Using Verbs to Characterize Noun-Noun Relations

Preslav Nakov and Marti Hearst

EECS and School of Information
University of California at Berkeley
Berkeley CA 94720, USA
{nakov@cs,hearst@sims}.berkeley.edu

Abstract. We present a novel, simple, unsupervised method for characterizing the semantic relations that hold between nouns in noun-noun compounds. The main idea is to discover *predicates* that make explicit the hidden relations between the nouns. This is accomplished by writing Web search engine queries that restate the noun compound as a relative clause containing a wildcard character to be filled in with a verb. A comparison to results from the literature suggest this is a promising approach.

Keywords: *Web statistics*, *noun compound*, *lexical semantics*, *componential analysis.*

1 Introduction

An important characteristic of technical literature is the abundance of long noun compounds like *bone marrow biopsy specimen* and *neck vein thrombosis*. While eventually mastered by domain experts, their interpretation poses a significant challenge for automated analysis, e.g., what is the relationship between *bone marrow* and *biopsy*? Between *biopsy* and *specimen*? Understanding relations between multiword expressions is important for many tasks, including question answering, textual entailment, machine translation, and information retrieval, among others.

In this paper we focus on the problem of determining the semantic relation(s) that holds within two-word English noun compounds. We introduce a novel approach for this problem: use paraphrases posed against an enormous text collection as a way to determine which predicates, represented as verbs, best characterize the relationship between the nouns.

Most algorithms that perform semantic interpretation place heavy reliance on the appearance of verbs, since they are the predicates which act as the backbone of the assertion being made. Noun compounds are terse elisions of the predicate; their structure assumes that the reader knows enough about the constituent nouns and about the world at large to be able to infer what the relationship between the words is. Our idea is to try to uncover the relationship between the noun pairs by, in essence, rewriting or paraphrasing the noun compound in such a way as to be able to determine the predicate(s) holding between the nouns. What is especially novel about this approach is paraphrasing noun compound semantics in terms of concrete verbs, rather than a fixed number of abstract predicates (e.g., `HAVE`, `MAKE`, `USE`), relations (e.g., `LOCATION`, `INSTRUMENT`, `AGENT`), or prepositions (e.g., `OF`, `FOR`, `IN`), as is traditional in the literature.

This idea builds on earlier work which shows that the vast size of the text available on the Web makes it likely that the same concept is stated in many different ways [1, 2]. This information in turn can be used to solve syntactic ambiguity problems [3, 4]. Here we extend that idea by applying it to determining semantic relations.

In our approach, we pose paraphrases for a given noun compound by rewriting it as a phrase that contains a wildcard where the verb would go. For example, we rewrite *neck vein* as `"vein that * neck"`, send this as a query to a Web search engine, and then parse the resulting snippets to find the verbs that appear in the place of the wildcard. Some of the most frequent verbs (+ prepositions) found for *neck vein* are: *emerge from*, *pass through*, *be found in*, *be terminated at*, *be in*, *flow in*, *run from*, *terminate in*, *descend in*, *come from*, etc. A comparison to examples from the literature suggest this is a promising approach with a broad range of potential applications.

In the remainder of this paper we first describe related work, then give details of the algorithm, present preliminary results as compared to other work in the literature, and discuss potential applications.

2 Related Work

There is currently no consensus as to which set of relations should hold between nouns in a noun compound, but most existing approaches make use of a set of a small number of abstract relations, typically less than 50. However, some researchers (e.g., Downing [5]), have proposed that an unlimited number is needed; in this paper we will hold a similar position.

One of the most important theoretical linguistic models is that of Levi [6], which states that noun compounds are derived through two basic processes: predicate nominalization (e.g., *'The president refused the appointment.'* → *presidential refusal*) and predicate deletion (*'pie made of apples'* → *apple pie*). According to Levi, predicate nominalizations can be subjective or objective, depending on whether the subject or the object is retained, and the relation can be further classified as `ACT`, `PRODUCT`, `AGENT` or `PATIENT`, depending on the thematic role between the nominalized verb and the retained argument. Predicate deletion, in turn, is limited to the following abstract predicates: five verbs (`CAUSE`, `HAVE`, `MAKE`, `USE` and `BE`) and four prepositions (`IN`, `FOR`, `FROM` and `ABOUT`). For example, according to Levi, *night flight* should be analyzed as `IN` (*flight at night*), and *tear gas* as `CAUSE` (*gas that causes tears*). A problem with this approach is that the predicates are too abstract, and can be ambiguous, e.g., *sand dune* is both `HAVE` and `BE`.

