Compound interpretation as a challenge for computational semantics

Díarmaíd Ó Séaghdha

ComAComA, Dublin
24 August 2014
Noun-noun compounding is very common in many languages. We can make new words out of old. Expanding vocabulary → lots of OOV problems! Compounding compresses information about semantic relations. Decompressing this information ("interpretation") is a non-trivial task. In this talk I focus on relational understanding.
The hut is located in the mountains
The hut is constructed out of timber
The camp produces timber
The hut is located in the mountains  LOCATION
The hut is constructed out of timber  MATERIAL
The camp produces timber  LOCATION/PRODUCER
The hut is located in the mountains  
The hut is constructed out of timber  
The camp produces timber  

We slept in a mountain hut  
We slept in a timber hut  
We slept in a timber camp
The hut is located in the mountains  LOCATION
The hut is constructed out of timber  MATERIAL
The camp produces timber  LOCATION/PRODUCER

We slept in a mountain hut  ??
We slept in a timber hut
We slept in a timber camp
Why compounds?

- Special but very frequent case of information extraction
- In order to interpret compounds, a system must be able to deal with:
  - Lexical semantics
  - Relational semantics
  - Implicit information
  - World knowledge
  - Handling sparsity
- Compound interpretation is an excellent testbed for computational semantics.
Thoughts and open questions
A brief history of compound semantics

- **500 BCE**: Sanskrit grammarians
- **1900**: Linguistics
- **1970**: NLP
Open questions

- ... almost all questions are still open!

- Some questions that I am interested in:
  - What are useful representations for compound semantics?
  - What are learnable representations for compound semantics?
  - Should we use representations that are not specific to compounds?
  - What are the applications of compound interpretation?
    - Paraphrasing/lexical expansion (for MT, search,...)
    - Machine reading/natural language understanding

- Many representation options, some more popular than others

- All have pros and cons
The lexical analysis

- **Idea:** Treat compounds as if they were words.
- **Frequent/idiomatic compounds (e.g., in WordNet)**
- **Pro:** Flexible
- **Con:** Productivity

![Graph showing the distribution of corpus frequency vs. number of types.](image)
The “pro-verb” analysis

➤ **Idea:** Underspecified single relation for all compounds

➤ Adequate when parsing to logical form or e.g. Minimal Recursion Semantics:

  - `car tyre`    `compound nn rel(car,tyre)`
  - `history book` `compound nn rel(history,book)`

➤ **Pro:** Easy to integrate with parsing/structured prediction

➤ **Con:** Not very expressive!
The inventory analysis

- **Idea:** Select a relation label from a (small) set of candidates

  - car tyre: **Part-Whole**
  - mountain hut: **Location**
  - cheese knife: **Purpose**
  - headache pill: **Purpose**

- Earliest, most common approach [Su, 1969; Russell, 1972; Nastase and Szpakowicz, 2003; Girju et al., 2005; Tratz and Hovy, 2010]

- Some relation extraction datasets span compounds and other constructions [Hendrickx et al., 2010]

- **Pro:** Learnable as multiclass classification; annotation is feasible

- **Con:** Conflates subtleties (*sleeping pill* vs *headache pill*); requires annotated training data
The vector analysis

- **Idea:** Represent a compound by composing vectors for each constituent to produce a new vector

- Lots of work on vector composition; some work on noun-noun composition [Mitchell and Lapata, 2010; Reddy et al., 2011; Ó Séaghdha and Korhonen, 2014]

- **Pro:** Learnable from unlabelled data

- **Con:** Difficult to interpret
The paraphrase analysis

- **Idea:** Represent the implicit relation(s) with a distribution over explicit paraphrases.
- Allowable paraphrases can use prepositions [Lauer, 1995], verbs [Nakov, 2008; Butnariu et al., 2010], free paraphrases [Hendrickx et al., 2013]

  - virus that causes flu 38
  - virus that spreads flu 13
  - virus that creates flu 6
  - virus that gives flu 5
  - ...
  - virus that is made up of flu 1
  - virus that is observed in flu 1

- Suitable for similarity, data expansion
- **Pro:** Learnable from unannotated text
- **Con:** Paraphrases can be ambiguous/synonymous
The frame analysis

- We could recover implicit relational structure in terms of FrameNet-like frames:

