Paraphrasing Verbs for Noun Compound Interpretation

Preslav Nakov

Linguistic Modeling Department, Institute for Parallel Processing
Bulgarian Academy of Sciences
25A, Acad. G. Bonchev St., 1113 Sofia, Bulgaria
nakov@lml.bas.bg

Abstract

An important challenge for the automatic analysis of English written text is the abundance of noun compounds: sequences of nouns acting as a single noun. In our view, their semantics is best characterized by the set of all possible paraphrasing verbs, with associated weights, e.g., *malaria mosquito* is *carry (23)*, *spread (16)*, *cause (12)*, *transmit (9)*, etc. Using Amazon's Mechanical Turk, we collect paraphrasing verbs for 250 noun-noun compounds previously proposed in the linguistic literature, thus creating a valuable resource for noun compound interpretation. Using these verbs, we further construct a dataset of pairs of sentences representing a special kind of textual entailment task, where a binary decision is to be made about whether an expression involving a verb and two nouns can be transformed into a noun compound, while preserving the sentence meaning.

1. Introduction

An important challenge for the automatic analysis of English written text is posed by noun compounds – sequences of nouns acting as a single noun[1], e.g., *colon cancer tumor suppressor protein* – which are abundant in English: Baldwin and Tanaka (2004) calculated that noun compounds comprise 3.9% and 2.6% of all tokens in the *Reuters corpus* and the *British National Corpus*[2], respectively.

Understanding noun compounds' syntax and semantics is difficult but important for many natural language applications (NLP) including question answering, machine translation, information retrieval, and information extraction. For example, a question answering system might need to determine whether '*protein acting as a tumor suppressor*' is a good paraphrase for *tumor suppressor protein*, and an information extraction system might need to decide whether *neck vein thrombosis* and *neck thrombosis* could possibly co-refer when used in the same document. Similarly, a machine translation system facing the unknown noun compound *WTO Geneva headquarters* might benefit from being able to paraphrase it as *Geneva headquarters of the WTO* or as *WTO headquarters located in Geneva*. Given a query like *migraine treatment*, an information retrieval system could use suitable paraphrasing verbs like *relieve* and *prevent* for page ranking and query refinement.

2. Noun Compound Interpretation

The dominant view in theoretical linguistics is that noun compound semantics can be expressed by a small set of abstract relations. For example, in the theory of Levi (1978), complex nominals (a more general notion than noun compounds) can be derived by two processes – predicate deletion (e.g., *pie made of apples* → *apple pie*) and predicate nominalization (e.g., *the President refused general MacArthur's request* → *presidential refusal*). The former can only delete the 12 abstract recoverably deletable predicates (RDPs) shown in Table 1.

RDP	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

Table 1: **Levi's recoverably deletable predicates (RDPs).** Column 3 shows modifier's function in the relative clause.

Similarly, in the theory of Warren (1978), noun compounds can express six major types of semantic relations (which are further divided into finer sub-relations): Constitute, Possession, Location, Purpose, Activity–Actor, and Resemblance.

A similar view is dominant in computational linguistics. For example, Nastase and Szpakowicz (2003) use 30 fine-grained relations (e.g., Cause, Effect, Purpose, Frequency, Direction, Location), grouped into 5 coarse-grained super-relations: QUALITY, SPATIAL, TEMPORALITY, CAUSALITY, and PARTICIPANT. Similarly, Girju et al. (2005) propose a set of 21 abstract relations (e.g., CAUSE, INSTRUMENT, PURPOSE, RESULT), and Rosario and Hearst (2001) use 18 abstract domain-specific biomedical relations (e.g., Defect, Material, Person Afflicted).

An alternative view is held by Lauer (1995), who defines the problem of noun compound interpretation as predicting which among the following eight prepositions best paraphrases the target noun compound: *of*, *for*, *in*, *at*, *on*, *from*, *with*, and *about*. For example, *olive oil* is *oil from olives*.

