SemEval-2013 Task 4: Free Paraphrases of Noun Compounds

Iris Hendrickx
Radboud University Nijmegen &
Universidade de Lisboa
iris@clul.ul.pt

Zornitsa Kozareva
University of Southern California
kozareva@isi.edu

Preslav Nakov
QCRI, Qatar Foundation
pnakov@qf.org.qa

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

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

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

Abstract

In this paper, we describe SemEval-2013 Task 4: the definition, the data, the evaluation and the results. The task is to capture some of the meaning of English noun compounds via paraphrasing. Given a two-word noun compound, the participating system is asked to produce an explicitly ranked list of its free-form paraphrases. The list is automatically compared and evaluated against a similarly ranked list of paraphrases proposed by human annotators, recruited and managed through Amazon's Mechanical Turk. The comparison of raw paraphrases is sensitive to syntactic and morphological variation. The "gold" ranking is based on the relative popularity of paraphrases among annotators. To make the ranking more reliable, highly similar paraphrases are grouped, so as to downplay superficial differences in syntax and morphology. Three systems participated in the task. They all beat a simple baseline on one of the two evaluation measures, but not on both measures. This shows that the task is difficult.

1 Introduction

A noun compound (NC) is a sequence of nouns which act as a single noun (Downing, 1977), as in these examples: *colon cancer*, *suppressor protein*, *tumor suppressor protein*, *colon cancer tumor suppressor protein*, etc. This type of compounding is highly productive in English. NCs comprise 3.9% and 2.6% of all tokens in the Reuters corpus and the British National Corpus (BNC), respectively (Baldwin and Tanaka, 2004).

The frequency spectrum of compound types follows a Zipfian distribution (Ó Séaghdha, 2008), so many NC tokens belong to a "long tail" of low-frequency types. More than half of the two-noun types in the BNC occur exactly once (Kim and Baldwin, 2006). Their high frequency and high productivity make robust NC interpretation an important goal for broad-coverage semantic processing of English texts. Systems which ignore NCs may give up on salient information about the semantic relationships implicit in a text. Compositional interpretation is also the only way to achieve broad NC coverage, because it is not feasible to list in a lexicon all compounds which one is likely to encounter. Even for relatively frequent NCs occurring 10 times or more in the BNC, static English dictionaries provide only 27% coverage (Tanaka and Baldwin, 2003).

In many natural language processing applications it is important to understand the syntax and semantics of NCs. NCs often are structurally similar, but have very different meaning. Consider *caffeine headache* and *ice-cream headache*: a lack of caffeine causes the former, an excess of ice-cream – the latter. Different interpretations can lead to different inferences, query expansion, paraphrases, translations, and so on. A question answering system may have to determine whether *protein acting as a tumor suppressor* is an accurate paraphrase for *tumor suppressor protein*. An information extraction system might need to decide whether *neck vein thrombosis* and *neck thrombosis* can co-refer in the same document. A machine translation system might paraphrase the unknown compound *WTO Geneva headquarters* as *WTO headquarters located in Geneva*.

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 138–143, Atlanta, Georgia, June 14-15, 2013. ©2013 Association for Computational Linguistics

Research on the automatic interpretation of NCs has focused mainly on common two-word NCs. The usual task is to classify the semantic relation underlying a compound with either one of a small number of predefined relation labels or a paraphrase from an open vocabulary. Examples of the former take on classification include (Moldovan et al., 2004; Girju, 2007; Ó Séaghdha and Copestake, 2008; Tratz and Hovy, 2010). Examples of the latter include (Nakov, 2008b; Nakov, 2008a; Nakov and Hearst, 2008; Butnariu and Veale, 2008) and a previous NC paraphrasing task at SemEval-2010 (Butnariu et al., 2010), upon which the task described here builds.

The assumption of a small inventory of predefined relations has some advantages – parsimony and generalization – but at the same time there are limitations on expressivity and coverage. For example, the NCs *headache pills* and *fertility pills* would be assigned the same semantic relation (*PURPOSE*) in most inventories, but their relational semantics are quite different (Downing, 1977). Furthermore, the definitions given by human subjects can involve rich and specific meanings. For example, Downing (1977) reports that a subject defined the NC *oil bowl* as "the bowl into which the oil in the engine is drained during an oil change", compared to which a minimal interpretation *bowl for oil* seems very reductive. In view of such arguments, linguists such as Downing (1977), Ryder (1994) and Coulson (2001) have argued for a fine-grained, essentially open-ended space of interpretations.

