

Standardised Evaluation of English Noun Compound Interpretation

Su Nam Kim^{♠◇}, Timothy Baldwin[◇]

National University of Singapore[♠]
Department of Computer Science, School of Computing, Singapore
kimsn@comp.nus.edu.sg
and
University of Melbourne[◇]
Department of CSSE, Carlton, Victoria, Melbourne
{snkim,tim}@csse.unimelb.edu.au

Abstract

We present a tagged corpus for English noun compound interpretation and describe the method used to generate them. In order to collect noun compounds, we extracted binary noun compounds (i.e. noun-noun pairs) by looking for sequences of two nouns in the POS tag data of the Wall Street Journal. We then manually filtered out all noun compounds which were incorrectly tagged or included proper nouns. This left us with a data set of 2,169 noun compounds, which we annotated using a set of 20 semantic relations defined by Barker and Szpakowicz (1998) allowing the annotators to assign multiple semantic relations if necessary. The initial agreement was 52.31%. The final data set contains 1,081 test noun compounds and 1,088 training noun compounds.

1. Introduction

Noun compounds (or NCs), such as *computer science* and *paper submission*, have received significant attention in the linguistic and computational linguistic literature over the past couple of decades, particularly in the area of interpreting the semantic relations between a head noun and its modifier(s). We define **noun compounds** as sequences of nouns contained in a single NP, where the rightmost noun is the NP head.

Semantic relations (or SRs) are directed binary predicates which represent the nature of the semantic link between the component nouns of an NC. For example, the semantic relation for *orange juice* is MATERIAL, indicating that the modifier, *orange* is the material from which the *juice* (head noun) is made. On the other hand, *morning juice* is of type TIME, indicating that the *juice* has some temporal significance (i.e. *morning*). Since NCs are both highly productive and semantically underspecified (Lapata, 2002), the task of interpreting them is not easy. Moreover, the interpretation can differ depending on contextual and pragmatic factors.

Interpreting SRs has been undertaken in the past from both linguistic and computational perspectives. Defining possible semantic relations in noun compounds has been studied and proposed from different perspectives (Levi, 1979; Finin, 1980; Vanderwende, 1994; Saghdha and Copestake, 2007). One approach has been to propose a predefined number of SRs to interpret NCs (Levi, 1979; Sparck Jones, 1983), while a second has been to suggest that there is an unbounded set of SRs, and propose a context-sensitive means of interpretation (Finin, 1980). The main issues here are the granularity and coverage of the SRs, and distribution of different SRs in a given dataset (Saghdha and Copestake, 2007). Since the set of SRs directly influences the automatic interpretation task, it is necessary to agree on a standard set of SRs. Unfortunately, however, this is still under active debate.

On the other hand, recent approaches to automatic interpretation have achieved success to a certain degree based

on supervised learning approaches (Moldovan et al., 2004; Kim and Baldwin, 2005; Kim and Baldwin, 2006; Nastase et al., 2006; Nakov and Hearst, 2006; Girju, 2007). The two main basic supervised approaches to the interpretation of NCs have been semantic similarity (Moldovan et al., 2004; Kim and Baldwin, 2005; Girju, 2007) ellipsed predicate recovery (Kim and Baldwin, 2006; Nakov and Hearst, 2006). In addition, the workshop on semantic evaluation 2007 (SemEval-2007) provided a standardised (sub)set of SRs and NC instances over which to compare approaches to NC interpretation. Although the problem was restricted to binary classification (i.e. is a given NC compatible with a given SR or not), it provided a great opportunity to gather many ideas and clearly showcased promising approaches for future study. Moreover, it allowed researchers to understand problems such as the impact of label and training data bias on interpretation (Girju et al., 2007).

Our goal in this work is to outline a standardised data set for noun compound interpretation, which we hope will complement the SemEval-2007 dataset in furthering empirical research on NC interpretation.

In the following sections, we describe where and how we obtained the NCs in the dataset (Section 2.), describe semantic relations used in the data set (Section 3.), describe the annotation procedure (Section 4.), and summarise our contribution (Section 5.).

2. Data

In this section, we describe the data collection process. We will also look at the statistics of the data set and data format, including examples.

2.1. Data Collection

First, we selected a sample set of NCs based on the gold-standard part-of-speech (POS) tags in the Wall Street Journal at Penn Treebank, by retrieving all binary noun compounds (i.e. noun-noun pairs). That is, if two nouns appeared contiguously at least once, we tagged that combination as an NC candidate. Second, we excluded NCs that

	Total	Test	Train
Total no. of NCs	2169	1081	1088
No. of (NC,SR) pairs	2347	1163	1184
No. of NCs with multiple SRs	178	82	96

Table 1: Composition of the data set

contained proper nouns such as country names, or names of people/companies from our data set (e.g. *Apple computer*, *Melbourne Cup*). We then manually filtered out any false positives from the remaining data. For example, in the sentence *.. was a library students used ..* we initially retrieved *library students* based on the simple POS information, which we later excluded on the basis of not occurring in the a single NP. We also excluded binary NCs that are part of larger NCs. For example, given *computer science department* we would not extract out either *computer science* or *science department*. Naturally, however, if *computer science* or *science department* occurred separately as a binary NC, they would be included in the data set.