  - **cheese knife**
    - Cutting(f) ∧ Instrument(f,knife) ∧ Item(f,cheese)
  - **kitchen knife**
    - Cutting(f) ∧ Instrument(f,knife) ∧ Place(f,kitchen)
  - **student demonstration**
    - Protest(f) ∧ Protestor(f,student)
  - **headache pill**
    - Cure(f) ∧ Affliction(f,headache) ∧ Medication(f,pill)

- Connection to cognitive/frame semantics [Ryder, 1994; Coulson, 2001]
- SRL usually assumes explicit verbal predicates or nominalisations
- **Pro**: More structured than paraphrases, more fine-grained than traditional relations
- **Con**: Annotation
The first part of this talk has no conclusion!
Experiments with a multi-granularity relation inventory
<table>
<thead>
<tr>
<th>Relation</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>BE</td>
<td>guide dog</td>
</tr>
<tr>
<td>HAVE</td>
<td>car tyre</td>
</tr>
<tr>
<td>IN</td>
<td>air disaster</td>
</tr>
<tr>
<td>ACTOR</td>
<td>committee discussion</td>
</tr>
<tr>
<td>INST</td>
<td>air filter</td>
</tr>
<tr>
<td>ABOUT</td>
<td>history book</td>
</tr>
</tbody>
</table>
Relation Inventory

- **COARSE**
  - BE
  - HAVE
  - IN
  - ACTOR
  - INST
  - ABOUT

- **DIRECTED**
  - HAVE \(_1\)
    - car tyre
  - HAVE \(_2\)
    - hotel owner
Relation Inventory

**COARSE**
- BE
- HAVE
- IN
- ACTOR
- INST
- ABOUT

**DIRECTED**
- HAVE
  - HAVE\(_1\)
    - POSSESSOR-POSSESSION\(_1\)
    - EXPERIENCER-CONDITION\(_1\)
    - OBJECT-PROPERTY\(_1\)
    - WHOLE-PART\(_1\)
    - GROUP-MEMBER\(_1\)
  - HAVE\(_2\)
    - POSSESSOR-POSSESSION\(_2\)
    - EXPERIENCER-CONDITION\(_2\)
    - OBJECT-PROPERTY\(_2\)
    - WHOLE-PART\(_2\)
    - GROUP-MEMBER\(_2\)

**FINE**
- family firm
- reader mood
- grass scent
- car tyre
- group member
- hotel owner
- coma victim
- quality puppy
- shelf unit
- lecture course
2,000 candidate two-noun compounds sampled from the British National Corpus

Filtered for extraction errors and idioms

1,443 unique compounds labelled with semantic relations at each level of granularity

<table>
<thead>
<tr>
<th>Granularity</th>
<th>Labels</th>
<th>Agreement ($\kappa$)</th>
<th>Random Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coarse</td>
<td>6</td>
<td>0.62</td>
<td>16.3%</td>
</tr>
<tr>
<td>Directed</td>
<td>10</td>
<td>0.61</td>
<td>10.0%</td>
</tr>
<tr>
<td>Fine</td>
<td>27</td>
<td>0.56</td>
<td>3.7%</td>
</tr>
</tbody>
</table>

Try it out yourself: http://www.cl.cam.ac.uk/~do242/Resources/1443_Compounds.tar.gz
Information sources for relation classification

**Lexical information:** Information about the individual constituent words of a compound.

**Relational information:** Information about how the entities denoted by a compound's constituents typically interact in the world.

**Contextual information:** Information derived from the context in which a compound occurs.

[Nastase et al., 2013]
Information sources for relation classification

**Lexical information:** Information about the individual constituent words of a compound.

**Relational information:** Information about how the entities denoted by a compound's constituents typically interact in the world.

**Contextual information:** Information derived from the context in which a compound occurs.

[Nastase et al., 2013]
### Information sources for kidney disease

<table>
<thead>
<tr>
<th><strong>Lexical:</strong></th>
<th><strong>Relational:</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><em>modifier (coord)</em></td>
<td>Stagnant water breeds fatal diseases of liver and kidney such as hepatitis</td>
</tr>
<tr>
<td><em>head (coord)</em></td>
<td>Chronic disease causes kidney function to worsen over time until dialysis is needed</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Context:</strong></th>
<th><strong>-</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>These include the elderly, people with chronic respiratory disease, chronic heart disease, kidney disease and diabetes, and health service staff</td>
<td><strong>-</strong></td>
</tr>
</tbody>
</table>
Information sources for holiday village

**Lexical:**
- **modifier (coord)**: weekend:507, sunday:198, holiday:180, day:159, event:115
- **head (coord)**: municipality:9417, parish:4786, town:4526, hamlet:1634, city:1263

**Relational:**
- He is spending the holiday at his grandmother’s house in the village of Busang in the Vosges region.
- The Prime Minister and his family will spend their holidays in Vernet, a village of 2,000 inhabitants located about 20 kilometers south of Toulouse.
- Other holiday activities include a guided tour of Panama City, a visit to an Indian village and a helicopter tour.

