

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

Previous research has shown that the meaning of many noun-noun compounds $N_1 N_2$ can be approximated reasonably well by paraphrasing clauses of the form ‘ N_2 that ... N_1 ’, where ‘...’ stands for a verb with or without a preposition. For example, *malaria mosquito* is a ‘*mosquito that carries malaria*’. Evaluating the quality of such paraphrases is the theme of Task 9 at SemEval-2010. This paper describes some background, the task definition, the process of data collection and the task results. We also venture a few general conclusions before the participating teams present their systems at the SemEval-2010 workshop. There were 5 teams who submitted 7 systems.

1 Introduction

Noun compounds (NCs) are sequences of two or more nouns that act as a single noun,¹ e.g., *stem cell*, *stem cell research*, *stem cell research organization*, etc. Lapata and Lascarides (2003) observe that NCs pose syntactic and semantic challenges for three basic reasons: (1) the compounding process is extremely productive in English; (2) the semantic relation between the head and the modifier is implicit; (3) the interpretation can be influenced by contextual and pragmatic factors. Corpus studies have shown that while NCs are very common in English, their frequency distribution follows a Zipfian or power-law distribution and the majority of NCs encountered will be rare types (Tanaka and Baldwin, 2003; Lapata and Lascarides, 2003; Baldwin and Tanaka, 2004; Ó Séaghdha, 2008). As a consequence, Natural Language Processing

(NLP) applications cannot afford either to ignore NCs or to assume that they can be handled by relying on a dictionary or other static resource.

Trouble with lexical resources for NCs notwithstanding, NC semantics plays a central role in complex knowledge discovery and applications, including but not limited to Question Answering (QA), Machine Translation (MT), and Information Retrieval (IR). For example, knowing the (implicit) semantic relation between the NC components can help rank and refine queries in QA and IR, or select promising translation pairs in MT (Nakov, 2008a). Thus, robust semantic interpretation of NCs should be of much help in broad-coverage semantic processing.

Proposed approaches to modelling NC semantics have used semantic similarity (Nastase and Szpakowicz, 2003; Moldovan et al., 2004; Kim and Baldwin, 2005; Nastase and Szpakowicz, 2006; Girju, 2007; Ó Séaghdha and Copestake, 2007) and paraphrasing (Vanderwende, 1994; Kim and Baldwin, 2006; Butnariu and Veale, 2008; Nakov and Hearst, 2008). The former body of work seeks to measure the similarity between known and unseen NCs by considering various features, usually context-related. In contrast, the latter group uses verb semantics to interpret NCs directly, e.g., *olive oil* as ‘*oil that is extracted from olive(s)*’, *drug death* as ‘*death that is caused by drug(s)*’, *flu shot* as a ‘*shot that prevents flu*’.

The growing popularity – and expected direct utility – of paraphrase-based NC semantics has encouraged us to propose an evaluation exercise for the 2010 edition of SemEval. This paper gives a bird’s-eye view of the task. Section 2 presents its objective, data, data collection, and evaluation method. Section 3 lists the participating teams. Section 4 shows the results and our analysis. In Section 5, we try to sum up our experience so far.

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

2 Task Description

2.1 The Objective

For the purpose of the task, we focused on two-word NCs which are modifier-head pairs of nouns, such as *apple pie* or *malaria mosquito*. There are several ways to “attack” the paraphrase-based semantics of such NCs.

We have proposed a rather simple problem: assume that many paraphrases can be found – perhaps via clever Web search – but their relevance is up in the air. Given sufficient training data, we seek to estimate the quality of candidate paraphrases in a test set. Each NC in the training set comes with a long list of verbs in the infinitive (often with a preposition) which may paraphrase the NC adequately. Examples of apt paraphrasing verbs: olive oil – *be extracted from*, drug death – *be caused by*, flu shot – *prevent*. These lists have been constructed from human-proposed paraphrases. For the training data, we also provide the participants with a quality score for each paraphrase, which is a simple count of the number of human subjects who proposed that paraphrase. At test time, given a noun compound and a list of paraphrasing verbs, a participating system needs to produce aptness scores that correlate well (in terms of relative ranking) with the held out human judgments. There may be a diverse range of paraphrases for a given compound, some of them in fact might be inappropriate, but it can be expected that the distribution over paraphrases estimated from a large number of subjects will indeed be representative of the compound’s meaning.