2.2. Data Split and Annotation

The total number of NCs in our data set is 2,169. We split the data into approximately 50% for test and the remainder for training. The number of test and training NCs are 1,081 and 1,088, respectively.

In order to annotate the data set, we hired two human annotators and trained them over 200 held-out NCs to familiarise them with the annotation task and set of SRs (see Section 4.). The SRs used for the annotation were taken from Barker and Szpakowicz (1998), as detailed in Section 3..

The annotators were instructed to tag with a unique SR where possible, but also that multiple SRs were allowed in instances of genuine ambiguity. For example, *cable operator* can be interpreted as corresponding to the SR TOPIC (as in *operator is concerned with cable(s)*) or alternatively OBJECT (as in *cable is acted on by operator*). On completion of the annotation, the two annotators were instructed to come together and resolve any disputes in annotation. In the final dataset, about 8.2% of the NCs were tagged with multiple semantic relations.

We present a breakdown of the final dataset in Table 1, in terms of the total number of NCs, the number of discrete pairings of NC and SR, and the number of noun compounds which have multiple SRs.

2.3. Data Format

The data format is simple, consisting of a single NC and its SR(s) per line. The nouns are space delimited, and the SRs are tab-delimited from the NC. In the files, the NCs are sorted in ascending order. An excerpt of the file is listed in Table 2.

3. Semantic Relations

To annotate our NCs, rather than attempting to define a new set of SRs, we used the set defined by Barker and Szpakowicz (1998). In detail, the authors defined 20 SRs for NCs and provided definitions and examples for each. Later, Nastase et al. (2006) classified these into 5 super-classes for

chest pain	source
computer expert	topic
printer tray	cause
student loan	object
student protest	agent

Table 2: Data sample

their own usage. The SRs are detailed in Table 3., along with the number of test and training instances containing in the data set. Note that the SRs were developed for a more general class of data than our NCs, including adjective-noun compounds. Hence, some of examples contain adjective as modifiers (e.g. *charitable compound* for BENEFICIARY and *late supper* for TIME).

In Table 3., the final column details the number of NCs tagged with each SR in the test and training data sets, respectively. The numbers in parentheses indicate the number of instances for each subset of the data that are tagged with multiple relations.

4. Annotation

We describe the annotation methodology in this section. First, we briefly describe our human annotators. Second, we outline the annotator training procedure. Finally, we describe the annotation procedure in detail.

4.1. Human Annotators

Our human annotators were two PhD students. One is an English native speaker with some experience in computational linguistics, and the other (the first author) is a non-native English speaker with wide experience in computational linguistics and various annotation tasks.

4.2. Training the Annotators

To train our human annotators, we collected 200 noun compounds not contained in our final data set. The training procedure started with an introduction to the set of semantic relations in Barker and Szpakowicz (1998) that enabled them to understand the semantic relations and differences between them. We then made the annotators independently tag the 200 held-out NCs, and had them come together to compare their annotations. In the case of disagreements, they discussed their respective annotations and tried to agreed upon a unique SR. However, when they could not come up with a unique mutually-agreeable SR, they were allowed to assign both SRs. We also allowed them to individually assign multiple SRs in instances of genuine ambiguity, such as *cotton bag*, which can be interpreted as either MATERIAL (*bag made of cotton*) or PURPOSE (*bag for cotton*).

4.3. Annotation Procedure

The process for annotating the data set was similar to that described for training the annotators. The final annotation was performed over the 2,169 noun compounds, allowing multiple SRs. For all cases of disagreement, the two annotators came together to discuss their respective annotations and agree on a finalise set of SRs.

<i>Relation</i>	<i>Definition</i>	<i>Example</i>	<i>Test/training instances</i>
AGENT	N_2 is performed by N_1	<i>student protest, band concert, military assault</i>	10(1)/5(0)
BENEFICIARY	N_1 benefits from N_2	<i>student price, charitable compound</i>	10(1)/7(1)
CAUSE	N_1 causes N_2	<i>printer tray, flood water, film music, story idea</i>	54(5)/74(3)
CONTAINER	N_1 contains N_2	<i>exam anxiety, overdue fine</i>	13(4)/19(3)
CONTENT	N_1 is contained in N_2	<i>paper tray, eviction notice, oil pan</i>	40(2)/34(2)
DESTINATION	N_1 is destination of N_2	<i>game bus, exit route, entrance stairs</i>	1(0)/2(0)
EQUATIVE	N_1 is also head	<i>composer arranger, player coach</i>	9(0)/17(1)
INSTRUMENT	N_1 is used in N_2	<i>electron microscope, diesel engine, laser printer</i>	6(0)/11(0)
LOCATED	N_1 is located at N_2	<i>building site, home town, solar system</i>	12(1)/16(2)
LOCATION	N_1 is the location of N_2	<i>lab printer, desert storm, internal combustion</i>	29(9)/24(4)
MATERIAL	N_2 is made of N_1	<i>carbon deposit, gingerbread man, water vapour</i>	12(0)/14(1)
OBJECT	N_1 is acted on by N_2	<i>engine repair, horse doctor</i>	88(6)/88(5)
POSSESSOR	N_1 has N_2	<i>student loan, company car, national debt</i>	33(1)/22(1)
PRODUCT	N_1 is a product of N_2	<i>automobile factory, light bulb, color printer</i>	27(0)/32(6)
PROPERTY	N_2 is N_1	<i>elephant seal, fairy penguin</i>	76(3)/85(3)
PURPOSE	N_2 is meant for N_1	<i>concert hall, soup pot, grinding abrasive</i>	159(13)/161(9)
RESULT	N_1 is a result of N_2	<i>storm cloud, cold virus, death penalty</i>	7(0)/8(0)
SOURCE	N_1 is the source of N_2	<i>chest pain, north wind, foreign capital</i>	86(11)/99(15)
TIME	N_1 is the time of N_2	<i>winter semester, morning class, late supper</i>	26(1)/19(0)
TOPIC	N_2 is concerned with N_1	<i>computer expert, safety standard, horror novel</i>	465(24)/447(39)

