Noun Compound Interpretation

By
Girishkumar Ponkiya

Mentor
Mr. Girish Palshikar

TRDDC, Pune

May-July 2014
Motivational Example

• Our website homepage logo design was finalized by that indian software designer team.

  (ROOT
   (S
    (NP (PRP$ Our) (NN website) (NN homepage) (NN logo) (NN design))
    (VP (VBD was)
      (VP (VBN finalized)
        (PP (IN by)
          (NP (DT that) (JJ indian) (NN software) (NN designer) (NN team))))))
  )
Motivational Example

- Our website homepage logo design was finalized by that Indian software designer team.
Some more examples..

- **Simple (??)**
  - bone marrow
  - web site design
  - internet connection speed test
  - plastic water bottle

- **Complicated (??)**
  - colon cancer tumor suppressor protein
Simplifying complexity

- colon cancer tumor suppressor protein

  - [tumor suppressor protein] which is implicated in [colon cancer]
    - (IN; LOCATION)

  - [protein] that acts as [tumor suppressor]
    - (IS; AGENT)

  - [suppressor] that inhibits [tumor(s)]
    - (OF; PURPOSE)

  - [cancer] that occurs in [(the) colon]
    - (OF; IN; LOCATION)
Corpus Statistics

• 2-4% of the tokens in various corpora are part of noun compounds (Baldwin and Tanaka, 2004)
  – 2.6% in the British National Corpus
  – 3.9% in the Reuters corpus
  – 2.9% in the Mainichi Shimbun Corpus

• 100M-word British National Corpus (BNC)
  – 939K distinct wordforms
  – 256K distinct noun compounds
Introduction

• Noun Compound (NC): “a sequence of two or more nouns”
  
  e.g. box juice, computer science department

• Individual nouns in the NC are known as “components”

• Three main problems:
  – Identifying noun compound
  – Syntactic analysis (bracketing)
  – Semantic Relation assignment
Bracketing

- Determining syntactic structure
- Examples:

  (1) *liver cell antibody*

  \[
  \begin{array}{c}
  \text{liver} \\
  \text{cell} \\
  \text{antibody}
  \end{array}
  \]

  (2) *liver cell line*

  \[
  \begin{array}{c}
  \text{liver} \\
  \text{cell} \\
  \text{line}
  \end{array}
  \]
Bracketing

• Methods
  e.g. computer science department, linguistics graduate program
  – **Adjacency model**
    based on frequency of (N1,N2) and (N2,N3) in bia-gram data
  – **Dependency model**
    based on frequency of (N1,N3) and (N2,N3) in dependency data
  – **Hybrid**
    • n-gram, adjacency, dependency, and some more features
Semantic Interpretation

• Approaches
  – Rule based (Vanderwende, 1994)
  – Statistical
    • Analogy based reasoning
      – “similar component words should have the same SR”
        e.g. $cat:meow \Leftrightarrow dog:bark$
    • Semantic disambiguation
      – Disambiguation relative to an underlying predicate or paraphrase
Levi's Theory (1978)

- Idea: study how noun compound can be derived
- Two syntactic processes:
  - predicate nominalization
    - For example, in sentence:
      
      ..the President refused General MacArthur’s request..
      
      $\rightarrow$ presidential refusal
  
  - predicate deletion
    - Example:
      
      pie made of apples $\rightarrow$ apple pie
    
    - Proposed set of abstract recoverably deletable predicates
## Recoverably Deletable Predicates

