

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

Su Nam Kim

Computer Science & Software Engineering
University of Melbourne
Melbourne, VIC 3010
Australia
snkim@csse.unimelb.edu.au

Preslav Nakov

Department of Computer Science
National University of Singapore
13 Computing Drive
Singapore 117417
nakov@comp.nus.edu.sg

Abstract

Responding to the need for semantic lexical resources in natural language processing applications, we examine methods to acquire noun compounds (NCs), e.g., *orange juice*, together with suitable fine-grained semantic interpretations, e.g., *squeezed from*, which are directly usable as paraphrases. We employ bootstrapping and web statistics, and utilize the relationship between NCs and paraphrasing patterns to jointly extract NCs and such patterns in multiple alternating iterations. In evaluation, we found that having one compound noun fixed yields both a higher number of semantically interpreted NCs and improved accuracy due to stronger semantic restrictions.

1 Introduction

Noun compounds (NCs) such as *malaria mosquito* and *colon cancer tumor suppressor protein* are challenging for text processing since the relationship between the nouns they are composed of is implicit. NCs are abundant in English and understanding their semantics is important in many natural language processing (NLP) applications. For example, a question answering system might need to know whether *protein acting as a tumor suppressor* is a good paraphrase for *tumor suppressor protein*. Similarly, a machine translation system facing the unknown noun compound *Geneva headquarters* might translate it better if it could first paraphrase it as *Geneva headquarters of the WTO*. Given a query for “*migraine treatment*”, an information retrieval system could use paraphrasing verbs like *relieve* and *prevent* for query expansion and result ranking.

Most work on noun compound interpretation has focused on two-word NCs. There have been two general lines of research: the first one derives the NC semantics from the semantics of the nouns it is made of (Rosario and Hearst, 2002; Moldovan et al., 2004; Kim and Baldwin, 2005; Girju, 2007; Séaghdha, 2009; Tratz and Hovy, 2010), while the second one models the relationship between the nouns directly (Vanderwende, 1994; Lapata, 2002; Kim and Baldwin, 2006; Nakov and Hearst, 2006; Nakov and Hearst, 2008; Butnariu and Veale, 2008).

In either case, the semantics of an NC is typically expressed by an abstract relation like CAUSE (e.g., *malaria mosquito*), SOURCE (e.g., *olive oil*), or PURPOSE (e.g., *migraine drug*), coming from a small fixed inventory. Some researchers however, have argued for a more fine-grained, even infinite, inventory (Finin, 1980). Verbs are particularly useful in this respect and can capture elements of the semantics that the abstract relations cannot. For example, while most NCs expressing MAKE, can be paraphrased by common patterns like *be made of* and *be composed of*, some NCs allow more specific patterns, e.g., *be squeezed from* for *orange juice*, and *be topped with* for *bacon pizza*.

Recently, the idea of using fine-grained paraphrasing verbs for NC semantics has been gaining popularity (Butnariu and Veale, 2008; Nakov, 2008b); there has also been a related shared task at SemEval-2010 (Butnariu et al., 2010). This interest is partly driven by practicality: verbs are directly usable as paraphrases. Still, abstract relations remain dominant since they offer a more natural generalization, which is useful for many NLP applications.

One good contribution to this debate would be a direct study of the relationship between fine-grained and coarse-grained relations for NC interpretation. Unfortunately, the existing datasets do not allow this since they are tied to one particular granularity; moreover, they only contain a few hundred NCs. Thus, our objective is to build a large-scale dataset of hundreds of thousands of NCs, each interpreted (1) by an abstract semantic relation and (2) by a set of paraphrasing verbs. Having such a large dataset would also help the overall advancement of the field.

Since there is no universally accepted abstract relation inventory in NLP, and since we are interested in NC semantics from both a theoretical and a practical viewpoint, we chose the set of abstract relations proposed in the theory of Levi (1978), which is dominant in theoretical linguistics and has been also used in NLP (Nakov and Hearst, 2008).

We use a two-step algorithm to jointly harvest NCs and patterns (verbs and prepositions) that interpret them for a given abstract relation. First, we extract NCs using a small number of seed patterns from a given abstract relation. Then, using the extracted NCs, we harvest more patterns. This is repeated until no new NCs and patterns can be extracted or for a pre-specified number of iterations. Our approach combines pattern-based extraction and bootstrapping, which is novel for NC interpretation; however, such combinations have been used in other areas, e.g., named entity recognition (Riloff and Jones, 1999; Thelen and Riloff, 2002; Curran et al., 2007; McIntosh and Curran, 2009).

The remainder of the paper is organized as follows: Section 2 gives an overview of related work, Section 3 motivates our semantic representation, Sections 4, 5, and 6 explain our method, dataset and experiments, respectively, Section 7 discusses the results, Section 8 provides error analysis, and Section 9 concludes with suggestions for future work.

2 Related Work

As we mentioned above, the implicit relation between the two nouns forming a noun compound can often be expressed overtly using verbal and prepositional paraphrases. For example, *student loan* is “loan given to a student”, while *morning tea* can be paraphrased as “tea in the morning”.