Lauer's approach is attractive since it is simple and yields prepositions representing paraphrases directly usable in NLP applications. However, it is also problematic since

[1] This is Downing (1977)'s definition of noun compounds.

[2] There are 256K distinct noun compounds out of the 939K distinct wordforms in the 100M-word *British National Corpus*.

mapping between prepositions and abstract relations is hard (Girju et al., 2005), e.g., *in*, *on*, and *at*, all can refer to both LOCATION and TIME.

Using abstract relations like CAUSE is problematic as well. First, it is unclear which relation inventory is the best one. Second, being both abstract and limited, such relations capture only part of the semantics, e.g., classifying *malaria mosquito* as CAUSE obscures the fact that mosquitos do not directly cause malaria, but just transmit it. Third, in many cases, multiple relations are possible, e.g., in Levi's theory, *sand dune* can be interpreted as both HAVE and BE.

Some of these issues are addressed by Finin (1980), who proposes to use a specific verb, e.g., *salt water* is interpreted as *dissolved in*. In a number of publications (Nakov and Hearst, 2006; Nakov, 2007; Nakov and Hearst, 2008), we introduced and advocated an extension of this idea, where noun compounds are characterized by the set of all possible paraphrasing verbs, with associated weights, e.g., *malaria mosquito* is *carry (23)*, *spread (16)*, *cause (12)*, *transmit (9)*, *etc*. These verbs are fine-grained, directly usable as paraphrases, and using multiple of them for a given noun compound approximates its semantics better.

Following this line of research, below we present two noun compound interpretation datasets which use human-derived paraphrasing verbs and are consistent with the view of an infinite inventory of relations. By making these resources publicly available, we hope to inspire further research in paraphrase-based noun compound interpretation.

3. Manual Annotations

We used a subset of the 387 examples listed in the appendix of (Levi, 1978). As we mentioned above, Levi's theory targets complex nominals, which include not only nominal compounds (e.g., *peanut butter*, *snowball*), but also nominalizations (e.g., *dream analysis*), and nonpredicate noun phrases (e.g., *electric shock*). We kept the former two categories since they are composed of nouns only and thus are noun compounds under our definition, but we removed the nonpredicate noun phrases, which have an adjectival modifier. We further excluded all concatenations (e.g., *silkworm*), thus ending up with 250 noun-noun compounds.

We then defined a paraphrasing task which asks human subjects to produce verbs, possibly followed by prepositions, that could be used in a paraphrase involving *that*. For example, *come from*, *be obtained from*, and *be from* are good paraphrasing verbs for *olive oil* since they can be used in paraphrases like '*oil that comes from olives*', '*oil that is obtained from olives*' or '*oil that is from olives*'. Note that this task definition implicitly allows for prepositional paraphrases when the verb is to *be* and is followed by a preposition. For example, the last paraphrase above is equivalent to '*oil from olives*'.

In an attempt to make the task as clear as possible and to ensure high quality of the results, we provided detailed instructions, we stated explicit restrictions, and we gave several example paraphrases. We instructed the participants to propose at least three paraphrasing verbs per noun-noun compound, if possible. We used the *Amazon Mechanical Turk* Web service[3], which represents a cheap and easy way to recruit subjects for various tasks that require human intelligence; it provides an API allowing a computer program to ask a human to perform a task and return the results.

We randomly distributed the noun-noun compounds into groups of 5 and we requested 25 different human subjects per group. We had to reject some of the submissions, which were empty or were not following the instructions, in which cases we requested additional workers in order to obtain about 25 good submissions per HIT (Human Intelligence Task). Each human subject was allowed to work on any number of groups, but was not permitted to do the same one twice, which is controlled by the *Amazon Mechanical Turk* Web Service. A total of 174 different human subjects produced 19,018 verbs. After removing the empty and the bad submissions, and after normalizing the verbs, we ended up with 17,821 verb annotations for the 250 examples. See Nakov (2007) for further details on the process of extraction and cleansing.

