SemEval-2010 Task 9: The Interpretation of Noun Compounds Using Paraphrasing Verbs and Prepositions

Cristina Butnariu
University College Dublin
ioana.butnariu@ucd.ie

Su Nam Kim
University of Melbourne
nkim@csse.unimelb.edu.au

Preslav Nakov
National University of Singapore
nakov@comp.nus.edu.sg

Diarmuid Ó Séaghdha
University of Cambridge
do242@cam.ac.uk

Stan Szpakowicz
University of Ottawa
Polish Academy of Sciences
szpak@site.uottawa.ca

Tony Veale
University College Dublin
tony.veale@ucd.ie

Abstract

We present a brief overview of the main challenges in understanding the semantics of noun compounds and consider some known methods. We introduce a new task to be part of SemEval-2010: the interpretation of noun compounds using paraphrasing verbs and prepositions. The task is meant to provide a standard testbed for future research on noun compound semantics. It should also promote paraphrase-based approaches to the problem, which can benefit many NLP applications.

1 Introduction

Noun compounds (NCs) – sequences of two or more nouns acting as a single noun,[1] e.g., *colon cancer tumor suppressor protein* – are abundant in English and pose a major challenge to the automatic analysis of written text. Baldwin and Tanaka (2004) calculated that 3.9% and 2.6% of the tokens in the *Reuters corpus* and the *British National Corpus* (*BNC*), respectively, are part of a noun compound. Compounding is also an extremely productive process in English. The frequency spectrum of compound types follows a Zipfian or power-law distribution (Ó Séaghdha, 2008), so in practice many compound tokens encountered belong to a "long tail" of low-frequency types. For example, over half of the two-noun NC types in the BNC occur just once (Lapata and Lascarides, 2003). Even for relatively frequent NCs that occur ten or more times in the BNC, static English dictionaries give only 27% coverage (Tanaka and Baldwin, 2003). Taken together,

the factors of high frequency and high productivity mean that achieving robust NC interpretation is an important goal for broad-coverage semantic processing. NCs provide a concise means of evoking a relationship between two or more nouns, and natural language processing (NLP) systems that do not try to recover these implicit relations from NCs are effectively discarding valuable semantic information. Broad coverage should therefore be achieved by post-hoc interpretation rather than pre-hoc enumeration, since it is impossible to build a lexicon of all NCs likely to be encountered.

The challenges presented by NCs and their semantics have generated significant ongoing interest in NC interpretation in the NLP community. Representative publications include (Butnariu and Veale, 2008; Girju, 2007; Kim and Baldwin, 2006; Nakov, 2008b; Nastase and Szpakowicz, 2003; Ó Séaghdha and Copestake, 2007). Applications that have been suggested include Question Answering, Machine Translation, Information Retrieval and Information Extraction. For example, a question-answering system may need to determine whether *headaches induced by caffeine withdrawal* is a good paraphrase for *caffeine headaches* when answering questions about the causes of headaches, while an information extraction system may need to decide whether *caffeine withdrawal headache* and *caffeine headache* refer to the same concept when used in the same document. Similarly, a machine translation system facing the unknown NC *WTO Geneva headquarters* might benefit from the ability to paraphrase it as *Geneva headquarters of the WTO* or as *WTO headquarters located in Geneva*. Given a query like *can-*

---

[1] We follow the definition in (Downing, 1977).

cer treatment, an information retrieval system could use suitable paraphrasing verbs like *relieve* and *prevent* for page ranking and query refinement.

In this paper, we introduce a new task, which will be part of the SemEval-2010 competition: NC interpretation using paraphrasing verbs and prepositions. The task is intended to provide a standard testbed for future research on noun compound semantics. We also hope that it will promote paraphrase-based approaches to the problem, which can benefit many NLP applications.

The remainder of the paper is organized as follows: Section 2 presents a brief overview of the existing approaches to NC semantic interpretation and introduces the one we will adopt for SemEval-2010 Task 9; Section 3 provides a general description of the task, the data collection, and the evaluation methodology; Section 4 offers a conclusion.