The idea of working with fine-grained paraphrases for NC semantics has recently grown in popularity among NLP researchers (Butnariu and Veale, 2008; Nakov and Hearst, 2008; Nakov, 2008a). Task 9 at SemEval-2010 (Butnariu et al., 2010) was devoted to this methodology. In that previous work, the paraphrases provided by human subjects were required to fit a restrictive template admitting only verbs and prepositions occurring between the NC's constituent nouns. Annotators recruited through Amazon Mechanical Turk were asked to provide paraphrases for the dataset of NCs. The gold standard for each NC was the ranked list of paraphrases given by the annotators; this reflects the idea that a compound's meaning can be described in different ways, at different levels of granularity and capturing different interpretations in the case of ambiguity.

For example, a *plastic saw* could be a *saw made of plastic* or a *saw for cutting plastic*. Systems participating in the task were given the set of attested paraphrases for each NC, and evaluated according to how well they could reproduce the humans' ranking.

The design of this task, SemEval-2013 Task 4, is informed by previous work on compound annotation and interpretation. It is also influenced by similar initiatives, such as the English Lexical Substitution task at SemEval-2007 (McCarthy and Navigli, 2007), and by various evaluation exercises in the fields of paraphrasing and machine translation. We build on SemEval-2010 Task 9, extending the task's flexibility in a number of ways. The restrictions on the form of annotators' paraphrases was relaxed, giving us a rich dataset of close-to-freeform paraphrases (Section 3). Rather than ranking a set of attested paraphrases, systems must now both generate and rank their paraphrases; the task they perform is essentially the same as what the annotators were asked to do. This new setup required us to innovate in terms of evaluation measures (Section 4).

We anticipate that the dataset and task will be of broad interest among those who study lexical semantics. We believe that the overall progress in the field will significantly benefit from a public-domain set of free-style NC paraphrases. That is why our primary objective is the challenging endeavour of preparing and releasing such a dataset to the research community. The common evaluation task which we establish will also enable researchers to compare their algorithms and their empirical results.

2 Task description

This is an English NC interpretation task, which explores the idea of interpreting the semantics of NCs via free paraphrases. Given a noun-noun compound such as *air filter*, the participating systems are asked to produce an explicitly ranked list of free paraphrases, as in the following example:

1 filter for air
2 filter of air
3 filter that cleans the air
4 filter which makes air healthier
5 a filter that removes impurities from the air
. . .

Such a list is then automatically compared and evaluated against a similarly ranked list of paraphrases proposed by human annotators, recruited and managed via Amazon's Mechanical Turk. The comparison of raw paraphrases is sensitive to syntactic and morphological variation. The ranking of paraphrases is based on their relative popularity among different annotators. To make the ranking more reliable, highly similar paraphrases are grouped so as to downplay superficial differences in syntax and morphology.

3 Data collection

We used Amazon's *Mechanical Turk* service to collect diverse paraphrases for a range of "gold-standard" NCs.[1] We paid the workers a small fee ($0.10) per compound, for which they were asked to provide five paraphrases. Each paraphrase should contain the two nouns of the compound (in singular or plural inflectional forms, but not in another derivational form), an intermediate non-empty linking phrase and optional preceding or following terms. The paraphrasing terms could have any part of speech, so long as the resulting paraphrase was a well-formed noun phrase headed by the NC's head.

We gave the workers feedback during data collection if they appeared to have misunderstood the nature of the task. Once raw paraphrases had been collected from all workers, we collated them into a spreadsheet, and we merged identical paraphrases in order to calculate their overall frequencies. Ill-formed paraphrases – those violating the syntactic restrictions described above – were manually removed following a consensus decision-making procedure; every paraphrase was checked by at least two task organizers. We did not require that the paraphrases be semantically felicitous, but we performed minor edits on the remaining paraphrases if they contained obvious typos.

The remaining well-formed paraphrases were sorted by frequency separately for each NC. The most frequent paraphrases for a compound are assigned the highest rank 0, those with the next-highest frequency are given a rank of 1, and so on.

	Total	**Min / Max / Avg**
Trial/Train (174 NCs)		
paraphrases	6,069	1 / 287 / 34.9
unique paraphrases	4,255	1 / 105 / 24.5
Test (181 NCs)		
paraphrases	9,706	24 / 99 / 53.6
unique paraphrases	8,216	21 / 80 / 45.4

Table 1: Statistics of the trial and test datasets: the total number of paraphrases with and without duplicates, and the minimum / maximum / average per noun compound.