Thus, many NLP approaches to NC semantics have used verbs and prepositions as a fine-grained semantic representation or as features when predicting coarse-grained abstract relations. For example, Vanderwende (1994) associated verbs extracted from definitions in an online dictionary with abstract relations. Lauer (1995) expressed NC semantics using eight prepositions. Kim and Baldwin (2006) predicted abstract relations using verbs as features. Nakov and Hearst (2008) proposed a fine-grained NC interpretation using a distribution over Web-derived verbs, prepositions and coordinating conjunctions; they also used this distribution to predict coarse-grained abstract relations. Butnariu and Veale (2008) adopted a similar fine-grained verb-centered approach to NC semantics. Using a distribution over verbs as a semantic interpretation was also carried out in a recent challenge: SemEval-2010 Task 9 (Butnariu et al., 2009; Butnariu et al., 2010).

In noun compound interpretation, verbs and prepositions can be seen as patterns connecting the two nouns in a paraphrase. Similar pattern-based approaches have been popular in information extraction and ontology learning. For example, Hearst (1992) extracted hyponyms using patterns such as *X, Y, and/or other Zs*, where *Z* is a hypernym of *X* and *Y*. Berland and Charniak (1999) used similar patterns to extract meronymy (part-whole) relations, e.g., *parts/NNS of/IN wholes/NNS* matches *basements of buildings*. Unfortunately, matches are rare, which makes it difficult to build large semantic inventories. In order to overcome data sparseness, pattern-based approaches are often combined with bootstrapping. For example, Riloff and Jones (1999) used a multi-level bootstrapping algorithm to learn both a semantic lexicon and extraction patterns, e.g., *owned by X* extracts *COMPANY* and *facilities in X* extracts *LOCATION*. That is, they learned semantic lexicons using extraction patterns, and then, alternatively, they extracted new patterns using these lexicons. They also introduced a second level of bootstrapping to retain the most reliable examples only. While the method enables the extraction of large lexicons, its quality degrades rapidly, which makes it impossible to run for too many iterations. Recently, Curran et al. (2007) and McIntosh and Curran (2009) proposed ways to control degradation using simultaneous learning and weighting.

Bootstrapping has been applied to noun compound extraction as well. For example, Kim and Baldwin (2007) used it to produce a large number of semantically interpreted noun compounds from a small number of seeds. In each iteration, the method replaced one component of an NC with its synonyms, hypernyms and hyponyms to generate a new NC. These new NCs were further filtered based on their semantic similarity with the original NC. While the method acquired a large number of noun compounds without significant semantic drifting, its accuracy degraded rapidly after each iteration. More importantly, the variation of the sense pairs was limited since new NCs had to be semantically similar to the original NCs.

Recently, Kozareva and Hovy (2010) combined patterns and bootstrapping to learn the selectional restrictions for various semantic relations. They used patterns involving the coordinating conjunction *and*, e.g., “* *and John fly to* *”, and learned arguments such as *Mary/Tom* and *France/New York*. Unlike in NC interpretation, it is not necessary for their arguments to form an NC, e.g., *Mary France* and *France Mary* are not NCs. Rather, they were interested in building a semantic ontology with a pre-defined set of semantic relations, similar to YAGO (Suchanek et al., 2007), where the pattern *work for* would have arguments like *a company/UNICEF*.

3 Semantic Representation

Inspired by (Finin, 1980), Nakov and Hearst (2006) and (Nakov, 2008b) proposed that NC semantics is best expressible using paraphrases involving verbs and/or prepositions. For example, *bronze statue* is a statue that *is made of*, *is composed of*, *consists of*, *contains*, *is of*, *is*, *is handcrafted from*, *is dipped in*, *looks like* bronze. They further proposed that selecting one such paraphrase is not enough and that multiple paraphrases are needed for a fine-grained representation. Finally, they observed that not all paraphrases are equally good (e.g., *is made of* is arguably better than *looks like* or *is dipped in* for MAKE), and thus proposed that the semantics of a noun compound should be expressed as a *distribution* over multiple possible paraphrases. This line of research was later adopted by SemEval-2010 Task 9 (Butnariu et al., 2010).

It easily follows that the semantics of abstract relations such as MAKE that can hold between the nouns in an NC can be represented in the same way: as a distribution over paraphrasing verbs and prepositions. Note, however, that some NCs are paraphrasable by more specific verbs that do not necessarily support the target abstract relation. For example, *malaria mosquito*, which expresses CAUSE, can be paraphrased using verbs like *carry*, which do not imply direct causation. Thus, while we will be focusing on extracting NCs for a particular abstract relation, we are interested in building semantic representations that are specific for these NCs and do not necessarily apply to *all* instances of that relation.