Lauer [7] simplifies Levi's idea by redefining the semantic relation identification problem as one of predicting which among the 8 prepositions is most likely to be associated with the compound when rewritten: *of*, *for*, *in*, *at*, *on*, *from*, *with* and *about*. Lapata and Keller [1] improve on Lauer's results (whose accuracy was 40%) by using his problem definition along with Web statistics to estimate ($noun_1$, *prep*, $noun_2$) trigram frequencies, achieving 55.71% accuracy. However, the preposition-oriented approach is problematic because the same preposition can indicate several different relations, and conversely, the same relation can be indicated by several different prepositions. For example, *in*, *on*, and *at* can all refer to both `LOCATION` and `TIME`.

Rosario and Hearst [8] show that a discriminative classifier can work quite well at assigning relations from a pre-defined set if training data is supplied in a domain-specific setting (60% accuracy, 18 classes). In later work, [9] provide a semi-supervised approach for characterizing the relation between two nouns in a bioscience noun-noun compound based on the semantic category each of the constituent nouns belongs to. Although this "descent of hierarchy" approach achieved a precision of 90% for finding the correct level of generalization, it does not assign *names* to the relations.

Girju et al. [10] apply both classic (SVM and decision trees) and novel supervised models (semantic scattering and iterative semantic specialization), using WordNet, word sense disambiguation, and a set of linguistic features. They test their system against both Lauer's 8 prepositional paraphrases and another set of 21 semantic relations, achieving up to 54% accuracy on the latter.

Lapata [11] focuses on the disambiguation of nominalizations. Using partial parsing, sophisticated smoothing and contextual information, she achieved 86.1% accuracy (baseline 61.5%) on the binary decision task of whether the modifier used to be the subject or the object of the nominalized verb (the head).

Girju et al. [12] present an SVM-based approach for the automatic classification of semantic relations in nominalized noun phrases (where either the head or the modifier has been derived from a verb). Their classification schema consists of 35 abstract semantic relations and has been also used by [13] for the semantic classification of noun phrases in general.

Turney and Littman [14] characterize the relation between two words, X and Y, as a vector whose coordinates correspond to Web frequencies for 128 phrases like "X for Y", "Y for X", etc., derived from a fixed set of 64 joining terms (e.g. "for", "such as", "not the", "is *", etc.). These vectors are then used in a nearest-neighbor classifier, which maps them to a set of fixed relations. He achieved an F-value of 26.5% (random guessing 3.3%) with 30 relations, and 43.2% (random: 20%) with 5 relations.

In work to appear, Turney [15] presents an unsupervised algorithm for mining the Web for patterns expressing implicit semantic relations. For example, `CAUSE` (e.g. *cold virus*) is best characterized by "*Y * causes X*", and "*Y in * early X*" is the best pattern for `TEMPORAL` (e.g. *morning frost*). He obtains an F-value 50.2% for 5 classes. This approach is the closest to our proposal.

Most other approaches to noun compound interpretation used hand-coded rules for at least one component of the algorithm [16], or rules combined with lexical resources [17] (52% accuracy, 13 relations). [18] make use of the identity of the two nouns and a number of syntactic clues in a nearest-neighbor classifier with 60-70% accuracy.

3 Using Verbs to Characterize Noun-Noun Relations

As we have described above, traditionally, the semantics of noun compounds have been represented as a set of abstract relations. This is problematic for several reasons. First, it is unclear which is the best set, and mapping between different sets has proven challenging [10]. Second, being both abstract and limited, these sets only capture a small part of the semantics, often multiple meanings are possible, and sometimes none of the pre-defined meanings are suitable for a given example. Finally, it is unclear how useful the proposed sets are, as researchers have often fallen short of demonstrating practical uses.