2.2 The Datasets

Following Nakov (2008b), we took advantage of the *Amazon Mechanical Turk*² (MTurk) to acquire paraphrasing verbs from human annotators. The service offers inexpensive access to subjects for tasks which require human intelligence. Its API allows a computer program to run tasks easily and collate the subjects’ responses. MTurk is becoming a popular means of eliciting and collecting linguistic intuitions for NLP research; see Snow et al. (2008) for an overview and a further discussion.

Even though we recruited human subjects, whom we required to take a qualification test,³

²www.mturk.com

³We soon realized that we also had to offer a version of our assignments without a qualification test (at a lower pay rate) since very few people were willing to take a test. Over-

data collection was time-consuming since many annotators did not follow the instructions. We had to monitor their progress and to send them timely messages, pointing out mistakes. Although the MTurk service allows task owners to accept or reject individual submissions, rejection was the last resort since it has the triply unpleasant effect of (1) denying the worker her fee, (2) negatively affecting her rating, and (3) lowering our rating as a requester. We thus chose to try and educate our workers “on the fly”. Even so, we ended up with many examples which we had to correct manually by labor-intensive post-processing. The flaws were not different from those already described by Nakov (2008b). Post-editing was also necessary to lemmatize the paraphrasing verbs systematically.

Trial Data. At the end of August 2009, we released as trial data the previously collected paraphrase sets (Nakov, 2008b) for the *Levi-250* dataset (after further review and cleaning). This dataset consisted of 250 noun-noun compounds form (Levi, 1978), each paraphrased by 25-30 MTurk workers (without a qualification test).

Training Data. The training dataset was an extension of the trial dataset. It consisted of the same 250 noun-noun compounds, but the number of annotators per compound increased significantly. We aimed to recruit at least 30 additional MTurk workers per compound; for some compounds we managed to get many more. For example, when we added the paraphrasing verbs from the trial dataset to the newly collected verbs, we had 131 different workers for *neighborhood bars*, compared to just 50 for *tear gas*. On the average, we had 72.7 workers per compound. Each worker was instructed to try to produce at least three paraphrasing verbs, so we ended up with 191.8 paraphrasing verbs per compound, 84.6 of them being unique. See Table 1 for more details.

Test Data. The test dataset consisted of 388 noun compounds collected from two data sources: (1) the Nastase and Szpakowicz (2003) dataset; and (2) the Lauer (1995) dataset. The former contains 328 noun-noun compounds (there are also a number of adjective-noun and adverb-noun pairs), while the latter contains 266 noun-noun compounds. Since these datasets overlap between themselves and with the training dataset, we had to exclude some examples. In the end, we had 388

all, we found little difference in the quality of work of subjects recruited with and without the test.

	Training: 250 NCs		Testing: 388 NCs		All: 638 NCs	
	Total	Min/Max/Avg	Total	Min/Max/Avg	Total	Min/Max/Avg
MTurk workers	28,199	50/131/72.7	17,067	57/96/68.3	45,266	50/131/71.0
Verb types	32,832	25/173/84.6	17,730	41/133/70.9	50,562	25/173/79.3
Verb tokens	74,407	92/462/191.8	46,247	129/291/185.0	120,654	92/462/189.1

Table 1: **Statistics about the the training/test datasets.** Shown are the total number of verbs proposed as well as the minimum, maximum and average number of paraphrasing verb types/tokens per compound.

unique noun-noun compounds for testing, distinct from those used for training. We aimed for 100 human workers per testing NC, but we could only get 68.3, with a minimum of 57 and a maximum of 96; there were 185.0 paraphrasing verbs per compound, 70.9 of them being unique, which is close to what we had for the training data.