Paraphrases with a frequency of 1 – proposed for a given NC by only one annotator – always occupy the lowest rank on the list for that compound.

We used 174+181 noun-noun compounds from the NC dataset of Ó Séaghdha (2007). The trial dataset, which we initially released to the participants, consisted of 4,255 human paraphrases for 174 noun-noun pairs; this dataset was also the training dataset. The test dataset comprised paraphrases for 181 noun-noun pairs. The "gold standard" contained 9,706 paraphrases of which 8,216 were unique for those 181 NCs. Further statistics on the datasets are presented in Table 1.

Compared with the data collected for the SemEval-2010 Task 9 on the interpretation of noun compounds, the data collected for this new task have a far greater range of variety and richness. For example, the following (selected) paraphrases for *work area* vary from parsimonious to expansive:

- area for work
- area of work
- area where work is done
- area where work is performed
- ...
- an area cordoned off for persons responsible for work
- an area where construction work is carried out
- an area where work is accomplished and done
- area where work is conducted
- office area assigned as a work space
- ...

[1] Since the annotation on Mechanical Turk was going slowly, we also recruited four other annotators to do the same work, following exactly the same instructions.

4 Scoring

Noun compounding is a generative aspect of language, but so too is the process of NC interpretation: human speakers typically generate a range of possible interpretations for a given compound, each emphasizing a different aspect of the relationship between the nouns. Our evaluation framework reflects the belief that there is rarely a single right answer for a given noun-noun pairing. Participating systems are thus expected to demonstrate some generativity of their own, and are scored not just on the accuracy of individual interpretations, but on the overall breadth of their output.

For evaluation, we provided a scorer implemented, for good portability, as a Java class. For each noun compound to be evaluated, the scorer compares a list of system-suggested paraphrases against a "gold-standard" reference list, compiled and rank-ordered from the paraphrases suggested by our human annotators. The score assigned to each system is the mean of the system's performance across all test compounds. Note that the scorer removes all determiners from both the reference and the test paraphrases, so a system is neither punished for not reproducing a determiner or rewarded for producing the same determiners.

The scorer can match words identically or non-identically. A match of two identical words W_{gold} and W_{test} earns a score of 1.0. There is a partial score of (2 $|P|$ / ($|PW_{gold}|$ + $|PW_{test}|$))^2 for a match of two words PW_{gold} and PW_{test} that are not identical but share a common prefix P, $|P| > 2$, e.g., $wmatch($*cutting*$,$ *cuts*$)$ = $(6/11)^2$ = 0.297.

Two n-grams N_{gold} = $[GW_1, \ldots, GW_n]$ and N_{test} = $[TW_1, \ldots, TW_n]$ can be matched if $wmatch(GW_i, TW_i) > 0$ for all i in $1..n$. The score assigned to the match of these two n-grams is then $\sum_i wmatch(GW_i, TW_i)$. For every n-gram N_{test} = $[TW_1, \ldots, TW_n]$ in a system-generated paraphrase, the scorer finds a matching n-gram N_{gold} = $[GW_1, \ldots, GW_n]$ in the reference paraphrase $Para_{gold}$ which maximizes this sum.

The overall n-gram overlap score for a reference paraphrase $Para_{gold}$ and a system-generated paraphrase $Para_{test}$ is the sum of the score calculated for all n-grams in $Para_{test}$, where n ranges from 1 to the size of $Para_{test}$.

This overall score is then normalized by dividing by the maximum value among the n-gram overlap score for $Para_{gold}$ compared with itself and the n-gram overlap score for $Para_{test}$ compared with itself. This normalization step produces a paraphrase match score in the range [0.0 – 1.0]. It punishes a paraphrase $Para_{test}$ for both over-generating (containing more words than are found in $Para_{gold}$) and under-generating (containing fewer words than are found in $Para_{gold}$). In other words, $Para_{test}$ should ideally reproduce everything in $Para_{gold}$, and nothing more or less.

The reference paraphrases in the "gold standard" are ordered by rank; the highest rank is assigned to the paraphrases which human judges suggested most often. The rank of a reference paraphrase matters because a good participating system will aim to reproduce the top-ranked "gold-standard" paraphrases as produced by human judges. The scorer assigns a multiplier of $R/(R + n)$ to reference paraphrases at rank n; this multiplier asymptotically approaches 0 for the higher values of n of ever lower-ranked paraphrases. We choose a default setting of $R = 8$, so that a reference paraphrase at rank 0 (the highest rank) has a multiplier of 1, while a reference paraphrase at rank 5 has a multiplier of $8/13$ = 0.615.