Traditionally, the semantics of a noun compound have been represented as an abstract relation drawn from a small closed set. Unfortunately, no such set is universally accepted, and mapping between sets has proven challenging (Girju et al., 2005). Moreover, being both abstract and limited, such sets capture only part of the semantics; often multiple meanings are possible, and sometimes none of the pre-defined ones suits a given example. Finally, it is unclear how useful these sets are since researchers have often fallen short of demonstrating practical uses.

Arguably, verbs have more expressive power and are more suitable for semantic representation: there is an infinite number of them (Downing, 1977), and they can capture fine-grained aspects of the meaning. For example, while both *wrinkle treatment* and *migraine treatment* express the same abstract relation TREATMENT-FOR-DISEASE, fine-grained differences can be revealed using verbs, e.g., *smooth* can paraphrase the former, but not the latter.

In many theories, verbs play an important role in NC derivation (Levi, 1978). Moreover, speakers often use verbs to make the hidden relation between the noun in a noun compound overt. This allows for simple extraction and for straightforward use in NLP tasks like textual entailment (Tatu and Moldovan, 2005) and machine translation (Nakov, 2008a).

Finally, a single verb is often not enough, and the meaning is better approximated by a collection of verbs. For example, while *malaria mosquito* expresses CAUSE (and is paraphrasable using *cause*), further aspects of the meaning can be captured with more verbs, e.g., *carry*, *spread*, *be responsible for*, *be infected with*, *transmit*, *pass on*, etc.

4 Method

We harvest noun compounds expressing some target abstract semantic relation (in the experiments below, this is Levi’s MAKE₂), starting from a small number of initial seed patterns: paraphrasing verbs and/or prepositions. Optionally, we might also be given a small number of noun compounds that instantiate the target abstract relation. We then learn more noun compounds and patterns for the relation by alternating between the following two bootstrapping steps, using the Web as a corpus. First, we extract more noun compounds that are paraphrasable with the available patterns (see Section 4.1). We then look for new patterns that can paraphrase the newly-extracted noun compounds (see Section 4.2). These two steps are repeated until no new noun compounds can be extracted or until a pre-determined number of iterations has been reached. A schematic description of the algorithm is shown in Figure 1.

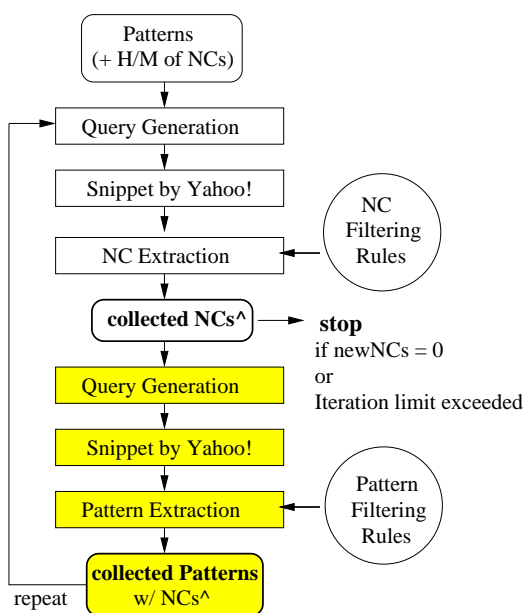


Figure 1: Our bootstrapping algorithm.

4.1 Bootstrapping Step 1: Noun Compound Extraction

Given a list of patterns (verbs and/or prepositions), we mine the Web to extract noun compounds that match these patterns. We experiment with the following three bootstrapping strategies for this step:

- **Loose bootstrapping** uses the available patterns and imposes no further restrictions.
- **Strict bootstrapping** requires that, in addition to the patterns themselves, some noun compounds matching each pattern be made available as well. A pattern is only instantiated in the context of either the head or the modifier of a noun compound that is known to match it.
- **NC-only strict bootstrapping** is a stricter version of *strict bootstrapping*, where the list of patterns is limited to the initial seeds.

Below we describe each of the sub-steps of the NC extraction process: query generation, snippet harvesting, and noun compound acquisition & filtering.

4.1.1 Query Generation

We generate generalized exact-phrase queries to be used in a Web search engine (we use *Yahoo!*):

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

where PATTERN is an inflected form of a verb, MOD and HEAD are inflected forms the modifier and the head of a noun compound that is paraphrasable by the pattern, *that* is the word *that*, and * is the search engine’s star operator.

We use the first pattern for *loose bootstrapping* and the other two for both *strict bootstrapping* and *NC-only strict bootstrapping*.

Note that the above queries are generalizations of the actual queries we use against the search engine. In order to instantiate these generalizations, we further generate the possible inflections for the verbs and the nouns involved. For nouns, we produce singular and plural forms, while for verbs, we vary not only the number (singular and plural), but also the tense (we allow present, past, and present perfect). When inflecting verbs, we distinguish between active verb forms like *consist of* and passive ones like *be made from* and we treat them accordingly. Overall, in the case of *loose bootstrapping*, we generate about 14 and 20 queries per pattern for active and passive patterns, respectively, while for *strict bootstrapping* and *NC-only strict bootstrapping*, the instantiations yield about 28 and 40 queries for active and passive patterns, respectively.