We believe verbs have more expressive power and are better tailored for the task of semantic representation: there is an infinite number of them (according to [5]) and they can capture fine-grained aspects of the meaning. For example, while *wrinkle treatment* and *migraine treatment* express the same abstract relation `TREATMENT-FOR-DISEASE`, some fine-grained differences can be shown by specific verbs e.g., *smooth* is possible in a verbal paraphrase of the former, but not of the latter.

In many theories, verbs play an important role in the process of noun compound derivation, and they are frequently used to make the hidden relation overt. This allows not only for simple and effective extraction (as we have seen above), but also for straightforward uses of the extracted verbs and paraphrases in NLP tasks like machine translation, information retrieval, etc.

We further believe that a single verb often is not enough and that the meaning is approximated better by a collection of verbs. For example, while *malaria mosquito* can very well be characterized as `CAUSE` (or *cause*), further aspects of the meaning, can be captured by adding some additional verbs e.g., *carry*, *spread*, *transmit*, *be responsible for*, *be infected with*, *pass on*, etc.

In the next section, we describe our algorithm for discovering predicate relations that hold between nouns in a compound.

4 Method

In a typical noun-noun compound "$noun_1$ $noun_2$", $noun_2$ is the head and $noun_1$ is a modifier, attributing a property to it. Our idea is to preserve the head-modifier relation by substituting the pre-modifier $noun_1$ with a suitable post-modifying relative phrase; e.g., "*tear gas*" can be transformed into "*gas that causes tears*", "*gas that brings tears*", "*gas which produces tears*", etc. Using all possible inflections of $noun_1$ and $noun_2$ as found in WordNet [19], we issue exact phrase Google queries of the following type:

```
"noun2 THAT * noun1"
```

where THAT can be *that*, *which* or *who*. The Google * operator is a one-word wildcard substitution; we issue queries with up to 8 stars.

We collect the text snippets (summaries) from the search results pages (up to 1000 per query) and we only keep the ones for which the sequence of words following noun1 is non-empty and contains at least one non-noun, thus ensuring the snippet includes the entire noun phrase. To help POS tagging and shallow parsing of the snippet, we further substitute the part before noun2 by the fixed phrase "*We look at the*". We then perform POS tagging [20] and shallow parsing[1], and extract the verb (and the following preposition, if any) between THAT and noun1. We allow for adjectives and participles to fall between the verb and the preposition, but not nouns; we ignore the modals, and the auxiliaries, but retain the passive *be*, and we make sure there is exactly one verb phrase (thus disallowing complex paraphrases like "*gas that makes the eyes fill with tears*"). Finally, we convert the main verb to an infinitive using WordNet [19].

The proposed method is similar to previous paraphrase acquisition approaches which search for similar/fixed endpoints and collect the intervening material. Lin and Pantel [21] extract paraphrases from dependency tree paths whose ends contain similar sets of words by generalizing over these ends. For example, for "*X solves Y*" they extract paraphrasing templates like "*Y is resolved by X*", "*X resolves Y*", "*X finds a solution to Y*" and "*X tries to solve Y*". The idea is extended by Shinyama et al. [22], who use named entities of matching semantic class as anchors, e.g., LOCATION, ORGANIZATION, etc. However, the goal of these approaches is to create summarizing paraphrases, while we are interested in finding noun compound semantics.

Table 1 shows a subset of the verbs found using our extraction method for *cancer treatment*, *migraine treatment*, *wrinkle treatment* and *herb treatment*. We can see that *herb treatment* is very different from the other compounds and shares no features with them: it *uses* and *contains* herb, but does not *treat* it. Further, while migraine and wrinkles cannot be *cured*, they can be *reduced*. Migraines can also be *prevented*, and wrinkles can be *smoothed*. Of course, these results are merely suggestive and should not be taken as ground truth, especially the absence of indicators. Still they seem to capture interesting fine-grained semantic distinctions, which normally require deep knowledge of the semantics of the two nouns and/or about the world.

5 Evaluation

5.1 Comparison with Girju et al., 2005

In order to test this approach, we compared it against examples from the literature. In this preliminary evaluation, we manually determined if verbs accurately reflected each paper's set of semantic relations.