When a system-generated paraphrase $Para_{test}$ is matched with a reference paraphrase $Para_{gold}$, their normalized n-gram overlap score is scaled by the rank multiplier attaching to the rank of $Para_{gold}$ relative to the other reference paraphrases provided by human judges. The scorer automatically chooses the reference paraphrase $Para_{gold}$ for a test paraphrase $Para_{test}$ so as to maximize this product of normalized n-gram overlap score and rank multiplier.

The overall score assigned to each system for a specific compound is calculated in two different ways: using *isomorphic matching* of suggested paraphrases to the "gold-standard's" reference paraphrases (on a *one-to-one* basis); and using *non-isomorphic matching* of system's paraphrases to the "gold-standard's" reference paraphrases (in a potentially *many-to-one* mapping).

Isomorphic matching rewards both precision and recall. It rewards a system for accurately reproducing the paraphrases suggested by human judges, and for reproducing as many of these as it can, and in much the same order.

In isomorphic mode, system's paraphrases are matched 1-to-1 with reference paraphrases on a first-come first-matched basis, so ordering can be crucial.

Non-isomorphic matching rewards only precision. It rewards a system for accurately reproducing the top-ranked human paraphrases in the "gold standard". A system will achieve a higher score in a non-isomorphic match if it reproduces the top-ranked human paraphrases as opposed to lower-ranked human paraphrases. The ordering of system's paraphrases is thus not important in non-isomorphic matching.

Each system is evaluated using the scorer in both modes, *isomorphic* and *non-isomorphic*. Systems which aim only for precision should score highly on non-isomorphic match mode, but poorly in isomorphic match mode. Systems which aim for precision *and* recall will face a more substantial challenge, likely reflected in their scores.

A naïve baseline

We decided to allow preposition-only paraphrases, which are abundant in the paraphrases suggested by human judges in the crowdsourcing Mechanical Turk collection process. This abundance means that the top-ranked paraphrase for a given compound is often a preposition-only phrase, or one of a small number of very popular paraphrases such as *used for* or *used in*. It is thus straightforward to build a naïve baseline generator which we can expect to score reasonably on this task, at least in *non-isomorphic matching* mode. For each test compound *M H*, the baseline system generates the following paraphrases, in this precise order: *H of M*, *H in M*, *H for M*, *H with M*, *H on M*, *H about M*, *H has M*, *H to M*, *H used for M*, *H used in M*.

This naïve baseline is truly unsophisticated. No attempt is made to order paraphrases by their corpus frequencies or by their frequencies in the training data. The same sequence of paraphrases is generated for each and every test compound.

5 Results

Three teams participated in the challenge, and all their systems were supervised. The MELODI system relied on semantic vector space model built from the UKWAC corpus (window-based, 5 words). It used only the features of the right-hand head noun to train a maximum entropy classifier.

Team	isomorphic	non-isomorphic
SFS	23.1	17.9
IIITH	23.1	25.8
MELODI-Primary	13.0	54.8
MELODI-Contrast	13.6	53.6
Naive Baseline	*13.8*	*40.6*

Table 2: Results for the participating systems; the baseline outputs the same paraphrases for all compounds.

The IIITH system used the probabilities of the preposition co-occurring with a relation to identify the class of the noun compound. To collect statistics, it used Google n-grams, BNC and ANC.

The SFS system extracted templates and fillers from the training data, which it then combined with a four-gram language model and a MaxEnt reranker. To find similar compounds, they used Lin's WordNet similarity. They further used statistics from the English Gigaword and the Google n-grams.

Table 2 shows the performance of the participating systems, SFS, IIITH and MELODI, and the naïve baseline. The baseline shows that it is relatively easy to achieve a moderately good score in non-isomorphic match mode by generating a fixed set of paraphrases which are both common and generic: two of the three participating systems, SFS and IIITH, under-perform the naïve baseline in non-isomorphic match mode, but outperform it in isomorphic mode. The only system to surpass this baseline in non-isomorphic match mode is the MELODI system; yet, it under-performs against the same baseline in isomorphic match mode. No participating team submitted a system which would outperform the naïve baseline in both modes.

6 Conclusions

The conclusions we draw from the experience of organizing the task are mixed. Participation was reasonable but not large, suggesting that NC paraphrasing remains a niche interest – though we believe it deserves more attention among the broader lexical semantics community and hope that the availability of our freeform paraphrase dataset will attract a wider audience in the future.