For example, given the seed *be made of*, we could generate "** that were made of **". If we are further given the NC *orange juice*, we could also produce "*juice that was made of **" and "** that is made of oranges*".

4.1.2 Snippet Extraction

We execute the above-described instantiations of the generalized queries against a search engine as exact phrase queries, and, for each one, we collect the snippets for the top 1,000 returned results.

4.1.3 NC Extraction and Filtering

Next, we process the snippets returned by the search engine and we acquire potential noun compounds from them. Then, in each snippet, we look for an instantiation of the pattern used in the query and we try to extract suitable noun(s) that occupy the position(s) of the ***.

For *loose bootstrapping*, we extract two nouns, one from each end of the matched pattern, while for *strict bootstrapping* and for *NC-only strict bootstrapping*, we only extract one noun, either preceding or following the pattern, since the other noun is already fixed. We then lemmatize the extracted noun(s) and we form NC candidates from the two arguments of the instantiated pattern, taking into account whether the pattern is active or passive.

Due to the vast number of snippets we have to process, we decided not to use a syntactic parser or a part-of-speech (POS) tagger¹; thus, we use heuristic rules instead. We extract "phrases" using simple indicators such as punctuation (e.g., comma, period), coordinating conjunctions² (e.g., *and*, *or*), prepositions (e.g., *at*, *of*, *from*), subordinating conjunctions (e.g., *because*, *since*, *although*), and relative pronouns (e.g., *that*, *which*, *who*). We then extract the nouns from these phrases, we lemmatize them using WordNet, and we form a list of NC candidates.

While the above heuristics work reasonably well in practice, we perform some further filtering, removing all NC candidates for which one or more of the following conditions are met:

¹In fact, POS taggers and parsers are unreliable for Web-derived snippets, which often represent parts of sentences and contain errors in spelling, capitalization and punctuation.

²Note that filtering the arguments using such indicators indirectly subsumes the pattern "*X PATTERN Y and*" proposed in (Kozareva and Hovy, 2010).

1. the candidate NC is one of the seed examples or has been extracted on a previous iteration;
2. the head and the modifier are the same;
3. the head or the modifier are not both listed as nouns in WordNet (Fellbaum, 1998);
4. the candidate NC occurs less than 100 times in the *Google Web 1T 5-gram corpus*³;
5. the NC is extracted less than *N* times (we tried 5 and 10) in the context of the pattern for all instantiations of the pattern.

4.2 Bootstrapping Step 2: Pattern Extraction

This is the second step of our bootstrapping algorithm as shown on Figure 1. Given a list of noun compounds, we mine the Web to extract patterns: verbs and/or prepositions that can paraphrase each NC. The idea is to turn the NC's pre-modifier into a post-modifying relative clause and to collect the verbs and prepositions that are used in such clauses. Below we describe each of the sub-steps of the NC extraction process: query generation, snippet harvesting, and NC extraction & filtering.

4.2.1 Query Generation

The process of extraction starts with exact-phrase queries issued against a Web search engine (again *Yahoo!*) using the following generalized pattern:

"HEAD THAT? * MOD"

where MOD and HEAD are inflected forms of NC's modifier and head, respectively, THAT? stands for *that*, *which*, *who* or the empty string, and * stands for 1-6 instances of search engine's star operator.

For example, given *orange juice*, we could generate queries like "*juice that * oranges*", "*juices which * * * * * oranges*", and "*juices * * * orange*".

4.2.2 Snippet Extraction

The same as in Section 4.1.2 above.

³<http://www ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13>

4.2.3 Pattern Extraction and Filtering

We split the extracted snippets into sentences, and filter out all incomplete ones and those that do not contain (a possibly inflected version of) the target nouns. We further make sure that the word sequence following the second mentioned target noun is non-empty and contains at least one non-noun, thus ensuring the snippet includes the entire noun phrase. We then perform shallow parsing, and we extract all verb forms, and the following preposition, between the target nouns. We allow for adjectives and participles to fall between the verb and the preposition but not nouns; we further ignore modal verbs and auxiliaries, but we retain the passive *be*, and we make sure there is exactly one verb phrase between the target nouns. Finally, we lemmatize the verbs to form the patterns candidates, and we apply the following pattern selection rules:

1. we filter out all patterns that were provided as initial seeds or were extracted previously;
2. we select the top 20 most frequent patterns;
3. we filter out all patterns that were extracted less than N times (we tried 5 and 10) and with less than M NCs per pattern (we tried 20 and 50).

5 Target Relation and Seed Examples

As we mentioned above, we use the inventory of abstract relations proposed in the popular theoretical linguistics theory of Levi (1978). In this theory, noun compounds are derived from underlying relative clauses or noun phrase complement constructions by means of two general processes: predicate deletion and predicate nominalization. Given a two-argument predicate, *predicate deletion* removes that predicate, but retains its arguments to form an NC, e.g., *pie made of apples* → *apple pie*. In contrast, *predicate nominalization* creates an NC whose head is a nominalization of the underlying predicate and whose modifier is either the subject or the object of that predicate, e.g., *The President refused General MacArthur's request.* → *presidential refusal*.