Table 3: The set of semantic relations (N_1 = modifier, N_2 = head noun)

The initial agreement for the two annotators was 52.31%, with instances of the annotators agreeing on at least one SR being classified as agreement. Common confusion pairs amongst the initial disagreements were SOURCE and CAUSE, PURPOSE and TOPIC, and OBJECT and TOPIC.

5. Summary

In this paper, we have presented a dataset for English noun compound interpretation. We collected 2,169 English noun compounds from the POS-tagged Wall Street Journal at Penn Treebank, and annotated each NC type according to the 20 SRs defined by Barker and Szpakowicz (1998). Finally, we split the overall dataset into 1,081 and 1,088 instances for test and training, respectively.

During the annotation task, we confirmed that the agreement between human annotators for the NC interpretation task is low (Moldovan et al., 2004; Saghdha and Copestake, 2007). We also noticed that some NCs can be interpreted with multiple SRs, according to context (Downing, 1977). Finally, we reaffirm that defining and annotating SRs for NCs is a non-trivial task, and we hope that our data provides a reliable resource for further research.

6. References

- Ken Barker and Stan Szpakowicz. 1998. Semi-automatic recognition of noun modifier relationships. In *Proceedings of the 17th International Conference on Computational Linguistics (COLING-1998)*, pages 96–102, Montreal, Canada.
- Pamela Downing. 1977. On the creation and use of English compound nouns. *Language*, 53(4):810–842.
- Timothy Wilking Finin. 1980. *The semantic interpretation of compound nominals*. Ph.D. thesis, University of Illinois, Urbana, Illinois, USA.
- 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 the 4th Semantic Evaluation Workshop (SemEval-2007)*, pages 13–18, Prague, Czech Republic.
- 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*, pages 568–575, Prague, Czech Republic.
- Su Nam Kim and Timothy Baldwin. 2005. Automatic interpretation of compound nouns using wordnet::similarity. In *Proceedings of 2nd International Joint Conference on Natural Language Processing (IJCNLP-2005)*, pages 945–956, Jeju, Korea.
- Su Nam Kim and Timothy Baldwin. 2006. Interpreting semantic relations in noun compounds via verb semantics. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics and 21st International Conference on Computational Linguistics (COLING/ACL-2006)*, pages 491–498, Sydney, Australia.
- Maria Lapata. 2002. The disambiguation of nominalizations. *Computational Linguistics*, 28(3):357–388.
- Judith Levi. 1979. *The Syntax and Semantics of Complex Nominals*. Academic Press, New York, New York, USA.
- Dan Moldovan, Adriana Badulescu, Marta Tatu, Daniel Antohe, and Roxana Girju. 2004. Models for the semantic classification of noun phrases. In *Proceedings of HLT-NAACL 2004: Workshop on Computational Lexical Semantics*, pages 60–67, Boston, Massachusetts, USA.
- Preslav Nakov and Marti Hearst. 2006. Using verbs to

- characterize noun-noun relations. In *Proceedings of the 12th International Conference on Artificial Intelligence: Methodology, Systems, Applications (AIMSA)*, pages 233–244, Bularia.
- Vivi Nastase, Jelber Sayyad-Shirabad, Marina Sokolova, and Stan Szpakowicz. 2006. Learning noun-modifier semantic relations with corpus-based and wordnet-based features. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI)*, pages 781–787, Boston, Massachusetts, USA.
- Karen Sparck Jones. 1983. *Compound noun interpretation problems*. Prentice-Hall, Englewood Cliffs, NJ, USA.
- Diarmuid Saghda and Ann Copestake. 2007. Co-occurrence contexts for noun compound interpretation. In *Proceedings of the ACL-2007 Workshop on A Broader Perspective on Multiword Expressions*, pages 57–64, Prague, Czech Republic.
- Lucy Vanderwende. 1994. Algorithm for automatic interpretation of noun sequences. In *Proceedings of the 15th Conference on Computational linguistics*, pages 782–788, Kyoto, Japan.