2 Models of Relational Semantics in NCs

2.1 Inventory-Based Semantics

The prevalent view in theoretical and computational linguistics holds that the semantic relations that implicitly link the nouns of an NC can be adequately enumerated via a small inventory of abstract relational categories. In this view, *mountain hut*, *field mouse* and *village feast* all express 'location in space', while the relation implicit in *history book* and *nativity play* can be characterized as 'topicality' or 'aboutness'. A sample of some of the most influential relation inventories appears in Table 1.

Levi (1978) proposes that complex nominals – a general concept grouping together nominal compounds (e.g., *peanut butter*), nominalizations (e.g., *dream analysis*) and non-predicative noun phrases (e.g., *electric shock*) – are derived through the complementary processes of *recoverable predicate deletion* and *nominalization*; each process is associated with its own inventory of semantic categories. Table 1 lists the categories for the former.

Warren (1978) posits a hierarchical classification scheme derived from a large-scale corpus study of NCs. The top-level relations in her hierarchy are listed in Table 1, while the next level subdivides CONSTITUTE into SOURCE-RESULT, RESULT-SOURCE and COPULA; COPULA is then further subdivided at two additional levels.

In computational linguistics, popular inventories of semantic relations have been proposed by Nastase and Szpakowicz (2003) and Girju et al. (2005), among others. The former groups 30 fine-grained relations into five coarse-grained super-categories, while the latter is a flat list of 21 relations. Both schemes are intended to be suitable for broad-coverage analysis of text. For specialized applications, however, it is often useful to use domain-specific relations. For example, Rosario and Hearst (2001) propose 18 abstract relations for interpreting NCs in biomedical text, e.g., DEFECT, MATERIAL, PERSON AFFILIATED, ATTRIBUTE OF CLINICAL STUDY.

Inventory-based analyses offer significant advantages. Abstract relations such as 'location' and 'possession' capture valuable generalizations about NC semantics in a parsimonious framework. Unlike paraphrase-based analyses (Section 2.2), they are not tied to specific lexical items, which may themselves be semantically ambiguous. They also lend themselves particularly well to automatic interpretation methods based on multi-class classification.

On the other hand, relation inventories have been criticized on a number of fronts, most influentially by Downing (1977). She argues that the great variety of NC relations makes listing them all impossible; creative NCs like *plate length* ('what your hair is when it drags in your food') are intuitively compositional, but cannot be assigned to any standard inventory category. A second criticism is that restricted inventories are too impoverished a representation scheme for NC semantics, e.g., *headache pills* and *sleeping pills* would both be analyzed as FOR in Levi's classification, but express very different (indeed, contrary) relationships. Downing writes (p. 826): "*These interpretations are at best reducible to underlying relationships. . . , but only with the loss of much of the semantic material considered by subjects to be relevant or essential to the definitions.*" A further drawback associated with sets of abstract relations is that it is difficult to identify the "correct" inventory or to decide whether one proposed classification scheme should be favored over another.

2.2 Interpretation Using Verbal Paraphrases

An alternative approach to NC interpretation associates each compound with an explanatory para-

Author(s)	Relation Inventory
Levi (1978)	CAUSE, HAVE, MAKE, USE, BE, IN, FOR, FROM, ABOUT
Warren (1978)	POSSESSION, LOCATION, PURPOSE, ACTIVITY-ACTOR, RESEMBLANCE, CONSTITUTE
Nastase and Szpakowicz (2003)	CAUSALITY (cause, effect, detraction, purpose), PARTICIPANT (agent, beneficiary, instrument, object_property, object, part, possessor, property, product, source, whole, stative), QUALITY (container, content, equative, material, measure, topic, type), SPATIAL (direction, location_at, location_from, location), TEMPORALITY (frequency, time_at, time_through)
Girju et al. (2005)	POSSESSION, ATTRIBUTE-HOLDER, AGENT, TEMPORAL, PART-WHOLE, IS-A, CAUSE, MAKE/PRODUCE, INSTRUMENT, LOCATION/SPACE, PURPOSE, SOURCE, TOPIC, MANNER, MEANS, THEME, ACCOMPANIMENT, EXPERIENCER, RECIPIENT, MEASURE, RESULT
Lauer (1995)	OF, FOR, IN, AT, ON, FROM, WITH, ABOUT

Table 1: Previously proposed inventories of semantic relations for noun compound interpretation. The first two come from linguistic theories; the rest have been proposed in computational linguistics.

phrase. Thus, *cheese knife* and *kitchen knife* can be expanded as a *knife for cutting cheese* and a *knife used in a kitchen*, respectively. In the paraphrase-based paradigm, semantic relations need not come from a small set; it is possible to have many subtle distinctions afforded by the vocabulary of the paraphrasing language (in our case, English). This paradigm avoids the problems of coverage and representational poverty, which Downing (1977) observed in inventory-based approaches. It also reflects cognitive-linguistic theories of NC semantics, in which compounds are held to express underlying *event frames* and whose constituents are held to denote event participants (Ryder, 1994).