According to Levi, predicate deletion can be applied to abstract predicates, whose semantics can be roughly approximated using five paraphrasing verbs (CAUSE, HAVE, MAKE, USE, and BE) and four prepositions (IN, FOR, FROM, and ABOUT).

Typically, in predicate deletion, the modifier is derived from the object of the underlying relative clause; however, the first three verbs also allow for it to be derived from the subject. Levi expresses the distinction using indexes. For example, *music box* is MAKE₁ (object-derived), i.e., the *box makes music*, while *chocolate bar* is MAKE₂ (subject-derived), i.e., *the bar is made of chocolate* (note the passive).

Due to time constraints, we focused on one relation of Levi's, MAKE₂, which is among the most frequent relations an NC can express and is present in some form in many relation inventories (Warren, 1978; Barker and Szpakowicz, 1998; Rosario and Hearst, 2001; Nastase and Szpakowicz, 2003; Girju et al., 2005; Girju et al., 2007; Girju et al., 2009; Hendrickx et al., 2010; Tratz and Hovy, 2010).

In Levi's theory, MAKE₂ means that the head of the noun compound is made up of or is a product of its modifier. There are three subtypes of this relation (we do not attempt to distinguish between them):

- (a) the modifier is a unit and the head is a configuration, e.g., *root system*;
- (b) the modifier represents a material and the head is a mass or an artefact, e.g., *chocolate bar*;
- (c) the head represents human collectives and the modifier specifies their membership, e.g., *worker teams*.

There are 20 instances of MAKE₂ in the appendix of (Levi, 1978), and we use them all as *seed NCs*. As *seed patterns*, we use a subset of the human-proposed paraphrasing verbs and prepositions corresponding to these 20 NCs in the dataset in (Nakov, 2008b), where each NC is paraphrased by 25-30 annotators. For example, for *chocolate bar*, we find the following list of verbs (the number of annotators who proposed each verb is shown in parentheses):

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).

Seed NCs: 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
Seed patterns: 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

Table 1: Our seed examples: 20 noun compounds and 18 verb patterns.

As we can see, the most frequent patterns are of highest quality, e.g., *be made of* (16), while the less frequent ones can be wrong, e.g., *serve* (1). Therefore, we filtered out all verbs that were proposed less than five times with the 20 seed NCs. We further removed the verb *be*, which is too general, thus ending up with 18 seed patterns. Note that some patterns can paraphrase multiple NCs: the total number of seed NC-pattern pairs is 84.

The seed NCs and patterns are shown in Table 1. While some patterns, e.g., *taste like* do not express the target relation MAKE_2 , we kept them anyway since they were proposed by several human annotators and since they do express the fine-grained semantics of some particular instances of that relation; thus, we thought they might be useful, even for the general relation. For example, *taste like* has been proposed 8 times for *candy cigarette*, 7 times for *chocolate bar*, and 2 times for *sugar cube*, and thus it clearly correlates well with some seed examples, even if it does not express MAKE_2 in general.

6 Experiments and Evaluation

Using the NCs and patterns in Table 1 as initial seeds, we ran our algorithm for three iterations of *loose bootstrapping* and *strict bootstrapping*, and for two iterations of *NC-only strict bootstrapping*. We only performed up to three iterations because of the huge number of noun compounds extracted for *NC-only strict bootstrapping* (which we only ran for two iterations) and because of the low number of new NCs extracted by *loose bootstrapping* on iteration 3. While we could have run *strict bootstrapping* for more iterations, we opted for a comparable number of iterations for all three methods.

Examples of noun compounds that we have extracted are *bronze bell* (be made of, be made from) and *child team* (be composed of, include). Example patterns are *be filled with* (cotton bag, water cup) and *use* (water sculpture, wood statue).

Limits (see 4.2.3)	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	–	–

Table 2: Total number and accuracy in % for NCs, patterns and NC-pattern pairs extracted and retained for each of the three methods over all iterations.

Tables 2 and 3 show the overall results. As we mentioned in section 4.2.3, at each iteration, we filtered out all patterns that were extracted less than N times or with less than M NCs. Note that we only used the 10 most frequent NCs per pattern as NC seeds for NC extraction in the next iteration of *strict bootstrapping* and *NC-only strict bootstrapping*. Table 3 shows the results for two value combinations of $(N;M)$: (5;50) and (10;20). Note also that if some NC was extracted by several different patterns, it was only counted once. Patterns are subject to particular NCs, and thus we show (1) the number of patterns extracted with all NCs, i.e., unique NC-pattern pairs, (2) the accuracy of these pairs,⁴ and (3) the number of unique patterns retained after filtering, which will be used to extract new noun compounds on the second step of the current iteration.