Table 3 shows the results comparing against the examples of 21 relations that appear in [10]. In two cases, the most frequent verb is the copula, but

[1] OpenNLP tools: http://opennlp.sourceforge.net

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

Table 1. Some verbs found for different kinds of treatments.

	man	**woman**	**boy**	**bull**
ANIMATE	+	+	+	+
HUMAN	+	+	+	−
MALE	+	−	+	+
ADULT	+	+	−	+

Table 2. Example componential analysis for *man*, *woman*, *boy* and *bull.*

the following most frequent verbs are appropriate semantic characterizations of the compound. In the case of *"malaria mosquito"*, one can argue that the CAUSE relation, assigned by [10] is not really correct, in that the disease is only indirectly caused by the mosquitos, but rather is carried by them, and the proposed most frequent verbs *carry* and *spread* more accurately represent an AGENT relation. Nevertheless, *cause* is the third most frequent verb, indicating that it is common to consider the indirect relation as causal. In the case of *combustion gas*, the most frequent verb *support*, while being a good paraphrase of the noun compound, is not directly applicable to the relation assigned by [10] as RESULT, but the other verbs are.

In all other cases shown, the most frequent verbs accurately capture the relation assigned by [10]. In some cases, less frequent verbs indicate other logical entailments from the noun combination.

For the following examples, no meaningful verbs were found (in most cases there appears not to be a meaningful predicate for the particular nouns paired, or a nominalization plays the role of the predicate): *quality sound, crew investigation, image team, girl mouth, style performance, worker fatalities,* and *session day.*

5.2 Comparison with Barker and Szpakowicz, 1998

Table 4 shows comparison to examples from [18]. Due to space limitations, here we discuss the first 8 relations only. We also omitted *charitable donation* and

Sem. relation	Example	Extracted Verbs
POSSESSION	*family estate*	~~be in(29)~~, **be held by(9)**, **be owned by(7)**
TEMPORAL	*night flight*	**arrive at(19)**, **leave at(16)**, **be at(6)**, **be conducted at(6)**, **occur at(5)**
IS-A(HYPERNYMY)	*Dallas city*	**include(9)**
CAUSE	*malaria mosquito*	*carry(23)*, *spread(16)*, **cause(12)**, *transmit(9)*, *bring(7)*, *have(4)*, *be infected with(3)*, **be responsible for(3)**, *test positive for(3)*, **infect many with(3)**, *be needed for(3)*, **pass on(2)**, **give(2)**, **give out(2)**
MAKE/PRODUCE	*shoe factory*	**produce(28)**, **make(13)**, **manufacture(11)**
INSTRUMENT	*pump drainage*	**be controlled through(3)**, **use(2)**
LOCATION/SPACE	*Texas university*	~~be(5)~~, **be in(4)**
PURPOSE	*migraine drug*	**treat(11)**, **be used for(9)**, **prevent(7)**, **work for(6)**, **stop(4)**, **help(4)**, ~~work(4)~~ **be prescribed for(3)**, **relieve(3)**, **block(3)**, *be effective for(3)*, *be for(3)*, **help ward off(3)**, *seem effective against(3)*, **end(3)**, **reduce(2)**, **cure(2)**
SOURCE	*olive oil*	**come from(13)**, **be obtained from(11)**, **be extracted from(10)**, **be made from(9)**, **be produced from(7)**, **be released from(4)**, *taste like(4)*, **be beaten from(3)**, **be produced with(3)**, **emerge from(3)**
TOPIC	*art museum*	**focus on(29)**, *display(16)*, **bring(14)**, **highlight(11)**, *house(10)*, *exhibit(9)* **demonstrate(8)**, **feature(7)**, *show(5)*, **tell about(4)**, **cover(4)**, **concentrate in(4)**
MEANS	*bus service*	**use(14)**, **operate(6)**, *include(6)*
EXPERIENCER	*disease victim*	**spread(12)**, **acquire(12)**, **suffer from(8)**, **die of(7)**, *develop(7)*, **contract(6)**, **catch(6)**, **be diagnosed with(6)**, *have(5)*, *beat(5)*, **be infected by(4)**, **survive(4)**, **die from(4)**, **get(4)**, **pass(3)**, **fall by(3)**, *transmit(3)*, *avoid(3)*
THEME	*car salesman*	**sell(38)**, ~~mean inside(13)~~, **buy(7)**, **travel by(5)**, **pay for(4)**, **deliver(3)**, **push(3)**, **demonstrate(3)**, **purr(3)**, ~~bring used(3)~~, *know more about(3)*, *pour through(3)*
RESULT	*combustion gas*	*support(22)*, **result from(14)**, **be produced during(11)**, **be produced by(8)**, **be formed from(8)**, **form during(8)**, **be created during(7)**, **originate from(6)**, **be generated by(6)**, **develop with(6)**, **come from(5)**, ~~be cooled(5)~~