Data format. We distribute the training data as a raw text file. Each line has the following tab-separated format:

```
NC paraphrase frequency
```

where NC is a noun-noun compound (e.g., *apple cake*, *flu virus*), paraphrase is a human-proposed paraphrasing verb optionally followed by a preposition, and frequency is the number of annotators who proposed that paraphrase. Here is an illustrative extract from the training dataset:

```
flu virus cause 38
flu virus spread 13
flu virus create 6
flu virus give 5
flu virus produce 5
...
flu virus be made up of 1
flu virus be observed in 1
flu virus exacerbate 1
```

The test file has a similar format, except that the frequency is not included and the paraphrases for each noun compound appear in random order:

```
...
chest pain originate
chest pain start in
chest pain descend in
chest pain be in
...
```

License. All datasets are released under the *Creative Commons Attribution 3.0 Unported license*.⁴

⁴creativecommons.org/licenses/by/3.0

2.3 Evaluation

All evaluation was performed by computing an appropriate measure of similarity/correlation between system predictions and the compiled judgements of the human annotators. We did it on a compound-by-compound basis and averaged over all compounds in the test dataset. Section 4 shows results for three measures: Spearman rank correlation, Pearson correlation, and cosine similarity.

Spearman Rank Correlation (ρ) was adopted as the official evaluation measure for the competition. As a rank correlation statistic, it does not use the numerical values of the predictions or human judgements, only their relative ordering encoded as integer ranks. For a sample of n items ranked by two methods x and y , the rank correlation ρ is calculated as follows:

$$\rho = \frac{n \sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}} \quad (1)$$

where x_i, y_i are the ranks given by x and y to the i th item, respectively. The value of ρ ranges between -1.0 (total negative correlation) and 1.0 (total positive correlation).

Pearson Correlation (r) is a standard measure of correlation strength between real-valued variables. The formula is the same as (1), but with x_i, y_i taking real values rather than rank values; just like ρ , r 's values fall between -1.0 and 1.0.

Cosine similarity is frequently used in NLP to compare numerical vectors:

$$\cos = \frac{\sum_i^n x_i y_i}{\sqrt{\sum_i^n x_i^2} \sqrt{\sum_i^n y_i^2}} \quad (2)$$

For non-negative data, the cosine similarity takes values between 0.0 and 1.0. Pearson's r can be viewed as a version of the cosine similarity which performs centering on x and y .

Baseline: To help interpret these evaluation measures, we implemented a simple baseline. A distribution over the paraphrases was estimated by

summing the frequencies for all compounds in the training dataset, and the paraphrases for the test examples were scored according to this distribution. Note that this baseline entirely ignores the identity of the nouns in the compound.

3 Participants

The task has attracted five teams, one of which (UCD-GOGGLE) submitted three runs. The participants are listed in Table 2 along with brief system descriptions; for more details please see the teams’ own description papers.

4 Results and Discussion

The task results appear in Table 3. In an evaluation by Spearman’s ρ (the official ranking measure), the winning system was UVT-MEPHISTO, which scored 0.450. UVT also achieved the top Pearson’s r score. UCD-PN is the top-scoring system according to the cosine measure. One participant submitted part of his results after the official deadline, which is marked by an asterisk.

The participants used a variety of information sources and estimation methods. UVT-MEPHISTO is a supervised system that uses frequency information from the Google N-Gram Corpus and features from WordNet (Fellbaum, 1998) to rank candidate paraphrases. On the other hand, UCD-PN uses no external resources and no supervised training, yet came within 0.009 of UVT-MEPHISTO in the official evaluation. The basic idea of UCD-PN – that one can predict the plausibility of a paraphrase simply by knowing which other paraphrases have been given for that compound *regardless of their frequency* – is clearly a powerful one. Unlike the other systems, UCD-PN used information about the test examples (not their ranks, of course) for model estimation; this has similarities to “transductive” methods for semi-supervised learning. Post-hoc analysis shows UCD-PN would have preserved its rank if it had estimated its model on the training data only.

The other systems are comparable to UVT-MEPHISTO in that they use corpus frequencies to evaluate paraphrases and apply some kind of semantic smoothing to handle sparsity. However, UCD-GOGGLE-I, UCAM and NC-INTERP are unsupervised systems. UCAM uses the 100-million word BNC corpus, while the other systems use Web-scale resources; this has presumably exacerbated sparsity issues and contributed to a rela-

tively poor performance.