⁴One of the reviewers suggested that evaluating the accuracy of NC-pattern pairs could potentially conceal some of the drift of our algorithm. For example, while *water cup / be filled with* is a correct NC-pattern pair, *water cup* is incorrect for MAKE_2 ; it is probably an instance of Levi’s FOR. Thus, the same bootstrapping technique evaluated against a fixed set of semantic relations (which is the more traditional approach) could arguably show bootstrapping going “off the rails” more quickly than what we observe here. However, our goal, as stated in Section 3, is to find NC-specific paraphrases, and our evaluation methodology is more adequate with respect to this goal.

Limits (see 4.2.3)	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	–	–

Table 3: *Evaluation results for up to three iterations.* For NCs, we show the number of unique NCs extracted and their accuracy in %. For patterns, we show the number of unique NC-pattern pairs extracted, their accuracy in %, and the number of unique patterns retained and used to extract NCs on the second step of the current iteration. The first column shows the pattern filtering thresholds used (see Section 4.2.3 for details).

The above accuracies were calculated based on human judgments by an experienced, well-trained annotator. We also hired a second annotator for a small subset of the examples.

For NCs, the first annotator judged whether each NC is an instance of MAKE₂. All NCs were judged, except for iteration 2 of *NC-only strict bootstrapping*, where their number was prohibitively high and only the most frequent noun compounds extracted for each modifier and for each head were checked: 9,004 NCs for $N=5$ and 4,262 NCs for $N=10$.

For patterns, our first annotator judged the correctness of the unique NC-pattern pairs, i.e., whether the NC is paraphrasable with the target pattern. Given the large number of NC-pattern pairs, the annotator only judged patterns with their top 10 most frequent NCs. For example, if there were 5 patterns extracted, then the NC-pattern pairs to be judged would be no more than $5 \times 10 = 50$.

Our second annotator judged 340 random examples: 100 NCs and 20 patterns with their top 10 NCs for each iteration. The Cohen’s kappa (Cohen, 1960) between the two annotators is .66 (85% initial agreement), which corresponds to substantial agreement (Landis and Koch, 1977).

7 Discussion

Tables 2 and 3 show that fixing one of the two nouns in the pattern, as in *strict bootstrapping* and *NC-only strict bootstrapping*, yields significantly higher accuracy (χ^2 test) for both NC and NC-pattern pair extraction compared to *loose bootstrapping*.

The accuracy for *NC-only strict bootstrapping* is a bit higher than for *strict bootstrapping*, but the actual differences are probably smaller since the evaluation of the former on iteration 2 was done for the most frequent NCs, which are more accurate.

Note that the number of extracted NCs is much higher with the strict methods because of the higher number of possible instantiations of the generalized query patterns. For *NC-only strict bootstrapping*, the number of extracted NCs grows exponentially since the number of patterns does not diminish as in the other two methods. The number of extracted patterns is similar for the different methods since we select no more than 20 of them per iteration.

Overall, the accuracy for all methods decreases from one iteration to the next since errors accumulate; still, the degradation is slow. Note also the exception of *loose bootstrapping* on iteration 3.

Comparing the results for $N=5$ and $N=10$, we can see that, for all three methods, using the latter yields a sizable drop in the number of extracted NCs and NC-pattern pairs; it also tends to yield a slightly improved accuracy. Note, however, the exception of *loose bootstrapping* for the first two iterations, where the less restrictive $N=5$ is more accurate.

As a comparison, we implemented the method of Kim and Baldwin (2007), which generates new semantically interpreted NCs by replacing either the head or the modifier of a seed NC with suitable synonyms, hypernyms and sister words from WordNet, followed by similarity filtering using WordNet : : Similarity (Pedersen et al., 2004).

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

Table 4: Number of extracted noun compounds and accuracy in % for the method of Kim and Baldwin (2007). The abbreviations *Syn.*, *Hyp.*, and *Sis.* indicate using synonyms, hypernyms, and sister words, respectively.

The results for three bootstrapping iterations using the same list of 20 initial seed NCs as in our previous experiments, are shown in Table 4. We can see that the overall accuracy of their method is slightly better than ours. Note, however, that our method acquired a much larger number of NCs, while allowing more variety in the NC semantics. Moreover, for each extracted noun compound, we also generated a list of fine-grained paraphrasing verbs.

8 Error Analysis

Below we analyze the errors of our method.

Many problems were due to wrong POS assignment. For example, on Step 2, because of the omission of *that* in “*the statue has such high quality gold (that) demand is ...*”, *demand* was tagged as a noun and thus extracted as an NC modifier instead of *gold*. The problem also arose on Step 1, where we used WordNet to check whether the NC candidates were composed of two nouns. Since words like *clear*, *friendly*, and *single* are listed in WordNet as nouns (which is possible in some contexts), we extracted wrong NCs such as *clear cube*, *friendly team*, and *single chain*. There were similar issues with verb-particle constructions since some particles can be used as nouns as well, e.g., *give back*, *break down*.

Some errors were due to semantic transparency issues, where the syntactic and the semantic head of a target NP were mismatched (Fillmore et al., 2002; Fontenelle, 1999). For example, from the sentence “*This wine is made from a range of white grapes.*”, we would extract *range* rather than *grapes* as the potential modifier of *wine*.