<table>
<thead>
<tr>
<th>RDP</th>
<th>Example</th>
<th>Subj/obj</th>
<th>Traditional Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAUSE₁</td>
<td>tear gas</td>
<td>object</td>
<td>causative</td>
</tr>
<tr>
<td>CAUSE₂</td>
<td>drug deaths</td>
<td>subject</td>
<td>causative</td>
</tr>
<tr>
<td>HAVE₁</td>
<td>apple cake</td>
<td>object</td>
<td>possessive/dative</td>
</tr>
<tr>
<td>HAVE₂</td>
<td>lemon peel</td>
<td>subject</td>
<td>possessive/dative</td>
</tr>
<tr>
<td>MAKE₁</td>
<td>silkworm</td>
<td>object</td>
<td>productive/composit.</td>
</tr>
<tr>
<td>MAKE₂</td>
<td>snowball</td>
<td>subject</td>
<td>productive/composit.</td>
</tr>
<tr>
<td>USE</td>
<td>steam iron</td>
<td>object</td>
<td>instrumental</td>
</tr>
<tr>
<td>BE</td>
<td>soldier ant</td>
<td>object</td>
<td>essive/appositional</td>
</tr>
<tr>
<td>IN</td>
<td>field mouse</td>
<td>object</td>
<td>locative</td>
</tr>
<tr>
<td>FOR</td>
<td>horse doctor</td>
<td>object</td>
<td>purposive/benefactive</td>
</tr>
<tr>
<td>FROM</td>
<td>olive oil</td>
<td>object</td>
<td>source/ablative</td>
</tr>
<tr>
<td>ABOUT</td>
<td>price war</td>
<td>object</td>
<td>topic</td>
</tr>
</tbody>
</table>
O Seaghdha's Theory (2007)

- Revised the inventory of Levi (1978)
  - The inventory of relations should have good **coverage**
    - *history teacher, woman driver*
  - Relations should be disjunct, and should describe a **coherent** concept
    - Overlapping category boundaries
    - Annotation guidelines
  - The **class distribution** should not be overly skewed or sparse
  - The concepts underlying the relations should **generalize** to other linguistic phenomena
    - The guidelines should make the **annotation process** as simple as possible
    - The categories should provide useful semantic information.

- 2000 samples in dataset
Warren's Theory (1978)

- Based on study of Brown corpus
- Abstract semantic relations organized into a four-level hierarchy
  - **CONSTITUTE**: A is something that wholly constitutes B, or vice-versa
    - Source-Result, Result-Source, Copula
  - **POSSESSION**: A is something of which B is a part or a feature or vice versa
    - Part-Whole, Whole-Part, Size-Whole
  - **LOCATION**: A is the location or origin of B (in time or space)
    - Place-OBJ, Time-OBJ, Origin-OBJ
  - **ACTIVITY-ACTOR**: The comment indicates the activity or interest with which B is habitually concerned
  - **RESEMBLANCE**: A indicates something that B resembles
    - Comparant-Compared
  - **PURPOSE**: A is purpose of B, or vice-versa.
Improving Warren's Theory

• Barker & Szpakowicz (1998)
  – Flat 20 relations
  – From Wall Street Journal (Kim and Baldwin, 2005)
    • 2,169 unique 2-term NC
    • 1,571 unique 3-term NC

• Nastase & Szpakowicz (2003)
  – 5 coarse-grained super-relations
  – 30 fine-grained relations
  – 600 samples in dataset
A Lexical Semantic Approach to Interpreting and Bracketing English Noun Compounds

Su Nam Kim and Timothy Baldwin
Overview

• Goal
  – Automatic NC interpretation

• Approach
  – Analogical, based on WordNet similarity

• Other
  – NC interpretation helps bracketing
Semantic Relations

- Used the set of 20 SRs proposed by Barker and Szpakowicz (1998)
  - Relatively well-established in NLP research
  - Found to adequately capture the dataset used in this paper