[3] http://www.mturk.com

4. Lexicons of Paraphrasing Verbs

We make freely available three lexicons of paraphrasing verbs for noun compound interpretation: two generated by human subjects recruited with *Amazon Mechanical Turk*, and a third one automatically extracted from the Web, as described in (Nakov and Hearst, 2008).

4.1. Human-Proposed: All

The dataset is provided as a text file containing a separate line for each of the 250 noun-noun compounds, ordered lexicographically. Each line begins with an example number (e.g., 94), followed by a noun compound (e.g., *flu virus*), the original Levi's RDP (e.g., $\texttt{CAUSE}_1$; see Table 1), and a list of paraphrasing verbs. The verbs are separated by a semicolon and each one is followed in parentheses by a count indicating the total number of distinct human annotators that proposed it. Here is an example line:

```
94 flu virus CAUSE1 cause(19); spread(4); give(4);
result in(3); create(3); infect with(3); contain(3);
be(2); carry(2); induce(1); produce(1); look like(1);
make(1); incubate into(1); exacerbate(1); turn into(1);
happen from(1); transmit(1); be made of(1); involve(1);
generate(1); breed(1); be related to(1); sicken with(1);
lead to(1); intensify be(1); disseminate(1); come
from(1); be implicated in(1); appear(1); instigate(1);
be conceived by(1); bring about(1)
```

4.2. Human-Proposed: First Only

As we mentioned above, the human subjects recruited to work on *Amazon Mechanical Turk* (workers) were instructed to provide at least three paraphrasing verbs per noun-noun compound. Sometimes this was hard, and many introduced some bad verbs in order to fulfill this requirement. Assuming that the very first verb is the most likely one to be correct, we created a second dataset in the same format, where only the first verb from each worker is considered. For example, line 94 in that new text file becomes:

```
94 flu virus CAUSE1 cause(17), be(1), carry(1),
involve(1), come from(1)
```

4.3. Automatically Extracted from the Web

Finally, we provide a text file in the same format, where the verbs are automatically extracted from the Web using the method described in (Nakov and Hearst, 2008). This dataset might be found useful by other researchers for comparison purposes. The corresponding line 94 in that file starts as follows (here truncated due to a very long tail):

```
94 flu virus CAUSE1 cause(906); produce(21);
give(20); differentiate(17); be(16); have(13);
include(11); spread(7); mimic(7); trigger(6); induce(5);
develop from(4); be like(4); be concealed by(3); be
characterized by(3); bring(3); carry(3); become(3); be
associated with(3); ...
```

4.4. Comparing the Human-Proposed and the Program-Generated Paraphrasing Verbs

In this section, we describe a comparison of the human- and the program-generated verbs aggregated by Levi's RDP (see Table 1). Given an RDP like $\texttt{HAVE}_1$, we collected all verbs belonging to noun-noun compounds from that RDP together with their frequencies. From a vector-space model point of view, we summed their corresponding frequency vectors. We did this separately for the human- and the program-generated verbs, and we compared them for each RDP. Figure 4.4. shows the cosine correlations (in %s) between the human- and the program-generated verbs by Levi's RDP: using all human-proposed verbs vs. using the first verb from each worker only. As we can see, there is a very-high correlation (mid 70s to mid 90s) for RDPs like $\texttt{CAUSE}_1$, $\texttt{MAKE}_1$, and BE, but low correlation 11-30% for reverse RDPs like $\texttt{HAVE}_2$ and $\texttt{MAKE}_2$, and for rare RDPs like ABOUT. Interestingly, using the first verb only improves the results for RDPs with high cosine correlation, but damages low-correlated ones. This suggests that when the RDP is more homogeneous, the first verbs proposed by the workers are good enough and the following ones only introduce noise, but when it is more heterogeneous, the additional verbs are more likely to be useful.

