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*)

- **Verbs** (Finin 1980; Vanderwende 1994; Kim & Baldwin 2006; Butnariu & Veale 2008; Nakov & Hearst 2008)
  - ⋆ **malaria mosquito**: *carries, spreads, causes, transmits, brings, has*
  - ⋆ **olive oil**: *comes from, is made from, is derived from*

Abstract Relations vs. Paraphrasing Verbs

	cancer treatment	migraine treatment	wrinkle treatment	herb treatment
treat	+	+	+	−
prevent	+	+	−	−
cure	+	−	−	−
reduce	−	+	+	−
smooth	−	−	+	−
cause	+	−	−	−
contain	−	−	−	+
use	−	−	−	+

- 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**
  - ⋆ $\texttt{MAKE}_2$: *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**
  - ⋆ $\texttt{MAKE}_2$: *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:** $\texttt{MAKE}_2$
  - ⋆ **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

Rel.	Example	Subj/obj	Traditional Name
CAUSE$_1$	*tear gas*	object	causative
CAUSE$_2$	*drug deaths*	subject	causative
HAVE$_1$	*apple cake*	object	possessive/dative
HAVE$_2$	*lemon peel*	subject	possessive/dative
MAKE$_1$	*silkworm*	object	productive/composit.
MAKE$_2$	*snowball*	subject	productive/composit.
USE	*steam iron*	object	instrumental
BE	*soldier ant*	object	essive/appositional
IN	*field mouse*	object	locative
FOR	*horse doctor*	object	purposive/benefactive
FROM	*olive oil*	object	source/ablative
ABOUT	*price war*	object	topic

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 MAKE_2 (from Levi'78); 25-30 annotators
  - ⋆ kept: verbs proposed by $\geq$5 annotators
  - ⋆ removed *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 MAKE$_2$

84 NC-pattern pairs

bronze statue: be made of, be composed of, contain

cable network: consist of, be made of

candy cigarette: be made of, taste like, be made from, look like

chocolate bar: contain, be made of, be made from, taste like, be composed of, consist of

copper coin: be made of, be made from, contain, be composed of

daisy chain: be made of, be made from, contain, consist of

glass eye: be made of, be composed of, be made from

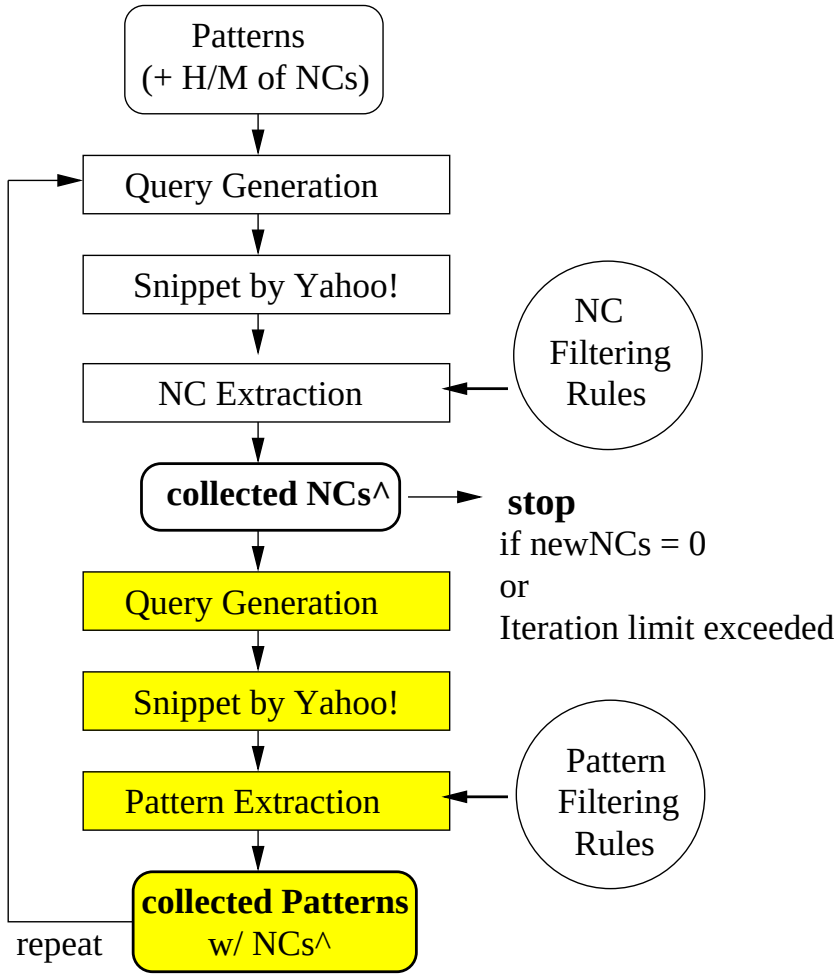
...

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



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 *" (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**
  - ⋆ **bronze 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

Limits	**Seeds**		**Iteration 1**		**Iteration 2**		**Iteration 3**	
	Patt.	NCs	Patt.	NCs	Patterns	NCs	Patterns	NCs
Loose Bootstrapping								
N=5, M=50	–	18	–	1,144 / 63.11	1,136 / 64.44 / 9	390 / 58.72	201 / 70.00 / 3	128 / 57.03
N=10, M=20	–	18	–	502 / 61.55	294 / 62.50 / 8	78 / 60.26	22 / 90.00 / 1	10 / 70.00
Strict Bootstrapping								
N=5, M=50	20	18	–	7,011 / 70.65	5,312 / 74.00 / 10	11,214 / 67.15	4,448 / 60.00 / 6	7,150 / 64.69
N=10, M=20	20	18	–	4,826 / 71.26	2,838 / 79.38 / 10	7,371 / 67.26	2,188 / 78.33 / 6	3,893 / 66.48
NC-only Strict Bootstrapping								
N=5	20	18	–	7,011 / 70.65	–	198,448 / 69.55	–	–
N=10	20	18	–	4,826 / 71.26	–	95,524 / 70.59	–	–

Evaluation: Overall, for all Three Iterations

Limits	Extracted & Retained		
	NCs	Patterns	Patt.+NC
Loose Bootstrapping			
N=5, M=50	1,662 / 61.67	12 / 65.83	1,337
N=10, M=20	590 / 61.52	9 / 65.56	316
Strict Bootstrapping			
N=5, M=50	25,375 / 67.42	16 / 71.43	9,760
N=10, M=20	16,090 / 68.27	16 / 78.98	5,026
NC-only Strict Bootstrapping			
N=5	205,459 / 69.59	–	–
N=10	100,550 / 70.43	–	–

Comparison to Kim & Baldwin (2007)

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

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

- 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 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**