- List of SRs in next slide
<table>
<thead>
<tr>
<th>Relation</th>
<th>Definition</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGENT</td>
<td>$N_2$ is performed by $N_1$</td>
<td>student protest, band concert, military assault</td>
</tr>
<tr>
<td>BENEFICIARY</td>
<td>$N_1$ benefits from $N_2$</td>
<td>student price, charitable compound</td>
</tr>
<tr>
<td>CAUSE</td>
<td>$N_1$ causes $N_2$</td>
<td>printer tray, flood water, film music, story idea</td>
</tr>
<tr>
<td>CONTAINER</td>
<td>$N_1$ contains $N_2$</td>
<td>exam anxiety, overdue fine</td>
</tr>
<tr>
<td>CONTENT</td>
<td>$N_1$ is contained in $N_2$</td>
<td>paper tray, eviction notice, oil pan</td>
</tr>
<tr>
<td>DESTINATION</td>
<td>$N_1$ is destination of $N_2$</td>
<td>game bus, exit route, entrance stairs</td>
</tr>
<tr>
<td>EQUATEVIVE</td>
<td>$N_1$ and $N_2$</td>
<td>composer arranger, player coach</td>
</tr>
<tr>
<td>INSTRUMENT</td>
<td>$N_1$ is used in $N_2$</td>
<td>electron microscope, diesel engine, laser printer</td>
</tr>
<tr>
<td>LOCATED</td>
<td>$N_1$ is located at $N_2$</td>
<td>building site, home town, solar system</td>
</tr>
<tr>
<td>LOCATION</td>
<td>$N_1$ is the location of $N_2$</td>
<td>lab printer, desert storm, internal combustion</td>
</tr>
<tr>
<td>MATERIAL</td>
<td>$N_2$ is made of $N_1$</td>
<td>carbon deposit, gingerbread man, water vapour</td>
</tr>
<tr>
<td>OBJECT</td>
<td>$N_1$ is acted on by $N_2$</td>
<td>engine repair, horse doctor</td>
</tr>
<tr>
<td>POSSESSOR</td>
<td>$N_1$ has $N_2$</td>
<td>student loan, company car, national debt</td>
</tr>
<tr>
<td>PRODUCT</td>
<td>$N_1$ is a product of $N_2$</td>
<td>automobile factory, light bulb, color printer</td>
</tr>
<tr>
<td>PROPERTY</td>
<td>$N_2$ is $N_1$</td>
<td>elephant seal, blue car, big house, fast computer</td>
</tr>
<tr>
<td>PURPOSE</td>
<td>$N_2$ is meant for $N_1$</td>
<td>concert hall, soup pot, grinding abrasive</td>
</tr>
<tr>
<td>RESULT</td>
<td>$N_1$ is a result of $N_2$</td>
<td>storm cloud, cold virus, death penalty</td>
</tr>
<tr>
<td>SOURCE</td>
<td>$N_1$ is the source of $N_2$</td>
<td>chest pain, north wind, foreign capital</td>
</tr>
<tr>
<td>TIME</td>
<td>$N_1$ is the time of $N_2$</td>
<td>winter semester, morning class, late supper</td>
</tr>
<tr>
<td>TOPIC</td>
<td>$N_2$ is concerned with $N_1$</td>
<td>computer expert, safety standard, horror novel</td>
</tr>
</tbody>
</table>
NC Interpretation: Approach

- For 2-term NC

\[
S((N_{i,1}, N_{i,2}), (B_{j,1}, B_{j,2})) = \alpha S1 + (1 - \alpha) S2
\]
NC Interpretation: Example

- For 2-term NC

<table>
<thead>
<tr>
<th>Training noun</th>
<th>Test noun</th>
<th>$S_{ij}$</th>
<th>Combined Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_1$</td>
<td>apple</td>
<td>0.71</td>
<td></td>
</tr>
<tr>
<td>$N_2$</td>
<td>chocolate</td>
<td>0.83</td>
<td>0.77</td>
</tr>
<tr>
<td>$N_1$</td>
<td>juice</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$N_2$</td>
<td>milk</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$N_1$</td>
<td>morning</td>
<td>0.27</td>
<td>0.64</td>
</tr>
<tr>
<td>$N_2$</td>
<td>milk</td>
<td>1.00</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Training noun</th>
<th>Test noun</th>
<th>$S_{ij}$</th>
<th>Combined Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_1$</td>
<td>personal</td>
<td>0.32</td>
<td>0.58</td>
</tr>
<tr>
<td>$N_2$</td>
<td>interest</td>
<td>0.84</td>
<td></td>
</tr>
<tr>
<td>$N_1$</td>
<td>loan</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$N_2$</td>
<td>rate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$N_1$</td>
<td>bank</td>
<td>0.75</td>
<td>0.80</td>
</tr>
<tr>
<td>$N_2$</td>
<td>interest</td>
<td>0.84</td>
<td></td>
</tr>
<tr>
<td>$N_1$</td>
<td>loan</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$N_2$</td>
<td>rate</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
NC Interpretation: Approach