Table 3. Comparison to examples (14 out of 21) found in [10], showing the most frequently extracted verbs. Verbs expressing the target relation are in bold, those referring to a different, but semantically valid, are in italic, and errors are struck out.

Relation	Example	Extracted Verbs
AGENT	*student protest*	**be led by(6)**, **be sponsored by(6)**, **pit(4)**, *be(4)*, **be organized by(3)**, **be staged by(3)**, **be launched by(3)**, **be started by(3)**, **be supported by(3)**, *involve(3)*, *arise from(3)*
AGENT	*band concert*	*feature(17)*, *capture(10)*, *include(6)*, **be given by(6)**, *play of(4)*, *involve(4)*, ~~be than(4)~~ **be organized by(3)**, **be by(3)**, *start with(3)*, *bring(3)*, *take(3)*, *consist of(3)*
AGENT	*military assault*	**be initiated by(4)**, *shatter(2)*
BENEFICIARY	*student price*	*be(14)*, ~~mean(4)~~, ~~differ from(4)~~, **be unfair for(3)**, **be discounted for(3)**, **be for(3)**, **be affordable for(3)**, **be charged for(3)**, **please(3)**, **be shared with(3)**, *draw in(3)*
CAUSE	*exam anxiety*	*be generated during(3)*
CONTAINER	*printer tray*	**hold(12)**, *come with(9)*, *be folded(8)*, *fit under(6)*, **be folded into(4)**, **pull from(4)**, **be inserted into(4)**, *be mounted on(4)*, *be used by(4)*, **be inside(3)**, **feed into(3)**
CONTAINER	*flood water*	*cause(24)*, *produce(9)*, *remain after(9)*, *be swept by(6)*, *create(5)*, *bring(5)*, *reinforce(5)*
CONTAINER	*film music*	*fit(16)*, **be in(13)**, **be used in(11)**, **be heard in(11)**, *play throughout(9)*, *be written for(9)*
CONTAINER	*story idea*	*tell(20)*, *make(19)*, *drive(15)*, *become(13)*, *turn into(12)*, *underlie(12)*, **occur within(8)**, **hold(8)**, *tie(8)*, *be(8)*, *spark(8)*, **appear throughout(7)**, *tell(7)*, *move(7)*, *come from(6)*
CONTENT	*paper tray*	**feed(6)**, *be lined with(6)*, *stand up(6)*, **hold(4)**, **contain(4)**, *catch(4)*, **overflow with(3)**
CONTENT	*eviction notice*	*result in(10)*, *precede(3)*, *make(2)*
DESTINATION	*game bus*	*be in(6)*, **leave for(3)**, *be like(3)*, *be(3)*, *make playing(3)*, *lose(3)*
DESTINATION	*exit route*	*be indicated by(4)*, **reach(2)**, *have(1)*, *do(1)*
DESTINATION	*entrance stairs*	*look like(4)*, *stand outside(3)*, *have(3)*, *follow from(3)*, *be at(3)*, ~~be(3)~~, *descend from(2)*
EQUATIVE	*player coach*	*work with(42)*, *recruit(28)*, **be(19)**, *have(16)*, *know(16)*, *help(12)*, *coach(11)*, *take(11)*
INSTRUMENT	*electron microscope*	**use(27)**, *show(5)*, **work with(4)**, **utilize(4)**, **employ(4)**, *beam(3)*
INSTRUMENT	*diesel engine*	*be(18)*, **operate on(8)**, *look like(8)*, **use(7)**, *sound like(6)*, **run on(5)**, **be on(5)**
INSTRUMENT	*laser printer*	**use(20)**, *consist of(6)*, *be(5)*

Table 4. Comparison to examples (8 out of 20) from [18], showing the most frequently extracted verbs. Verbs expressing the target relation are in bold, those referring to a different, but semantically valid, are in italic, and errors are struck out.

overdue fine, as the modifier in these cases is an adjective, and *composer arranger*, because no results were found.