We also observed a varied response from our annotators in terms of embracing their freedom to generate complex and rich paraphrases; there are many possible reasons for this including laziness, time pressure and the fact that short paraphrases are often very appropriate paraphrases. The results obtained by our participants were also modest, demonstrating that compound paraphrasing is both a difficult task and a novel one that has not yet been "solved".

Acknowledgments

This work has partially supported by a small but effective grant from Amazon; the credit allowed us to hire sufficiently many Turkers – thanks! And a thank-you to our additional annotators Dave Carter, Chris Fournier and Colette Joubarne for their complete sets of paraphrases of the noun compounds in the test data.

References

Timothy Baldwin and Takaaki Tanaka. 2004. Translation by machine of complex nominals: Getting it right. *Proc. ACL04 Workshop on Multiword Expressions: Integrating Processing*, Barcelona, Spain, 24-31.

Cristina Butnariu and Tony Veale. 2008. A concept-centered approach to noun-compound interpretation. *Proc. 22nd International Conference on Computational Linguistics (COLING-08)*, Manchester, UK, 81-88.

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. *Proc. 5th International ACL Workshop on Semantic Evaluation*, Uppsala, Sweden, 39-44.

Seana Coulson. 2001. *Semantic Leaps: Frame-Shifting and Conceptual Blending in Meaning Construction*. Cambridge University Press, Cambridge, UK.

Pamela Downing. 1977. On the creation and use of English compound nouns. *Language*, **53**(4): 810-842.

Roxana Girju. 2007. Improving the interpretation of noun phrases with cross-linguistic information. *Proc. 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic, 568-575.

Su Nam Kim and Timothy Baldwin. 2006. Interpreting semantic relations in noun compounds via verb semantics. *Proc. ACL-06 Main Conference Poster Session*, Sydney, Australia, 491-498.

Diana McCarthy and Roberto Navigli. 2007. Semeval-2007 task 10: English lexical substitution task. *Proc. Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, Prague, Czech Republic, 48-53.

Dan Moldovan, Adriana Badulescu, Marta Tatu, Daniel Antohe, and Roxana Girju. 2004. Models for the semantic classification of noun phrases. Dan Moldovan and Roxana Girju, eds., *HLT-NAACL 2004: Workshop on Computational Lexical Semantics*, Boston, MA, USA, 60-67.

Preslav Nakov and Marti Hearst. 2008. Solving relational similarity problems using the Web as a corpus. *Proc. 46th Annual Meeting of the Association for Computational Linguistics ACL-08*, Columbus, OH, USA, 452-460.

Preslav Nakov. 2008a. Improved statistical machine translation using monolingual paraphrases. *Proc. 18th European Conference on Artificial Intelligence ECAI-08*, Patras, Greece, 338-342.

Preslav Nakov. 2008b. Noun compound interpretation using paraphrasing verbs: Feasibility study. *Proc. 13th International Conference on Artificial Intelligence: Methodology, Systems, Applications AIMSA-08*, Varna, Bulgaria, *Lecture Notes in Computer Science* **5253**, Springer, 103-117.

Diarmuid Ó Séaghdha. 2007. Designing and Evaluating a Semantic Annotation Scheme for Compound Nouns. In *Proceedings of the 4th Corpus Linguistics Conference*, Birmingham, UK.

Diarmuid Ó Séaghdha. 2008. *Learning compound noun semantics*. Ph.D. thesis, Computer Laboratory, University of Cambridge. Published as University of Cambridge Computer Laboratory Technical Report 735.

Diarmuid Ó Séaghdha and Ann Copestake. 2009. Using lexical and relational similarity to classify semantic relations. *Proc. 12th Conference of the European Chapter of the Association for Computational Linguistics EACL-09*, Athens, Greece, 621-629.

Diarmuid Ó Séaghdha and Ann Copestake. 2008. Semantic classification with distributional kernels. In *Proc. 22nd International Conference on Computational Linguistics (COLING-08)*, Manchester, UK.

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

Takaaki Tanaka and Tim Baldwin. 2003. Noun-noun compound machine translation: A feasibility study on shallow processing. *Proc. ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, Sapporo, Japan, 17-24.

Stephen Tratz and Eduard Hovy. 2010. A taxonomy, dataset, and classifier for automatic noun compound interpretation. *Proc. 48th Annual Meeting of the Association for Computational Linguistics ACL-10*, Uppsala, Sweden, 678-687.