The *hybrid* approach exemplified by UCD-GOGGLE-III combines the predictions of a system that models paraphrase correlations and one that learns from corpus frequencies and thus attains better performance. Given that the two top-scoring systems can also be characterized as using these two distinct information sources, it is natural to consider combining these systems. Simply normalizing and averaging the two sets of predictions for each compound does indeed give better scores: Spearman $\rho = 0.472$, $r = 0.431$, Cosine = 0.685.

The baseline from Section 2.3 turns out to be very strong. In Spearman’s ρ evaluation, only three systems outperform it. It is less competitive in the other evaluation measures though. This suggests that global paraphrase frequencies may be useful for telling sensible paraphrases from bad ones, but will not do for quantifying the plausibility of a paraphrase for a given noun compound.

5 Conclusion

Given that it has been a newly-proposed task, this initial experiment in paraphrasing noun compounds has been a moderate success. The participation rate has been sufficient for the purposes of comparing and contrasting different approaches to the role of paraphrases in the interpretation of noun-noun compounds. We have seen a variety of approaches applied to the same dataset, and we have been able to compare the performance of *pure* approaches to *hybrid* approaches, and of supervised approaches to unsupervised approaches. The actual results reported here are also encouraging for a first-time task, though clearly there is considerable room for improvement.

This task has established a high baseline for systems to beat. We can take heart from the fact that the best performance is apparently obtained from a combination of corpus-derived usage features and dictionary-derived linguistic knowledge. Although clever but simple approaches can do quite well on such a task, it is encouraging to note that the best results await those who employ the most robust and the most informed treatments of NCs and their paraphrases. Despite a good start, this is a challenge that remains resolutely open. We expect that the dataset created for the task will be a valuable resource for future research.

System	Institution	Team	Description
NC-INTERP	International Institute of Information Technology, Hyderabad	Prashant Mathur	Unsupervised model using verb-argument frequencies from parsed Web snippets and WordNet smoothing
UCAM	University of Cambridge	Clemens Heppner	Unsupervised model using verb-argument frequencies from the British National Corpus
UCD-GOGGLE-I	University College Dublin	Guofu Li	Unsupervised probabilistic model using pattern frequencies estimated from the Google N-Gram corpus
UCD-GOGGLE-II			Paraphrase ranking model learned from training data
UCD-GOGGLE-III			Combination of UCD-GOGGLE-I and UCD-GOGGLE-II
UCD-PN	University College Dublin	Paul Nulty	Scoring according to the probability of a paraphrase appearing in the same set as other paraphrases provided
UVT-MEPHISTO	Tilburg University	Sander Wubben	Supervised memory-based ranker using features from Google N-Gram Corpus and WordNet

Table 2: The participants in SemEval-2010 Task 9.

Rank	System	Supervised?	Hybrid?	Spearman ρ	Pearson r	Cosine
1	UVT-MEPHISTO	yes	no	0.450	0.411	0.635
2	UCD-PN	no	no	0.441	0.361	0.669
3	UCD-GOGGLE-III	yes	yes	0.432	0.395	0.652
4	UCD-GOGGLE-II	yes	no	0.418	0.375	0.660
5	UCD-GOGGLE-I	no	no	0.380	0.252	0.629
6	UCAM	no	no	0.267	0.219	0.374
7	NC-INTERP*	no	no	0.186	0.070	0.466
	Baseline	yes	no	0.425	0.344	0.524
	Combining UVT and UCD-PN	yes	yes	0.472	0.431	0.685

Table 3: Evaluation results for SemEval-2010 Task 9. Shown are Spearman Rank Correlation, Pearson Correlation and Cosine Similarity; the former is the official score. Overall, supervised systems tend to outperform unsupervised ones, with the notable exception of UCD-PN. Interestingly, our simple baseline scores quite high on Spearman Rank Correlation. There is also quite a lot to gain from combining the top two systems. (* denotes a late submission).