Lauer (1995) associates NC semantics with prepositional paraphrases. As Lauer only considers a handful of prepositions (about, at, for, from, in, of, on, with), his model is essentially inventory-based. On the other hand, noun-preposition co-occurrences can easily be identified in a corpus, so an automatic interpretation can be implemented through simple unsupervised methods. The disadvantage of this approach is the absence of a one-to-one mapping from prepositions to meanings; prepositions can be ambiguous (of indicates many different relations) or synonymous (at, in and on all express 'location'). This concern arises with all paraphrasing models, but it is exacerbated by the restricted nature of prepositions. Furthermore, many NCs cannot be paraphrased adequately with prepositions, e.g., *woman driver*, *honey bee*.

A richer, more flexible paraphrasing model is afforded by the use of verbs. In such a model, a *honey bee* is a *bee that produces honey*, a *sleeping pill* is a *pill that induces sleeping* and a *headache pill* is a *pill that relieves headaches*. In some previous computational work on NC interpretation, manually constructed dictionaries provided typical activities or functions associated with nouns (Finin, 1980; Isabelle, 1984; Johnston and Busa, 1996). It is, however, impractical to build large structured lexicons for broad-coverage systems; these methods can only be applied to specialized domains. On the other hand, we expect that the ready availability of large text corpora should facilitate the automatic mining of rich paraphrase information.

The SemEval-2010 task we present here builds on the work of Nakov (Nakov and Hearst, 2006; Nakov, 2007; Nakov, 2008b), where NCs are paraphrased by combinations of verbs and prepositions. Given the problem of synonymy, we do not provide a single correct paraphrase for a given NC but a probability distribution over a range of candidates. For example, highly probable paraphrases for *chocolate bar* are *bar made of chocolate* and *bar that tastes like chocolate*, while *bar that eats chocolate* is very unlikely. As described in Section 3.3, a set of gold-standard paraphrase distributions can be constructed by collating responses from a large number of human subjects.

In this framework, the task of interpretation becomes one of identifying the most likely paraphrases for an NC. Nakov (2008b) and Butnariu and Veale (2008) have demonstrated that paraphrasing information can be collected from corpora in an unsupervised fashion; we expect that participants in

SemEval-2010 Task 9 will further develop suitable techniques for this problem. Paraphrases of this kind have been shown to be useful in applications such as machine translation (Nakov, 2008a) and as an intermediate step in inventory-based classification of abstract relations (Kim and Baldwin, 2006; Nakov and Hearst, 2008). Progress in paraphrasing is therefore likely to have follow-on benefits in many areas.

3 Task Description

The description of the task we present below is preliminary. We invite the interested reader to visit the official Website of SemEval-2010 Task 9, where up-to-date information will be published; there is also a discussion group and a mailing list.[2]

3.1 Preliminary Study

In a preliminary study, we asked 25-30 human subjects to paraphrase 250 noun-noun compounds using suitable paraphrasing verbs. This is the *Levi-250* dataset (Levi, 1978); see (Nakov, 2008b) for details.[3] The most popular paraphrases tend to be quite apt, while some less frequent choices are questionable. For example, for *chocolate bar* we obtained the following paraphrases (the number of subjects who proposed each one is shown in parentheses):

> contain (17); be made of (16); be made from (10); taste like (7); be composed of (7); consist of (5); be (3); have (2); smell of (2); be manufactured from (2); be formed from (2); melt into (2); serve (1); sell (1); incorporate (1); be made with (1); be comprised of (1); be constituted by (1); be solidified from (1); be flavored with (1); store (1); be flavored with (1); be created from (1); taste of (1)

3.2 Objective

We propose a task in which participating systems must estimate the quality of paraphrases for a test set of NCs. A list of verb/preposition paraphrases will be provided for each NC, and for each list a participating system will be asked to provide aptness

scores that correlate well (in terms of frequency distribution) with the human judgments collated from our test subjects.

