{-\beta J(w_1, w_1')}
\]

where

\[
m(w) = P(\text{rel}, w_2 | w)
\]

\[
n(w_1, w_1') = \frac{1}{2} \left( m(w_1) + m(w_1') \right)
\]

\[
D\left( m(w_1) \| n(w_1, w_1') \right) = 
\]

\[
\sum_{\text{rel}, w_2} P(\text{rel}, w_2 | w_1) \log \frac{P(\text{rel}, w_2 | w_1)}{\frac{1}{2} \left( P(\text{rel}, w_2 | w_1) + P(\text{rel}, w_2 | w_1') \right)}
\]
Evaluation

- Annotator agreement = 89.7%
- Take 2,000 nearest neighbour verbs \( w_1' \) distance-weighted averaging methods, \( \beta = 5 \)
- Baseline accuracy of 61.5% (OBJ interpretation)
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
Classifying the Semantic Relations in Noun Compounds via a Domain-Specific Lexical Hierarchy

Barbara Rosario and Marti Hearst
EMNLP ’01
Basic Outline

- **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
Integrating Symbolic and Statistical Representations: The Lexicon-pragmatics Interface

Ann Copestake and Alex Lascarides
ACL/EACL ’97
Basic Outline

• 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

n_n_rule

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of

made-of
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
Overview of Course
Course Summary

- Meandering course through the wilds of computational word learning, covering subtasks of morphology, syntax, semantics and pragmatics

- Focus on methods rather than results

- Different methods pick up on different conceptualisations of same task
Looking back over the Road Covered

- **NLP technologies covered:** tagging, chunking, dependency parsing, lemmatisation, ...

- **Methods covered:** LSA (SVD), graph-based clustering, MDL, EM algorithm, mutual information, information-based similarity, ...

- **Learner paradigms touched upon:** transformation-based learning, decision lists, decision trees, memory-based learners (ID1), ...
• **Resources used:** WordNet, MeSH, UMLS, COMLEX, BNC, WSJ, Brown, LDOCE, EVCA, ALT-J/E, ERG, ...

• **Systems used:** Brill tagger, RASP, TiMBL, C4.5, fnTBL, ...
Lessons Learned

- There’s more than one way to tackle a given task
- NLP papers (hopefully) aren’t as intimidating as they first seem
- NLP is only really scratching the surface of the word learning tasks it has tackled to date, and there are still many new tasks out there waiting to be discovered
Where to from Here?

- NLP-related courses on offer at Stanford:

- NLP Reading Group

- Various projects and individuals at Stanford involved in NLP who would love to hear from you if you want to get your hands dirty!
Thanks for being part of a first-time course that has been fun to teach!

Tim and Dominic

1 December, 2003