Acknowledgements

This work is partially supported by grants from Amazon and from the Bulgarian National Science Foundation (D002-111/15.12.2008 – *SmartBook*).

References

- Timothy Baldwin and Takaaki Tanaka. 2004. Translation by Machine of Compound Nominals: Getting it Right. In *Proceedings of the ACL-04 Workshop on Multiword Expressions: Integrating Processing*, pages 24–31, Barcelona, Spain.
- Cristina Butnariu and Tony Veale. 2008. A Concept-Centered Approach to Noun-Compound Interpretation. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING-08)*, pages 81–88, Manchester, UK.
- Pamela Downing. 1977. On the creation and use of English compound nouns. *Language*, 53(4):810–842.
- Christiane Fellbaum, editor. 1998. *WordNet: an electronic lexical database*. MIT Press.
- Roxana Girju. 2007. Improving the Interpretation of Noun Phrases with Cross-linguistic Information. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL-07)*, pages 568–575, Prague, Czech Republic.
- Su Nam Kim and Timothy Baldwin. 2005. Automatic interpretation of noun compounds using WordNet similarity. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP-05)*, pages 945–956, Jeju Island, South Korea.
- Su Nam Kim and Timothy Baldwin. 2006. Interpreting Semantic Relations in Noun Compounds via Verb Semantics. In *Proceedings of the COLING-ACL-06 Main Conference Poster Sessions*, pages 491–498, Sydney, Australia.
- 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-03)*, pages 235–242, Budapest, Hungary.
- Mark Lauer. 1995. *Designing Statistical Language Learners: Experiments on Noun Compounds*. Ph.D. thesis, Macquarie University.
- Judith Levi. 1978. *The Syntax and Semantics of Complex Nominals*. Academic Press, New York, NY.
- Dan Moldovan, Adriana Badulescu, Marta Tatu, Daniel Antohe, and Roxana Girju. 2004. Models for the Semantic Classification of Noun Phrases. In *Proceedings of the HLT-NAACL-04 Workshop on Computational Lexical Semantics*, pages 60–67, Boston, MA.
- 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-08)*, pages 452–460, Columbus, OH.
- Preslav Nakov. 2008a. Improved Statistical Machine Translation Using Monolingual Paraphrases. In *Proceedings of the 18th European Conference on Artificial Intelligence (ECAI-08)*, pages 338–342, Patras, Greece.
- Preslav Nakov. 2008b. Noun Compound Interpretation Using Paraphrasing Verbs: Feasibility Study. In *Proceedings of the 13th International Conference on Artificial Intelligence: Methodology, Systems and Applications (AIMSA-08)*, pages 103–117, Varna, Bulgaria.
- Vivi Nastase and Stan Szpakowicz. 2003. Exploring noun-modifier semantic relations. In *Proceedings of the 5th International Workshop on Computational Semantics (IWCS-03)*, pages 285–301, Tilburg, The Netherlands.
- Vivi Nastase and Stan Szpakowicz. 2006. Matching syntactic-semantic graphs for semantic relation assignment. In *Proceedings of the 1st Workshop on Graph Based Methods for Natural Language Processing (TextGraphs-06)*, pages 81–88, New York, NY.
- Diarmuid Ó Séaghdha and Ann Copestake. 2007. Co-occurrence Contexts for Noun Compound Interpretation. In *Proceedings of the ACL-07 Workshop on A Broader Perspective on Multiword Expressions (MWE-07)*, pages 57–64, Prague, Czech Republic.
- Diarmuid Ó Séaghdha. 2008. *Learning Compound Noun Semantics*. Ph.D. thesis, University of Cambridge.
- 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-08)*, pages 254–263, Honolulu, HI.
- Takaaki Tanaka and Timothy Baldwin. 2003. Noun-noun compound machine translation: A feasibility study on shallow processing. In *Proceedings of the ACL-03 Workshop on Multiword Expressions (MWE-03)*, pages 17–24, Sapporo, Japan.
- Lucy Vanderwende. 1994. Algorithm for Automatic Interpretation of Noun Sequences. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING-94)*, pages 782–788, Kyoto, Japan.