We also performed an experiment using the verbs as features in a nearest-neighbor classifier, trying to predict the Levi's RDP the noun compound belongs to. We first filtered out all nominalizations, thus obtaining 214 noun compounds, each annotated with one of the 12 RDPs shown in Table 1, and we then used this dataset in a leave-one-out cross-validation. Using all human-proposed verbs, we achieved 73.71%±6.29% accuracy (here we also show the confidence interval). For comparison, using Web-derived verbs and prepositions only yields 50.47%±6.68% accuracy. Therefore, we can conclude that the performance with human-proposed verbs is an upper bound on what can be achieved with Web-derived ones. See (Nakov and Hearst, 2008) for additional details.

5. A Dataset for Textual Entailment

Collecting this dataset was motivated by the Pascal Recognizing Textual Entailment (RTE) Challenge,[4] which addresses a generic semantic inference task arguably needed by many NLP applications, including question answering,

[4] www.pascal-network.org/Challenges/RTE2/

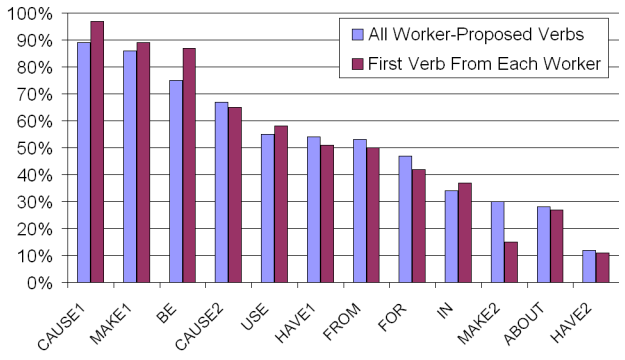

Figure 1: **Cosine correlation (in %s) between the human- and the program- generated verbs by Levi's RDP:** using all human-proposed verbs vs. using the first verb from each worker only.

information retrieval, information extraction, and multi-document summarization. Given two textual fragments, a text T and a hypothesis H, the goal is to recognize whether the meaning of H is entailed (can be inferred) from the meaning of T. Or, as the RTE2 task definition puts it:

> "*We say that T entails H if, typically, a human reading T would infer that H is most likely true. This somewhat informal definition is based on (and assumes) common human understanding of language as well as common background knowledge.*"

In many cases, solving such entailment problems requires deciding whether a noun compound can be paraphrased in a particular way.

The sentences in our Textual Entailment dataset are collected from the Web and involve some of the above-described human-derived paraphrasing verbs. These sentences are further manually annotated and provided in format that is similar to that used by RTE. Each example consists of three lines, all starting with the example number. The first line continues with T: (the text), followed by a sentence where the target nouns involved in a paraphrase are marked. The second line continues with H: (the hypothesis), followed by the same sentence but with the paraphrase re-written as a noun compound. The third line continues with A: (the answer), followed by either YES or NO, depending on whether T implies H.

The following example is positive since *professors that are women* is an acceptable paraphrase of the noun compound *women professors*:

```
17 T: I have friends that are organizing
to get more <e2>professors</e2> that are
<e1>women</e> and educate women to make
specific choices on where to get jobs.
17 H: I have friends that are organizing to
get more <e1>women</e1> <e2>professors</e2>
and educate women to make specific choices
on where to get jobs.
17 A: YES
```

The example below however is negative since a bad paraphrasing verb is used in the first sentence:

```
18 T: As McMillan collected, she also
quietly gave, donating millions of
dollars to create scholarships and
fellowships for black Harvard Medical
School students, African filmmakers,
and MIT <e2>professors</e2> who study
<e1>women</e1> in the developing world.
18 H: As McMillan collected, she also
quietly gave, donating millions of dollars
to create scholarships and fellowships
for black Harvard Medical School students,
African filmmakers, and MIT <e1>women</e1>
<e2>professors</e2> in the developing
world.
18 A: NO
```