**Context:**
- For FFr100m ($17.5m), American Express has bought a 2% stake in Club Méditerranée, a French group that ranks third among European tour operators, and runs holiday villages in exotic places.
Contextual information doesn’t help

- Contextual information does not have discriminative power for compound interpretation [Ó Séaghdha and Copestake, 2007]

We slept in a mountain hut
We slept in a timber hut
We slept in a timber camp

I cut it with the cheese knife
I cut it with the kitchen knife
I cut it with the steel knife

- Sparsity also an issue
- Not considered further here
Experimental setup

- 5-fold cross-validation on 1443-compounds
- All experiments use a Support Vector Machine classifier (LIBSVM)
- SVM cost parameter ($c$) set per fold by cross-validation on the training data
- Kernel derived from Jensen-Shannon divergence [Ó Séaghdha and Copestake, 2008; 2013]:

$$k_{JSD(\text{linear})}(\mathbf{p}, \mathbf{q}) = - \sum_i p_i \log_2 \left( \frac{p_i}{p_i + q_i} \right) + q_i \log_2 \left( \frac{q_i}{p_i + q_i} \right)$$
Lexical features

- Distributional features extracted from parsed BNC and Wikipedia corpora.

- One vector for each constituent:
  
  **Coordination** Distribution over nouns co-occurring in a coordination relation

  **All GRs** Distribution over all lexicalised grammatical relations involving a noun, verb, adjective or adverb

  **GR Clusters** 1000-dimensional representation learned with Latent Dirichlet Allocation from All GRs data

  [Ó Séaghdha and Korhonen, 2011; 2014]
Results - lexical features

![Accuracy and F-Score Bar Charts]

**Accuracy**
- Coarse: 63.0
- Directed: 62.2
- Fine: 51.2

**F-Score**
- Coarse: 61.0
- Directed: 57.4
- Fine: 47.1

Legend:
- Blue: Coordination
- Red: All GRs
- Green: GR Clusters
Relational features

- **Context set** for a compound $N_1 N_2$: the set of all contexts in a corpus where $N_1$ and $N_2$ co-occur
- Context sets for all compounds extracted from Gigaword and BNC corpora
- Embeddings for strings:
  - **Gap-weighted**: all discontinuous $n$-grams [Lodhi et al., 2002]
  - **PairClass**: fixed length (up to 7-word) patterns with wildcards [Turney, 2008]
- Context set representation is the average of its members’ embeddings
Results - relational features

Accuracy

- Coarse: 52.0
- Directed: 49.8
- Fine: 37.8

F-Score

- Coarse (Gap-weighted): 49.8
- Directed (PairClass): 43.1
- Fine: 29.7
Results - combined features

Accuracy

<table>
<thead>
<tr>
<th></th>
<th>Coarse</th>
<th>Directed</th>
<th>Fine</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>65.4</td>
<td>64.4</td>
<td>53.5</td>
</tr>
</tbody>
</table>

F-Score

<table>
<thead>
<tr>
<th></th>
<th>Coarse</th>
<th>Directed</th>
<th>Fine</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>64.0</td>
<td>59.1</td>
<td>47.6</td>
</tr>
</tbody>
</table>
Head-only vs modifier-only features

![Bar chart showing F-Score comparison between Head-only and Modifier-only features]
Effect of context set size

F-Score (Coarse labels)

- Relational
- Lexical

Size of context set

0-199  200-399  400-599  600-799  800-999  1000+

0  20  40  60
Conclusions

- Compound interpretation is fun!
- Combining lexical and relation information leads to state-of-the-art performance.
- Previous best performance on 1443-Compounds: 63.6% accuracy on coarse labels [Tratz and Hovy, 2010]
- Our best:

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coarse</td>
<td>65.4</td>
<td>64.0</td>
</tr>
<tr>
<td>Directed</td>
<td>64.4</td>
<td>59.1</td>
</tr>
<tr>
<td>Fine</td>
<td>53.5</td>
<td>47.6</td>
</tr>
</tbody>
</table>