Large-Scale Noun Compound Interpretation Using Bootstrapping and the Web as a Corpus

Su Nam Kim† and Preslav Nakov‡

The University of Melbourne†
National University of Singapore‡
Overview

- **Task**: semantic interpretation of noun compounds

- **Goal**: compare abstract relations vs. paraphrasing verbs

- **Idea**: large dataset of noun compounds interpreted by both
  - an abstract relation, and
  - paraphrasing verbs

- **Approach**: bootstrapping from the Web

- **Result**: more restrictions yield higher accuracy, and more NCs
Noun Compounds

- **Definition**: Sequence of nouns that function as one noun.

- **Examples**:
  - silkworm
  - olive oil
  - healthcare reform
  - plastic water bottle
  - colon cancer tumor suppressor protein
Noun Compounds

- **Definition**: Sequence of nouns that function as one noun.

- **Examples**:
  - silkworm
  - olive oil
  - healthcare reform
  - plastic water bottle
  - colon cancer tumor suppressor protein

- **Our focus**: Semantics of two-word noun compounds
Noun Compounds: Properties

- **Encode implicit relations:** *hard to interpret*
  - *plastic bottle* – Material
  - *water bottle* – Container

- **Abundant:** *cannot be ignored*
  - cover 4% of the tokens in the Reuters corpus

- **Highly productive:** *cannot be listed in a dictionary*
  - 60% of the NCs in BNC occur just once
Noun Compounds: Applications

- Question Answering
- Information Extraction
- Information Retrieval
- Machine Translation
- Textual Entailment

* malaria mosquito can be paraphrased as
  * mosquito with malaria
  * mosquito spreading malaria
  * mosquito that causes malaria
  * malaria-spreading mosquito
Noun Compounds: Applications

• Question Answering
  Information Extraction
  Information Retrieval
  Machine Translation
  Textual Entailment

★ malaria mosquito can be paraphrased as
  * mosquito with malaria
  * mosquito spreading malaria
  * mosquito that causes malaria
  * malaria-spreading mosquito

• Relational Search

★ Query: Find all X such that X causes malaria (as in TEXTRUNNER)
★ Result: malaria mosquito, ...
Noun Compounds: Semantics

- **Abstract relations** (Nastase & Szpakowicz 2003; Kim & Baldwin 2005; Girju 2007; Ó Séaghdha & Copestake 2007)

  - *malaria mosquito*: **Cause**
  - *olive oil*: **Source**
Noun Compounds: Semantics

- **Abstract relations** (Nastase & Szpakowicz 2003; Kim & Baldwin 2005; Girju 2007; Ó Séaghdha & Copestake 2007)
  - malaria mosquito: **Cause**
  - olive oil: **Source**

- **Prepositions** (Lauer 1995)
  - malaria mosquito: *with*
  - olive oil: *from*
Noun Compounds: Semantics

- **Abstract relations** (Nastase & Szpakowicz 2003; Kim & Baldwin 2005; Girju 2007; Ó Séaghdha & Copestake 2007)
  - *malaria mosquito*: **Cause**
  - *olive oil*: **Source**

- **Prepositions** (Lauer 1995)
  - *malaria mosquito*: **with**
  - *olive oil*: **from**
  - *morning flight, field mouse*: **in** (also **is in**)
Noun Compounds: Semantics

- **Abstract relations** (Nastase & Szpakowicz 2003; Kim & Baldwin 2005; Girju 2007; Ó Séaghdha & Copestake 2007)
  - malaria mosquito: **Cause**
  - olive oil: **Source**

- **Prepositions** (Lauer 1995)
  - malaria mosquito: *with*
  - olive oil: *from*
  - morning flight, field mouse: *in* (also *is in)*

  - malaria mosquito: carries, spreads, causes, transmits, brings, has
  - olive oil: comes from, is made from, is derived from
Abstract Relations vs. Paraphrasing Verbs

<table>
<thead>
<tr>
<th></th>
<th>cancer treatment</th>
<th>migraine treatment</th>
<th>wrinkle treatment</th>
<th>herb treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>treat</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>prevent</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>cure</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>reduce</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>smooth</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>cause</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>contain</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>use</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
</tbody>
</table>

- **TREATMENT-FOR-DISEASE**
  - cancer/migraine/wrinkle treatment: positive
  - herb treatment: negative
Abstract Relations vs. Paraphrasing Verbs

- **Abstract relations**
  - Useful generalization
  - BUT
    - no universal set
    - too coarse grained
    - issues with ambiguity and coverage
    - limited use in real tasks

- **Verbs**
  - Fine grained distinctions
  - Directly usable in paraphrases

- **Maybe we need both representations?**
Overview (again)

- **Goal**: compare abstract relations vs. paraphrasing verbs

- **Idea**: large dataset of noun compounds interpreted by both
  - an abstract relation – offers useful generalization, and
  - verbs – fine-grained semantics; directly usable in paraphrases.