In some cases, the NC-pattern pair was correct, but the NC did not express the target relation, e.g., while *contain* is a good paraphrase for *toy box*, the noun compound itself is not an instance of MAKE₂.

There were also cases where the pair of extracted nouns did not make a good NC, e.g., *worker work* or *year toy*. Note that this is despite our checking that the candidate NC occurred at least 100 times in the *Google Web 1T 5-gram corpus* (see Section 4.1.3). We hypothesized that such bad NCs would tend to have a low collocation strength. We tested this hypothesis using the Dice coefficient, calculated using the *Google Web 1T 5-gram corpus*. Figure 2 shows a plot of the NC accuracy vs. collocation strength for *strict bootstrapping* with $N=5$, $M=50$ for all three iterations (the results for the other experiments show a similar trend). We can see that the accuracy improves slightly as the collocation strength increases: compare the left and the right ends of the graph (the results are mixed in the middle though).

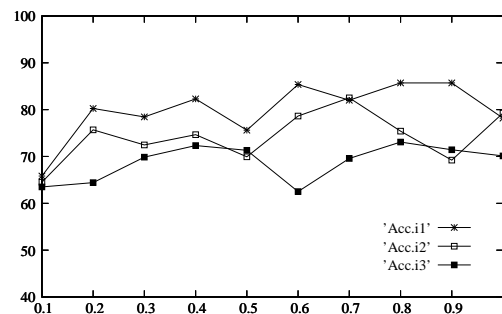


Figure 2: NC accuracy vs. collocation strength.

9 Conclusion and Future Work

We have presented a framework for building a very large dataset of noun compounds expressing a given target abstract semantic relation. For each extracted noun compound, we generated a corresponding fine-grained semantic interpretation: a frequency distribution over suitable paraphrasing verbs.

In future work, we plan to apply our framework to the remaining relations in the inventory of Levi (1978), and to release the resulting dataset to the research community. We believe that having a large-scale dataset of noun compounds interpreted with both fine- and coarse-grained semantic relations would be an important contribution to the debate about which representation is preferable for different tasks. It should also help the overall advancement of the field of noun compound interpretation.

Acknowledgments

This research is partially supported (for the second author) by the *SmartBook project*, funded by the Bulgarian National Science Fund under Grant D002-111/15.12.2008.

We would like to thank the anonymous reviewers for their detailed and constructive comments, which have helped us improve the paper.

References

- Ken Barker and Stan Szpakowicz. 1998. Semi-automatic recognition of noun modifier relationships. In *Proceedings of the 17th International Conference on Computational Linguistics*, pages 96–102.
- Matthew Berland and Eugene Charniak. 1999. Finding parts in very large corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, ACL '99, pages 57–64.
- Cristina Butnariu and Tony Veale. 2008. A concept-centered approach to noun-compound interpretation. In *Proceedings of the 22nd International Conference on Computational Linguistics*, COLING '08, pages 81–88.
- Cristina Butnariu, Su Nam Kim, Preslav Nakov, Diarmuid Ó Séaghdha, Stan Szpakowicz, and Tony Veale. 2009. Semeval-2010 task 9: The interpretation of noun compounds using paraphrasing verbs and prepositions. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, SEW '09, pages 100–105.
- Cristina Butnariu, Su Nam Kim, Preslav Nakov, Diarmuid Ó Séaghdha, Stan Szpakowicz, and Tony Veale. 2010. SemEval-2010 task 9: The interpretation of noun compounds using paraphrasing verbs and prepositions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval-2, pages 39–44.
- Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 1(20):37–46.
- James R. Curran, Tara Murphy, and Bernhard Scholz. 2007. Minimising semantic drift with mutual exclusion bootstrapping. In *Proceedings of the Conference of the Pacific Association for Computational Linguistics*, PAFLING '07, pages 172–180.
- Pamela Downing. 1977. On the creation and use of English compound nouns. *Language*, 53:810–842.
- Christiane Fellbaum, editor. 1998. *WordNet, An Electronic Lexical Database*. MIT Press, Cambridge, Massachusetts, USA.
- Charles J. Fillmore, Collin F. Baker, and Hiroaki Sato. 2002. Seeing arguments through transparent structures. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, volume III of *LREC '02*, pages 787–791.
- Timothy Wilking Finin. 1980. *The semantic interpretation of compound nominals*. Ph.D. thesis, University of Illinois at Urbana-Champaign, Champaign, IL, USA. AAI8026491.
- Thierry Fontenelle. 1999. Semantic resources for word sense disambiguation: a sine qua non. *Linguistica e Filologia*, 9:25–41.
- Roxana Girju, Dan Moldovan, Marta Tatu, and Daniel Antohe. 2005. On the semantics of noun compounds. *Computer Speech and Language*, 19(44):479–496.
- Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. 2007. Semeval-2007 task 04: Classification of semantic relations between nominals. In *Proceedings of the 4th Semantic Evaluation Workshop*, SemEval-1, pages 13–18.
- Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. 2009. Classification of semantic relations between nominals. *Language Resources and Evaluation*, 43(2):105–121.
- Roxana Girju. 2007. Improving the interpretation of noun phrases with cross-linguistic information. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, ACL '07, pages 568–575.
- Marti Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics*, COLING '92, pages 539–545.
- Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval-2, pages 33–38.
- Su Nam Kim and Timothy Baldwin. 2005. Automatic interpretation of compound nouns using WordNet similarity. In *Proceedings of 2nd International Joint Conference on Natural Language Processing*, IJCNLP '05, pages 945–956.
- Su Nam Kim and Timothy Baldwin. 2006. Interpreting semantic relations in noun compounds via verb semantics. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics and 21st International Conference on Computational Linguistics*, ACL-COLING '06, pages 491–498.