• For 2-term NC

\[ m = \arg\max_j S((N_{i,1}, N_{i,2}), (B_{j,1}, B_{j,2})) \]
Data Collection

- Source: Wall Street Journal
- Collected 2-term and 3-terms NCs
  - 2,169 unique 2-term NCs
  - 1,571 unique 3-term NCs
Data Annotation

- 2 trained human annotator
- First step: bracketing 3-term NC
- Second step: tagged outermost 2-term NC
  - (N2 N3) for ((N1 N2) N3), and
  - (N1 N3) for (N1 (N2 N3))
- Multiple SRs were assigned
  - e.g. *debt cost* : SOURCE or CAUSE ??
- Agreement for SR
  - 2-term: 52.31 %
  - 3-term: 49.28 %
<table>
<thead>
<tr>
<th>Relation</th>
<th>2-term NCs</th>
<th></th>
<th></th>
<th>3-term NCs</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Test N+</td>
<td>M</td>
<td>Training N+</td>
<td>M</td>
<td>Test N+</td>
<td>M</td>
</tr>
<tr>
<td>AGENT</td>
<td>10</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>BENEFICIARY</td>
<td>10</td>
<td>1</td>
<td>7</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>CAUSE</td>
<td>54</td>
<td>5</td>
<td>74</td>
<td>3</td>
<td>21</td>
<td>0</td>
</tr>
<tr>
<td>CONTAINER</td>
<td>13</td>
<td>4</td>
<td>19</td>
<td>3</td>
<td>13</td>
<td>1</td>
</tr>
<tr>
<td>CONTENT</td>
<td>40</td>
<td>2</td>
<td>34</td>
<td>2</td>
<td>23</td>
<td>0</td>
</tr>
<tr>
<td>DESTINATION</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>EQUATIVE</td>
<td>9</td>
<td>0</td>
<td>17</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>INSTRUMENT</td>
<td>6</td>
<td>0</td>
<td>11</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>LOCATED</td>
<td>12</td>
<td>1</td>
<td>16</td>
<td>2</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>LOCATION</td>
<td>29</td>
<td>9</td>
<td>24</td>
<td>4</td>
<td>19</td>
<td>0</td>
</tr>
<tr>
<td>MATERIAL</td>
<td>12</td>
<td>0</td>
<td>14</td>
<td>1</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>OBJECT</td>
<td>88</td>
<td>6</td>
<td>88</td>
<td>5</td>
<td>22</td>
<td>6</td>
</tr>
<tr>
<td>POSSESSOR</td>
<td>33</td>
<td>1</td>
<td>22</td>
<td>1</td>
<td>25</td>
<td>4</td>
</tr>
<tr>
<td>PRODUCT</td>
<td>27</td>
<td>0</td>
<td>32</td>
<td>6</td>
<td>27</td>
<td>1</td>
</tr>
<tr>
<td>PROPERTY</td>
<td>76</td>
<td>3</td>
<td>85</td>
<td>3</td>
<td>33</td>
<td>0</td>
</tr>
<tr>
<td>PURPOSE</td>
<td>159</td>
<td>13</td>
<td>161</td>
<td>9</td>
<td>89</td>
<td>7</td>
</tr>
<tr>
<td>RESULT</td>
<td>7</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>SOURCE</td>
<td>75</td>
<td>11</td>
<td>99</td>
<td>15</td>
<td>61</td>
<td>0</td>
</tr>
<tr>
<td>TIME</td>
<td>25</td>
<td>1</td>
<td>19</td>
<td>0</td>
<td>19</td>
<td>0</td>
</tr>
<tr>
<td>TOPIC</td>
<td>465</td>
<td>24</td>
<td>447</td>
<td>39</td>
<td>438</td>
<td>16</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>1163</strong></td>
<td><strong>82</strong></td>
<td><strong>1184</strong></td>
<td><strong>96</strong></td>
<td><strong>820</strong></td>
<td><strong>35</strong></td>
</tr>
</tbody>
</table>
Experiments #1