- **Approach**: bootstrapping from the Web
  - NCs that express a given abstract relation
  - verbs that interpret these NCs
Note 1: Paraphrasing Verbs

• Can paraphrase an NC

  ★ chocolate bar: be made of, contain, be composed of, taste like
Note 1: Paraphrasing Verbs

• Can paraphrase an NC
  ✤ chocolate bar: be made of, contain, be composed of, taste like

• Can also express an abstract relation
  ✤ MAKE₂: be made of, be composed of, consist of, be manufactured from
Note 1: Paraphrasing Verbs

- **Can paraphrase an NC**
  - chocolate bar: be made of, contain, be composed of, taste like

- **Can also express an abstract relation**
  - MAKE₂: be made of, be composed of, consist of, be manufactured from

- **... but can also be NC-specific**
  - orange juice: be squeezed from
  - bacon pizza: be topped with
  - chocolate bar: taste like
Note 2: Distribution over Verbs

- Single verb
  - malaria mosquito: cause
  - olive oil: be extracted from
Note 2: Distribution over Verbs

- Single verb
  - malaria mosquito: cause
  - olive oil: be extracted from

- Multiple verbs
  - malaria mosquito: cause, carry, spread, transmit, bring, have
  - olive oil: be extracted from, come from, be made from, contain, taste like
Note 2: Distribution over Verbs

- **Single verb**
  - *malaria mosquito*: cause
  - *olive oil*: be extracted from

- **Multiple verbs**
  - *malaria mosquito*: cause, carry, spread, transmit, bring, have
  - *olive oil*: be extracted from, come from, be made from, contain, taste like

- **Distribution over verbs** (SemEval-2010 Task 9)
  - *malaria mosquito*: carry (23), spread (16), cause (12), transmit (9), bring (7), have (4), be infected with (3), infect with (3), give (2), ... 
  - *olive oil*: come from (33), be made from (27), be derived from (10), be made of (7), contain (7), be pressed from (6), be extracted from (5), ...
Target Representation

* **NC semantics**
  * abstract relation
  * distribution over verbs that
    * can express the abstract relation, or
    * can be NC-specific
Target Representation

- **NC semantics**
  - *abstract relation*
  - *distribution over verbs that*
    - can express the abstract relation, or
    - can be NC-specific

- **chocolate bar**
  - *abstract relation: MAKE₂*
  - *verbs: be made of (16), contain (16), be made from (10), be composed of (7), taste like (7), consist of (5), ...*
## Relation Inventory: Levi’s Predicates

<table>
<thead>
<tr>
<th>Rel.</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>
Target Relation

- **MAKE$_2$** from Levi’s theory

- **Definition**: HEAD is made up of or is a product of MOD.

- **Subtypes**:

  (a) MOD: unit, HEAD: configuration, e.g., *root system*;  
  (b) MOD: material, HEAD: mass/artefact, e.g., *chocolate bar*;  
  (c) MOD: specifier, HEAD: human collectives, e.g., *worker teams*. 
Data: Verbs from SemEval-2010 Task 9

- Data
  - 20 examples of $\text{MAKE}_2$ (from Levi’78); 25-30 annotators
  - kept: verbs proposed by $\geq 5$ annotators
  - removed $\text{be}$

- Example
  - chocolate bar: be made of (16), contain (16), be made from (10), be composed of (7), taste like (7), consist of (5), be (3), have (2), melt into (2), be manufactured from (2), be formed from (2), smell of (2), be flavored with (1), sell (1), taste of (1), be constituted by (1), incorporate (1), serve (1), contain (1), store (1), be made with (1), be solidified from (1), be created from (1), be flavoured with (1), be comprised of (1)
The Initial Seed Examples for \text{MAKE}_2

84 NC-pattern pairs

\textbf{bronze statue: } be made of, be composed of, contain

\textbf{cable network: } consist of, be made of

\textbf{candy cigarette: } be made of, taste like, be made from, look like

\textbf{chocolate bar: } contain, be made of, be made from, taste like, be composed of, consist of

\textbf{copper coin: } be made of, be made from, contain, be composed of

\textbf{daisy chain: } be made of, be made from, contain, consist of

\textbf{glass eye: } be made of, be composed of, be made from

\ldots
The Initial Seed Examples for MAKE$_2$

- **20 NCs (from Levi’78):** bronze statue, cable network, candy cigarette, chocolate bar, concrete desert, copper coin, daisy chain, glass eye, immigrant minority, mountain range, paper money, plastic toy, sand dune, steel helmet, stone tool, student committee, sugar cube, warrior castle, water drop, worker team.