3.3 Datasets

Trial/Development Data. As trial/development data, we will release the previously collected paraphrase sets for the *Levi-250* dataset (after further review and cleaning). This dataset consists of 250 noun-noun compounds, each paraphrased by 25-30 human subjects (Nakov, 2008b).

Test Data. The test data will consist of approximately 300 NCs, each accompanied by a set of paraphrasing verbs and prepositions. Following the methodology of Nakov (2008b), we will use the *Amazon Mechanical Turk* Web service[4] to recruit human subjects. This service offers an inexpensive way to recruit subjects for tasks that require human intelligence, and provides an API which allows a computer program to easily run tasks and collate the responses from human subjects. The Mechanical Turk is becoming a popular means to elicit and collect linguistic intuitions for NLP research; see Snow et al. (2008) for an overview and a discussion of issues that arise.

We intend to recruit 100 annotators for each NC, and we will require each annotator to paraphrase at least five NCs. Annotators will be given clear instructions and will be asked to produce one or more paraphrases for a given NC. To help us filter out subjects with an insufficient grasp of English or an insufficient interest in the task, annotators will be asked to complete a short and simple multiple-choice pretest on NC comprehension before proceeding to the paraphrasing step.

Post-processing. We will manually check the trial/development data and the test data. Depending on the quality of the paraphrases, we may decide to drop the least frequent verbs.

License. All data will be released under the *Creative Commons Attribution 3.0 Unported license*[5].

3.4 Evaluation

Single-NC Scores. For each NC, we will compare human scores (our gold standard) with those proposed by each participating system. We have con-

---

[2]Please follow the Task #9 link at the SemEval-2010 homepage `http://semeval2.fbk.eu`

[3]This dataset is available from `http://sourceforge.net/projects/multiword/`

[4]`http://www.mturk.com`

[5]`http://creativecommons.org/licenses/by/3.0/`

sidered three scores: (1) Pearson's correlation, (2) cosine similarity, and (3) Spearman's rank correlation.

Pearson's correlation coefficient is a standard measure of the correlation strength between two distributions; it can be calculated as follows:

$$\rho = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - [E(X)]^2}\sqrt{E(Y^2) - [E(Y)]^2}} \quad (1)$$

where $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_n)$ are vectors of numerical scores for each paraphrase provided by the humans and the competing systems, respectively, n is the number of paraphrases to score, and $E(X)$ is the expectation of X.

Cosine correlation coefficient is another popular alternative and was used by Nakov and Hearst (2008); it can be seen as an uncentered version of Pearson's correlation coefficient:

$$\rho = \frac{X.Y}{\|X\|\|Y\|} \quad (2)$$

Spearman's rank correlation coefficient is suitable for comparing rankings of sets of items; it is a special case of Pearson's correlation, derived by considering rank indices (1,2,...) as item scores . It is defined as follows:

$$\rho = \frac{n \sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{n \sum x_i^2 - (\sum x_i)^2}\sqrt{n \sum y_i^2 - (\sum y_i)^2}} \quad (3)$$

One problem with using Spearman's rank coefficient for the current task is the assumption that swapping any two ranks has the same effect. The often-skewed nature of paraphrase frequency distributions means that swapping some ranks is intuitively less "wrong" than swapping others. Consider, for example, the following list of human-proposed paraphrasing verbs for *child actor*, which is given in Nakov (2007):

> be (22); look like (4); portray (3); start as (1); include (1); play (1); have (1); involve (1); act like (1); star as (1); work as (1); mimic (1); pass as (1); resemble (1); be classified as (1); substitute for (1); qualify as (1); act as (1)

Clearly, a system that swaps the positions for *be (22)* and *look like (4)* for *child actor* will have made a significant error, while swapping *contain (17)* and *be made of (16)* for *chocolate bar* (see Section 3.1) would be less inappropriate. However, Spearman's coefficient treats both alterations identically since it only looks at ranks; thus, we do not plan to use it for official evaluation, though it may be useful for post-hoc analysis.