We obtain very good results for AGENT and INSTRUMENT, but other relations are problematic, probably because the assigned classifications are of varying quality: *printer tray* and *film music* are probably correctly assigned to CONTAINER, but *flood water* and *story idea* are not; *entrance stairs* (DESTINATION) could be equally well analyzed as LOCATED or SOURCE; and *exam anxiety* (CAUSE) probably refers to TIME. Finally, although we find the verb *be* ranked third for *player coach*, the EQUATIVEs pose a problem in general, as the copula is not very frequent in this form of paraphrase.

5.3 Comparison with Rosario and Hearst, 2002

As we mentioned above, [9] characterize noun-noun compounds based on the semantic category, in the MeSH lexical hierarchy, each of the constituent nouns belongs to. For example, all noun compounds in which the first noun is classified under the A01 sub-hierarchy[2] (*Body Regions*), and the second one falls into A07 (*Cardiovascular System*), are hypothesized to express the same relation. Examples include *mesentery artery*, *leg vein*, *finger capillary*, etc.

By contrast, for the category pair A01-M01 (*Body Regions–Persons*) a distinction is needed between different kinds of persons and the algorithm needs to descend one level on the M01 side: M01.643 (*Patients*), M01.898 (*Donors*), M01.150 (*Disabled Persons*).

Table 5 shows some results of our comparison to [9]. Given a category pair (e.g., A01-A07), we consider all of the noun-noun compounds whose elements are in the corresponding MeSH sub-hierarchies, and we acquire a set of paraphrasing verbs+prepositions from the Web for each of them. We then aggregate the results from all such word pairs in order to obtain a set of paraphrasing verbs for the target category pair.

6 Potential Applications

The extracted verbs (+prepositions) have the potential to be useful for a number of important NLP tasks. For example, they may help in the process of noun compound translation [23]. They could be also directly integrated into a paraphrase-augmented machine translation system [24], machine translation evaluation system [25] [26], or summarization evaluation system [27].

Assuming annotated training data, the verbs could be used as features in the prediction of abstract relations like TIME and LOCATION, as is done by [14] and [15], who used the vector-space model and a nearest-neighbor classifier.

[2] In MeSH each concept is assigned one or more codes, corresponding to positions in the hierarchy e.g., A (*Anatomy*), A01 (*Body Regions*), A01.456 (*Head*), A01.456.505 (*Face*), A01.456.505.420 (*Eye*). *Eye* is ambiguous; it is also A09.371 (A09 represents *Sense Organs*).

Categ. Pair	Examples	Extracted Verbs
A01-A07 (Body Regions - Cardiovascular System)	*ankle artery* *foot vein* *forearm vein* *finger artery* *neck vein* *head vein* *leg artery* *thigh vein*	*feed*(133), *supply*(111), *drain*(100), *be in*(44), *run*(37), *appear on*(29), *be located in*(22), *be found in*(20), *run through*(19), *be behind*(19), *run from*(18), *serve*(15), *be felt with*(14), *enter*(14), *pass through*(12), *pass by*(12), *show on*(11), *be visible on*(11), *run along*(11), *nourish*(10), *be seen on*(10), *occur on*(10), *occur in*(9), *emerge from*(9), *go into*(9), ...
A01-M01.643 (Body Regions - Disabled Persons)	*arm patient* *eye outpatient* *abdomen patient*	*be*(54), *lose*(40), *have*(30), *be hit in*(11), *break*(9), *gouge out*(9), *injure*(8), *receive*(7), *be stabbed in*(7), *be shot in*(7), *need*(6), ...
A01-M01.150 (Body Regions - Disabled Persons)	*leg amputee* *arm amputee* *knee amputee*	*lose*(13), *grow*(6), *have cut off*(4), *miss*(2), *need*(1), *receive*(1), *be born without*(1)
A01-M01.898 (Body Regions - Donors)	*eye donor* *skin donor*	*give*(4), *provide*(3), *catch*(1)
D02-E05.272 (Organic Chemicals - Diet)	*choline diet* *methionine diet* *carotene diet* *saccharin diet*	*be low in*(18), *contain*(13), *be deficient in*(11), *be high in*(7), *be rich in*(6), *be sufficient in*(6), *include*(4), *be supplemented with*(3), *be in*(3), *be enriched with*(3), *contribute*(2), *miss*(2), ...