- For 2-term NC
- With equal weight for head and modifier similarities
- $k$-NN methods with various $k$ values
  - $k=1$ was found better
- Contribution of training-data size
**Experiment #1: Result**

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human annotation</td>
<td></td>
</tr>
<tr>
<td>Majority class</td>
<td></td>
</tr>
<tr>
<td>Path-based</td>
<td></td>
</tr>
<tr>
<td>Information content-based</td>
<td></td>
</tr>
<tr>
<td>Relatedness</td>
<td></td>
</tr>
<tr>
<td>Random</td>
<td></td>
</tr>
<tr>
<td>Inter-annotator agreement</td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>43.0%</td>
</tr>
<tr>
<td>WUP</td>
<td>53.3%</td>
</tr>
<tr>
<td>LCH</td>
<td>52.9%</td>
</tr>
<tr>
<td>JCN</td>
<td>46.7%</td>
</tr>
<tr>
<td>LIN</td>
<td>47.4%</td>
</tr>
<tr>
<td>LESK</td>
<td>42.4%</td>
</tr>
<tr>
<td>RANDOM</td>
<td>21.8%</td>
</tr>
</tbody>
</table>

Table 7. Accuracy of NC interpretation for the different WordNet-based scoring methods over our 2-term NC dataset
Experiment #1: Result

Fig. 3. Learning Curve with respect to the size of the training data
Experiment #2

- To check relative contribution of head and modifier

\[ S((N_{i,1}, N_{i,2}), (B_{j,1}, B_{j,2})) = \alpha S1 + (1 - \alpha) S2 \]

- For example
  - Head plays important role in PROPERTY relation e.g. *fairy penguin*
  - Modifier plays important role in TIME relation i.e. *winter coat*
Experiment #2: Result

Fig. 4. Classifier accuracy at different $\alpha$ values
Experiment #2: Result
Various Relational Approaches

- Using 8 prepositions (Lauer, 1995)
- Verbs + prepositions (Nakov and Hearst, 2006)
- Using mind pattern from web (Turney, 2006)
  
  e.g. “Y * causes X” for CAUSE

- Pattern from corpus analysis (Turney & Littman, 2005)
  - 128 fixed phrases using 64 joining-terms
0.87 “cooking utensils” FOR

**Human:**  
- be used for\(17\),  
- be used in\(9\),  
- facilitate\(4\),  
- help\(3\),  
- aid\(3\),  
- be required for\(2\),  
- be used during\(2\),  
- be found in\(2\),  
- be utilized in\(2\),  
- involve\(2\),  

**Progr.:**  
- be used for\(43\),  
- be used in\(11\),  
- make\(6\),  
- be suited for\(5\),  
- replace\(3\),  
- be used during\(2\),  
- facilitate\(2\),  
- turn\(2\),  
- keep\(2\),  
- be for\(1\),  

**Table 3. Human- and programme-proposed vectors, and cosines for sample noun-noun compounds.** The common verbs for each vector pair are underlined.
Use of Semantic Relation in NC

- Paraphrase-augmented machine translation
- Summarisation evaluation
- Textual entailment
- Information retrieval
  - index normalisation, query expansion, query refinement, results re-ranking, etc.
- Data mining
  - *Migraine treatment* → “*which prevents migraines*”
Our work

• Goal: extract “rules” for compound based on semantics of components
  – Used 20 relations proposed by Barker and Szpakowicz (1998)
• Explored ConceptNet, WordNet, and VerbNet
• Used CN2
References


• Vivi Nastase and Stan Szpakowicz. “Exploring noun-modifier semantic relations”. In Fifth international workshop on computational semantics (IWCS-5), pages 285–301, 2003
Thanks..