- Su Nam Kim and Timothy Baldwin. 2007. Interpreting noun compounds using bootstrapping and sense collocation. In *Proceedings of Conference of the Pacific Association for Computational Linguistics, PACLING '07*, pages 129–136.
- Zornitsa Kozareva and Eduard Hovy. 2010. Learning arguments and supertypes of semantic relations using recursive patterns. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL '10*, pages 1482–1491.
- Richard J. Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.
- Maria Lapata. 2002. The disambiguation of nominalizations. *Computational Linguistics*, 28(3):357–388.
- Mark Lauer. 1995. *Designing Statistical Language Learners: Experiments on Noun Compounds*. Ph.D. thesis, Dept. of Computing, Macquarie University, Australia.
- Judith Levi. 1978. *The Syntax and Semantics of Complex Nominals*. Academic Press, New York, USA.
- Tara McIntosh and James Curran. 2009. Reducing semantic drift with bagging and distributional similarity. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, ACL-IJCNLP '09*, pages 396–404.
- Dan Moldovan, Adriana Badulescu, Marta Tatu, Daniel Antohe, and Roxana Girju. 2004. Models for the semantic classification of noun phrases. In *Proceedings of the HLT-NAACL'04 Workshop on Computational Lexical Semantics*, pages 60–67.
- Preslav Nakov and Marti A. Hearst. 2006. Using verbs to characterize noun-noun relations. In *Proceedings of the 12th International Conference on Artificial Intelligence: Methodology, Systems, and Applications, AIMSA '06*, pages 233–244.
- Preslav Nakov and Marti Hearst. 2008. Solving relational similarity problems using the web as a corpus. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, ACL '08*, pages 452–460.
- Preslav Nakov. 2008a. Improved Statistical Machine Translation Using Monolingual Paraphrases. In *Proceedings of the 18th European Conference on Artificial Intelligence, ECAI '08*, pages 338–342.
- Preslav Nakov. 2008b. Noun compound interpretation using paraphrasing verbs: Feasibility study. In *Proceedings of the 13th international conference on Artificial Intelligence: Methodology, Systems, and Applications, AIMSA '08*, pages 103–117.
- Vivi Nastase and Stan Szpakowicz. 2003. Exploring noun-modifier semantic relations. In *Proceedings of the 5th International Workshop on Computational Semantics*, pages 285–301.
- Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. Wordnet::similarity - measuring the relatedness of concepts. In *Proceedings of the Nineteenth National Conference on Artificial Intelligence, AAAI '04*, pages 1024–1025.
- Ellen Riloff and Rosie Jones. 1999. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence, AAAI '99*, pages 474–479.
- Barbara Rosario and Marti Hearst. 2001. Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy. In *Proceedings of the 6th Conference on Empirical Methods in Natural Language Processing, EMNLP '01*, pages 82–90.
- Barbara Rosario and Marti Hearst. 2002. The descent of hierarchy, and selection in relational semantics. In *Proceedings of Annual Meeting of the Association for Computational Linguistics, ACL '02*, pages 247–254.
- Diarmuid Ó Séaghdha. 2009. Semantic classification with WordNet kernels. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers, NAACL '09*, pages 237–240.
- Fabian Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. YAGO: A core of semantic knowledge - unifying WordNet and Wikipedia. In *Proceedings of 16th International World Wide Web Conference, WWW '07*, pages 697–706.
- Marta Tatu and Dan Moldovan. 2005. A semantic approach to recognizing textual entailment. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, HLT-EMNLP '05*, pages 371–378.
- Michael Thelen and Ellen Riloff. 2002. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, EMNLP '02*, pages 214–221.
- Stephen Tratz and Eduard Hovy. 2010. A taxonomy, dataset, and classifier for automatic noun compound interpretation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL '10*, pages 678–687.
- Lucy Vanderwende. 1994. Algorithm for automatic interpretation of noun sequences. In *Proceedings of the 15th Conference on Computational linguistics*, pages 782–788.
- Beatrice Warren. 1978. Semantic patterns of noun-noun compounds. In *Gothenburg Studies in English 41, Goteburg, Acta Universtatis Gothoburgensis*.