- **18 patterns (from Nakov’08):** be composed of, be comprised of, be inhabited by, be lived in by, be made from, be made of, be made up of, be manufactured from, be printed on, consist of, contain, have, house, include, involve, look like, resemble, taste like.
Our Bootstrapping Method

Patterns
(+ H/M of NCs)

Query Generation

Snippet by Yahoo!

NC Extraction

collected NCs

stop
if newNCs = 0
or
Iteration limit exceeded

NC Filtering Rules

repeat

Pattern Extraction

collected Patterns
w/ NCs

NC Filtering Rules

Query Generation

Snippet by Yahoo!

Patterns
(+ H/M of NCs)
Bootstrapping Step 1: NC Acquisition (I)

- **1.1: Query Generation**
  - **Generalized query templates**
    
    
    - "* that PATTERN *" (loose)
    - "HEAD that PATTERN *" (strict)
    - "* that PATTERN MOD" (strict)
Bootstrapping Step 1: NC Acquisition (I)

1.1: Query Generation

- Generalized query templates

  "* that PATTERN *" (loose)
  "HEAD that PATTERN *" (strict)
  "* that PATTERN MOD" (strict)

- Example instantiations for be made of (and orange juice)

  "* that were made of *"
  "juice that was made of *"
  "* is made of oranges"
Bootstrapping Step 1: NC Acquisition (II)

- **Loose bootstrapping**: uses patterns only

  
  "* that PATTERN *" \( (\text{loose}) \)
Bootstrapping Step 1: NC Acquisition (II)

- **Loose bootstrapping**: uses patterns only
  
  "* that PATTERN *" (loose)

- **Strict bootstrapping**: uses NC-pattern pairs
  
  "HEAD that PATTERN *" (strict)
  "* that PATTERN MOD" (strict)
Bootstrapping Step 1: NC Acquisition (II)

- **Loose bootstrapping**: uses patterns only
  
  "* that PATTERN *" (loose)

- **Strict bootstrapping**: uses NC-pattern pairs
  
  "HEAD that PATTERN *" (strict)
  "* that PATTERN MOD" (strict)

- **NC-only strict bootstrapping**
  
  ⋅ strict bootstrapping limited to the initial 18 verbs
Bootstrapping Step 1: NC Acquisition (III)

• 1.2: Snippet Extraction: for the top 1,000 results.

• 1.3: Noun Compound Extraction

  ✴ Restrictions: reject the candidate noun compound if
  ✴ head == modifier;
  ✴ head/modifier is not a noun in WordNet;
  ✴ NC is a seed example or was already extracted;
  ✴ NC occurs fewer than 100 times in the Google Web 1T 5-gram corpus;
  ✴ NC is extracted fewer than $N$ (5; 10) times in the context of the pattern
Bootstrapping Step 2: Pattern Extraction (I)

- **2.1: Query Generation**
  
  ✪ Generalized query template
  
  
  "HEAD THAT? * MOD"

  ✪ THAT? is *that*, *which*, *who* or the empty string
  ✪ up to six stars

  ✪ Example instantiations for *orange juice*

  "juice *that* *oranges"
  "juices *which* *oranges"
  "juices *orange"
Bootstrapping Step 2: Pattern Extraction (II)

• **2.2: Snippet Extraction:** for the top 1,000 results.

• **2.3: Pattern Extraction**
  
  ★ **Select the top 20 most frequent patterns only.**
  ★ **Reject the pattern if**
    * is a seed example or was already extracted;
    * extracted fewer than $N$ times (5, 10) or with fewer than $M$ NCs (20, 50).

**Note:** patterns filtered for bootstrapping but not for the NCs
Examples: Found and Retained for Bootstrapping

- **NCs**
  - *brass bell* (be made of, be made from)
  - *child team* (be composed of, include)

- **Patterns**
  - *be filled with* (cotton bag, water cup)
  - *use* (water sculpture, wood statue)
Evaluation: Two Questions

- **For a noun compound**
  - Q: Does it express the target abstract relation?
  - Judged: all NCs, unless too many

- **For an NC-pattern pair**
  - Q: Does the pattern (verb) paraphrase the NC?
  - Judged: for each pattern, the 10 most frequent NCs
Evaluation

- **Single judge for the evaluation**

- **Second judge**
  - 340 random examples
    - 100 NCs
    - 20 patterns with the top 10 NCs for each iteration
  - Cohen’s kappa 0.66 – substantial agreement
Evaluation: For Individual Iterations

- **Shown are**
  - ⋆ for NCs: NCs extracted / accuracy
  - ⋆ for patterns: NC-pattern pairs extracted / accuracy / patterns retained