Final Score. A participating system's final score will be the average of the scores it achieves over all test examples.

Scoring Tool. We will provide an automatic evaluation tool that participants can use when training/tuning/testing their systems. We will use the same tool for the official evaluation.

4 Conclusion

We have presented a noun compound paraphrasing task that will run as part of SemEval-2010. The goal of the task is to promote and explore the feasibility of paraphrase-based methods for compound interpretation. We believe paraphrasing holds some key advantages over more traditional inventory-based approaches, such as the ability of paraphrases to represent fine-grained and overlapping meanings, and the utility of the resulting paraphrases for other applications such as Question Answering, Information Extraction/Retrieval and Machine Translation.

The proposed paraphrasing task is predicated on two important assumptions: first, that paraphrasing via a combination of verbs and prepositions provides a powerful framework for representing and interpreting the meaning of compositional nonlexicalized noun compounds; and second, that humans can agree amongst themselves about what constitutes a good paraphrase for any given NC. As researchers in this area and as proponents of this task, we believe that both assumptions are valid, but if the analysis of the task were to raise doubts about either assumption (e.g., by showing poor agreement amongst human annotators), then this in itself would be a meaningful and successful output of the task. As such, we anticipate that the task and its associated dataset will inspire further research, both on the theory and development of paraphrase-based compound interpretation and on its practical applications.

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.

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 2008)*, pages 81–88.

Pamela Downing. 1977. On the creation and use of English compound nouns. *Language*, 53(4):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. 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 2007)*, pages 568–575.

Pierre Isabelle. 1984. Another look at nominal compounds. In *Proceedings of the 10th International Conference on Computational Linguistics*, pages 509–516.

Michael Johnston and Frederica Busa. 1996. Qualia structure and the compositional interpretation of compounds. In *Proceedings of the ACL 1996 Workshop on Breadth and Depth of Semantic Lexicons*, pages 77–88.

Su Nam Kim and Timothy Baldwin. 2006. Interpreting semantic relations in noun compounds via verb semantics. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (COLING/ACL 2006) Main Conference Poster Sessions*, pages 491–498.

Mirella Lapata and Alex Lascarides. 2003. Detecting novel compounds: the role of distributional evidence. In *Proceedings of the 10th conference of the European chapter of the Association for Computational Linguistics (EACL 2003)*, pages 235–242.

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 A. Hearst. 2006. Using verbs to characterize noun-noun relations. In *LNCS vol. 4183: Proceedings of the 12th international conference on Artificial Intelligence: Methodology, Systems and Applications (AIMSA 2006)*, pages 233–244. Springer.

Preslav Nakov and Marti A. Hearst. 2008. Solving relational similarity problems using the web as a corpus. In *Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics (ACL 2008)*, pages 452–460.

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. 2008a. Improved statistical machine translation using monolingual paraphrases. In *Proceedings of the 18th European Conference on Artificial Intelligence (ECAI'2008)*, pages 338–342.

Preslav Nakov. 2008b. Noun compound interpretation using paraphrasing verbs: Feasibility study. In *LNAI vol. 5253: Proceedings of the 13th international conference on Artificial Intelligence: Methodology, Systems and Applications (AIMSA 2008)*, pages 103–117. Springer.

Vivi Nastase and Stan Szpakowicz. 2003. Exploring noun-modifier semantic relations. In *Proceedings of the 5th International Workshop on Computational Semantics*, pages 285–301.

Diarmuid Ó Séaghdha and Ann Copestake. 2007. Co-occurrence contexts for noun compound interpretation. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, pages 57–64.

Diarmuid Ó Séaghdha. 2008. *Learning Compound Noun Semantics*. Ph.D. thesis, University of Cambridge.

Barbara Rosario and Marti Hearst. 2001. Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing (EMNLP 2005)*, pages 82–90.

Mary Ellen Ryder. 1994. *Ordered Chaos: The Interpretation of English Noun-Noun Compounds*. University of California Press, Berkeley, CA.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*, pages 254–263.

Takaaki Tanaka and Timothy Baldwin. 2003. Noun-noun compound machine translation: a feasibility study on shallow processing. In *Proceedings of the ACL 2003 workshop on Multiword expressions*, pages 17–24.

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