Here is another kind of negative example, where the semantics is different due to a different phrase attachment. The first sentence refers to the action of giving, while the second one refers to the process of transfusion:

```
19 T: Rarely, the disease is transmitted
via transfusion of blood products from
a <e2>donor</e2> who gave <e1>blood</e1>
during the viral incubation period.
19 H: Rarely, the disease is transmitted
via transfusion of blood products from a
<e1>blood</e1> <e2>donor</e2> during the
viral incubation period.
19 A: NO
```

6. Conclusion

We have presented several novel resources consistent with the idea of characterizing noun compound semantics by the set of all possible paraphrasing verbs. These verbs are fine-grained, directly usable as paraphrases, and using multiple of them for a given noun compound approximates its semantics better. By making these resources publicly available, we hope to inspire further research in the direction of paraphrase-based noun compound interpretation, which opens the door to practical applications in a number of NLP tasks including but not limited to machine translation, text summarization, question answering, information retrieval, textual entailment, relational similarity, etc.

Unfortunately, the present situation with noun compound interpretation is similar to the situation with word sense disambiguation: in both cases, there is a general agreement that the research is important and much needed, there is a growing interest in performing further research, and a number of competitions are being organized, e.g., as part of SemEval (Girju et al., 2007). Still, despite that research interest, there is a lack of actual NLP applications using noun compound interpretation, with the notable exceptions of Tatu and Moldovan (2005) and Nakov (2008), who demonstrated improvements on textual entailment and machine translation, respectively. We believe that demonstrating more successful applications in real NLP problems is key for the advancement of the field, and we hope that other researchers will find the resources we release here helpful in this respect.

7. License

All datasets are released under the *Creative Commons License* (See `http://creativecommons.org/`).

Acknowledgments. This research was supported in part by NSF DBI-0317510 and by FP7-REGPOT-2007-1 SISTER.

8. References

Timothy Baldwin and Takaaki Tanaka. 2004. Translation by machine of compound nominals: Getting it right. In *Proceedings of the ACL 2004 Workshop on Multiword Expressions: Integrating Processing*, pages 24–31.

Pamela Downing. 1977. On the creation and use of English compound nouns. *Language*, (53):810–842.

Timothy Finin. 1980. *The Semantic Interpretation of Compound Nominals*. Ph.d. dissertation, University of Illinois, Urbana, Illinois.

Roxana Girju, Dan Moldovan, Marta Tatu, and Daniel Antohe. 2005. On the semantics of noun compounds. *Journal of Computer Speech and Language - Special Issue on Multiword Expressions*, 4(19):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 SemEval*, pages 13–18, Prague, Czech Republic.

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.

Preslav Nakov and Marti Hearst. 2006. Using verbs to characterize noun-noun relations. In *AIMSA*, volume 4183 of *Lecture Notes in Computer Science*, pages 233–244. Springer.

Preslav Nakov and Marti Hearst. 2008. Solving relational similarity problems using the web as a corpus. In *Proceedings of ACL'08: HLT*, Columbus, OH.

Preslav Nakov. 2007. *Using the Web as an Implicit Training Set: Application to Noun Compound Syntax and Semantics*. Ph.D. thesis, EECS Department, University of California, Berkeley, UCB/EECS-2007-173.

Preslav Nakov. 2008. Improving English-Spanish statistical machine translation: Experiments in domain adaptation, sentence paraphrasing, tokenization, and recasing. In *Proceedings of the Third Workshop on Statistical Machine Translation*, Columbus, OH.

Vivi Nastase and Stan Szpakowicz. 2003. Exploring noun-modifier semantic relations. In *Fifth International Workshop on Computational Semantics (IWCS-5)*, pages 285–301, Tilburg, The Netherlands.

Barbara Rosario and Marti Hearst. 2001. Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy. In *Proceedings of EMNLP*.

Marta Tatu and Dan Moldovan. 2005. A semantic approach to recognizing textual entailment. In *Proceedings of HLT*, pages 371–378.

Beatrice Warren. 1978. Semantic patterns of noun-noun compounds. In *Gothenburg Studies in English 41, Goteburg, Acta Universtatis Gothoburgensis*.