<table>
<thead>
<tr>
<th>Limits</th>
<th>Seeds</th>
<th>Iteration 1</th>
<th>Iteration 2</th>
<th>Iteration 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Patt.</td>
<td>NCs</td>
<td>Patt.</td>
<td>Patterns</td>
</tr>
<tr>
<td></td>
<td>Patt.</td>
<td>NCs</td>
<td>Patterns</td>
<td>NCs</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Loose Bootstrapping</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N=5, M=50</td>
<td>–</td>
<td>18</td>
<td>–</td>
<td>1,144</td>
</tr>
<tr>
<td>N=10, M=20</td>
<td>–</td>
<td>18</td>
<td>–</td>
<td>502</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>63.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strict Bootstrapping</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N=5, M=50</td>
<td>20</td>
<td>18</td>
<td>–</td>
<td>7,011</td>
</tr>
<tr>
<td>N=10, M=20</td>
<td>20</td>
<td>18</td>
<td>–</td>
<td>4,826</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>70.65</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NC-only Strict Bootstrapping</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N=5</td>
<td>20</td>
<td>18</td>
<td>–</td>
<td>7,011</td>
</tr>
<tr>
<td>N=10</td>
<td>20</td>
<td>18</td>
<td>–</td>
<td>4,826</td>
</tr>
</tbody>
</table>
Evaluation: Overall, for all Three Iterations

<table>
<thead>
<tr>
<th>Limits</th>
<th>Extracted &amp; Retained</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NCs</td>
</tr>
<tr>
<td><strong>Loose Bootstrapping</strong></td>
<td></td>
</tr>
<tr>
<td>( N=5, M=50 )</td>
<td>1,662 / 61.67</td>
</tr>
<tr>
<td>( N=10, M=20 )</td>
<td>590  / 61.52</td>
</tr>
<tr>
<td><strong>Strict Bootstrapping</strong></td>
<td></td>
</tr>
<tr>
<td>( N=5, M=50 )</td>
<td>25,375 / 67.42</td>
</tr>
<tr>
<td>( N=10, M=20 )</td>
<td>16,090 / 68.27</td>
</tr>
<tr>
<td><strong>NC-only Strict Bootstrapping</strong></td>
<td></td>
</tr>
<tr>
<td>( N=5 )</td>
<td>205,459 / 69.59</td>
</tr>
<tr>
<td>( N=10 )</td>
<td>100,550 / 70.43</td>
</tr>
</tbody>
</table>
Comparison to Kim & Baldwin (2007)

synonyms/hypernyms/sister words in WordNet → new interpreted NCs

<table>
<thead>
<tr>
<th>Rep.</th>
<th>Iter. 1</th>
<th>Iter. 2</th>
<th>Iter. 3</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syn.</td>
<td>11/81.81</td>
<td>3/66.67</td>
<td>0</td>
<td>14/78.57</td>
</tr>
<tr>
<td>Hyp.</td>
<td>27/85.19</td>
<td>35/77.14</td>
<td>33/66.67</td>
<td>95/75.79</td>
</tr>
<tr>
<td>Sis.</td>
<td>381/82.05</td>
<td>1,736/69.33</td>
<td>17/52.94</td>
<td>2,134/75.12</td>
</tr>
<tr>
<td>All</td>
<td>419/82.58</td>
<td>1,774/71.68</td>
<td>50/62.00</td>
<td>2,243/75.47</td>
</tr>
</tbody>
</table>

- Slightly higher accuracy

- BUT

  - less variety in semantics
  - no paraphrasing verbs
Error Analysis: Syntax

- **Wrong POS tags from tagger**
  
  e.g., in *the statue has such high quality gold (that) demand is ...*
  
  the NC modifier = *demand* vs. *gold*

- **Unlikely nouns in WordNet**
  
  *clear, friendly, single* are nouns in WordNet →
  
  wrong NCs such as *clear cube, friendly team, single chain*

- **Verb-particle constructions**
  
  some particles can be used as nouns in other contexts
  
  e.g., *give back, break down*.
Error Analysis: Semantics

- **Semantic transparency**
  - e.g., *This wine is made from a range of white grapes.*
    NC modifier = *range* vs. *grapes*

- **NC is not \text{MAKE}_2**
  - e.g., *toy box*

- **Extracted nouns do not make a good NC**
  - e.g., *worker work* or *year toy*
Summary

- **Framework for mining NCs, each interpreted by both**
  - an abstract relation
  - a distribution over paraphrasing verbs
- **Bootstrapping with slow degradation in accuracy**
- **More restrictions (strict bootstrapping) yield**
  - better accuracy, but also
  - more mined NCs
Future Work

- **Improve the accuracy**
  - better tests for nouns and NC
  - more restrictions
  - better seeds
  - better relation inventory (MAKE$_2$ is too ambiguous?)

- **Process the remaining relations of Levi (1978)**

- **Release the dataset**