Table 5. Comparison to [9] showing the most frequent verbs.

These relations in turn could play an important role in other applications, as demonstrated by [28], who achieved state-of-the-art results on the PASCAL Recognizing Textual Entailment challenge.

In information retrieval, the verbs could be used for index normalization [29] or query refinement, e.g., when querying for *migraine treatment*, pages containing good paraphrasing verbs, like *relieve* or *prevent*, would be preferred.

The verbs and prepositions, intervening between the two nouns could be also used to seed a Web search for whole classes of NPs [30], such as diseases, drugs, etc. For example, after finding that *prevent* is a good paraphrase for *migraine treatment*, we can use the query `"* which prevents migraines"` to obtain different treatments/drugs for migraine, e.g. *feverfew*, *Topamax*, *natural treatment*, *magnesium*, *Botox*, *Glucosamine*, etc.

Finally, the extracted verbs could be used for linguistic analysis. Note the similarity between Table 1 and Table 2. The latter shows a sample *componential analysis*, which represents word's semantics in terms of primitives, called components or features, thus making explicit relations like hyponymy, incompatibility, etc. [31–33]. Table 1 shows a similar semantic representation for noun-noun compounds. While the classic componential analysis has been criticized for being inherently subjective, a new *dynamic componential analysis* would extract the components automatically from a large corpus in a principled manner.

7 Conclusions and Future Work

We have presented a simple unsupervised approach to noun compound interpretation in terms of predicates characterizing the hidden relation, which could be useful for many NLP tasks.

A significant benefit of our approach is that it does not require knowledge of the meanings of constituent nouns in order to correctly assign relations. A potential drawback is that it will probably not work well for low-frequency words, so semantic class information will be needed for these cases.

In future we plan to apply full parsing to reduce the errors caused by shallow parsing and POS errors. We will also assess the results against a larger collection of manually labeled relations, and have an independent evaluation of the appropriateness of the verbs for those relations. We also plan to combine this work with the structural ambiguity resolution techniques of [4], and determine semantic relations among multi-word terms. Finally, we want to test the approach on some of the above-mentioned NLP tasks.

References

1. Lapata, M., Keller, F.: Web-based models for natural language processing. ACM Transactions on Speech and Language Processing **2** (2005) 1–31
2. Banko, M., Brill, E.: Scaling to very very large corpora for natural language disambiguation. Proceedings of the 39th Annual Meeting on Association for Computational Linguistics (2001) 26–33
3. Nakov, P., Hearst, M.: Search engine statistics beyond the n-gram: Application to noun compound bracketing. In: Proceedings of the 9th Conference on Computational Natural Language Learning. (2005) 17–24
4. Nakov, P., Hearst, M.: Using the Web as an implicit training set: Application to structural ambiguity resolution. In: Proceedings of HLT-EMNLP, Vancouver, British Columbia, Canada (2005) 835–842
5. Downing, P.: On the creation and use of English compound nouns. Language **53**(4) (1977) 810–842
6. Levi, J.: The Syntax and Semantics of Complex Nominals. Academic Press, New York (1978)
7. Lauer, M.: Designing Statistical Language Learners: Experiments on Noun Compounds. PhD thesis, Department of Computing Macquarie University NSW 2109 Australia (1995)
8. Rosario, B., Hearst, M.: Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy. In Lee, L., Harman, D., eds.: Proceedings of EMNLP (Empirical Methods in Natural Language Processing). (2001) 82–90
9. Rosario, B., Hearst, M., Fillmore, C.: The descent of hierarchy, and selection in relational semantics. In: Proceedings of ACL. (2002) 247–254
10. Girju, R., Moldovan, D., Tatu, M., Antohe, D.: On the semantics of noun compounds. Computer Speech and Language **19**(4) (2005) 479–496
11. Lapata, M.: The disambiguation of nominalisations. Computational Linguistics **28**(3) (2002) 357–388

12. Girju, R., Giuglea, A.M., Olteanu, M., Fortu, O., Bolohan, O., Moldovan, D.: Support vector machines applied to the classification of semantic relations in nominalized noun phrases. In: Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics, Boston, MA (2004) 68–75
13. Moldovan, D., Badulescu, A., Tatu, M., Antohe, D., Girju, R.: Models for the semantic classification of noun phrases. In: Proceedings of the Computational Lexical Semantics Workshop at HLT-NAACL. (2004) 60–67
14. Turney, P., Littman, M.: Corpus-based learning of analogies and semantic relations. Machine Learning Journal **60**(1-3) (2005) 251–278
15. Turney, P.: Expressing implicit semantic relations without supervision. In: Proceedings of COLING-ACL, Australia (2006)
16. Finin, T.: The Semantic Interpretation of Compound Nominals. Ph.d. dissertation, University of Illinois, Urbana, Illinois (1980)
17. Vanderwende, L.: Algorithm for automatic interpretation of noun sequences. In: Proceedings of COLING-94. (1994) 782–788
18. Barker, K., Szpakowicz, S.: Semi-automatic recognition of noun modifier relationships. In: Proceedings of COLING-ACL'98, Montreal, Canada (1998) 96–102
19. Fellbaum, C., ed.: WordNet: An Electronic Lexical Database. MIT Press (1998)
20. Toutanova, K., Klein, D., Manning, C., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of HLT-NAACL 2003. (2003) 252–259
21. Lin, D., Pantel, P.: Discovery of inference rules for question-answering. Natural Language Engineering **7**(4) (2001) 343–360
22. Shinyama, Y., Sekine, S., Sudo, K., Grishman, R.: Automatic paraphrase acquisition from news articles. In: Proceedings of Human Language Technology Conference (HLT 2002), San Diego, USA (2002) 40–46
23. Baldwin, T., Tanaka, T.: Translation by machine of complex nominals: Getting it right. In: Proceedings of the ACL04 Workshop on Multiword Expressions: Integrating Processing. (2004) 24–31
24. Callison-Burch, C., Koehn, P., Osborne, M.: Improved statistical machine translation using paraphrases. In: Proceedings of HLT-NAACL 2006. (2006) 17–24
25. Russo-Lassner, G., Lin, J., Resnik, P.: A paraphrase-based approach to machine translation evaluation. Technical Report LAMP-TR-125/CS-TR-4754/UMIACS-TR-2005-57, University of Maryland (2005)
26. Kauchak, D., Barzilay, R.: Paraphrasing for automatic evaluation. In: Proceedings of HLT-NAACL 2006. (2006) 455–462
27. Liang Zhou, Chin-Yew Lin, D.S.M., Hovy, E.: PARAEVAL: Using paraphrases to evaluate summaries automatically. In: Proceedings of HLT-NAACL 2006. (2006) 447–454
28. Tatu, M., Moldovan, D.: A semantic approach to recognizing textual entailment. In: Proceedings of HLT/EMNLP 2005. (2005) 371–378
29. Zhai, C.: Fast statistical parsing of noun phrases for document indexing. In: Proceedings of the fifth conference on Applied natural language processing, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc. (1997) 312–319
30. Etzioni, O., Cafarella, M., Downey, D., Popescu, A.M., Shaked, T., Soderland, S., Weld, D.S., Yates, A.: Unsupervised named-entity extraction from the web: an experimental study. Artificial Intelligence **165**(1) (2005) 91–134
31. Katz, J., Fodor, J.: The structure of a semantic theory. Language (39) (1963) 170–210
32. Jackendoff, R.: Semantics and Cognition. MIT Press, Cambridge, MA (1983)
33. Saeed, J.: Semantics. 2 edn. Blackwell (2003)