Annotating and Learning Compound Noun Semantics

Diarmuid Ó Séaghdha
University of Cambridge Computer Laboratory
15 JJ Thomson Avenue
Cambridge CB3 0FD
United Kingdom
do242@cam.ac.uk

Abstract

There is little consensus on a standard experimental design for the compound interpretation task. This paper introduces well-motivated general desiderata for semantic annotation schemes, and describes such a scheme for in-context compound annotation accompanied by detailed publicly available guidelines. Classification experiments on an open-text dataset compare favourably with previously reported results and provide a solid baseline for future research.

1 Introduction

There are a number of reasons why the interpretation of noun-noun compounds has long been a topic of interest for NLP researchers. Compounds occur very frequently in English and many other languages, so they cannot be avoided by a robust semantic processing system. Compounding is a very productive process with a highly skewed type frequency spectrum, and corpus information may be very sparse. Compounds are often highly ambiguous and a large degree of "world knowledge" seems necessary to understand them. For example, knowing that a **cheese knife** is (probably) a knife for cutting cheese and (probably) not a knife made of cheese (cf. **plastic knife**) does not just require an ability to identify the senses of **cheese** and **knife** but also knowledge about what one usually does with cheese and knives. These factors combine to yield a difficult problem that exhibits many of the challenges characteristic of lexical semantic processing in general. Recent research has made significant progress on solving the problem with statistical methods and often without the need for manually created lexical resources (Lauer, 1995; Lapata and Keller, 2004; Girju, 2006; Turney, 2006). The work presented here is part of an ongoing project that treats compound interpretation as a classification problem to be solved using machine learning.

2 Selecting an Annotation Scheme

For many classification tasks, such as part-of-speech tagging or word sense disambiguation, there is general agreement on a standard set of categories that is used by most researchers. For the compound interpretation task, on the other hand, there is little agreement and numerous classification schemes have been proposed. This hinders meaningful comparison of different methods and results. One must therefore consider how an appropriate annotation scheme should be chosen.

One of the problems is that it is not immediately clear what level of granularity is desirable, or even what kind of units the categories should be. Lauer (1995) proposes a set of 8 prepositions that can be used to paraphrase compounds: a **cheese knife** is a *knife FOR cheese* but a **kitchen knife** is a *knife (used) IN a kitchen*. An advantage of this approach is that preposition-noun co-occurrences can efficiently be mined from large corpora using shallow techniques. On the other hand, interpreting a paraphrase requires further disambiguation as one preposition can map onto many semantic relations.[1] Girju et al. (2005) and Nastase and Szpakowicz (2003) both present large inventories of semantic relations that describe noun-noun dependencies. Such relations provide richer semantic information, but it is harder for both humans and machines to identify their occurrence in text. Larger inventories can also suffer from class sparsity; for example, 14 of Girju et al.'s 35 relations do not occur in their dataset and 7 more occur in less than 1% of the data. Nastase and Szpakowicz' scheme mitigates

[1] The interpretation of prepositions is itself the focus of a Semeval task in 2007.

this problem by the presence of 5 supercategories.

Each of these proposals has its own advantages and drawbacks, and there is a need for principled criteria for choosing one. As the literature on semantic annotation "best practice" is rather small,[2] I devised a novel set of design principles based on empirical and theoretical considerations:

1. The inventory of informative categories should account for as many compounds as possible
2. The category boundaries should be clear and categories should describe a coherent concept
3. The class distribution should not be overly skewed or sparse
4. The concepts underlying the categories should generalise to other linguistic phenomena
5. The guidelines should make the annotation process as simple as possible
6. The categories should provide useful semantic information

These intuitively appear to be desirable principles for any semantic annotation scheme. The requirement of class distribution balance is motivated by the classification task. Where one category dominates, the most-frequent-class baseline can be difficult to exceed and care must be taken in evaluation to consider macro-averaged performance as well as raw accuracy. It has been suggested that classifiers trained on skewed data may perform poorly on minority classes (Zhang and Oles, 2001). Of course, this is not a justification for conflating concepts with little in common, and it may well be that the natural distribution of data is inherently skewed.

There is clearly a tension between these criteria, and only a best-fit solution is possible. However, it was felt that a new scheme might satisfy them more optimally than existing schemes. Such a proposal necessitates a method of evaluation. Not all the criteria are easily evaluable. It is difficult to prove generalisability and usefulness conclusively, but it can be maximised by building on more general work on semantic representation; for example, the guidelines introduced here use a conception of events and participants compatible with that of FrameNet (Baker et al., 1998). Good results on agreement and baseline classification will provide positive evidence for the coherence and balance of the classes; agreement measures can confirm ease of annotation.

Relation	Distribution	Example
BE	191 (9.55%)	**steel knife**
HAVE	199 (9.95%)	**street name**
IN	308 (15.40%)	**forest hut**
INST	266 (13.30%)	**rice cooker**
ACTOR	236 (11.80%)	**honey bee**
ABOUT	243 (12.15%)	**fairy tale**
REL	81 (4.05%)	**camera gear**
LEX	35 (1.75%)	**home secretary**
UNKNOWN	9 (0.45%)	**simularity crystal**
MISTAG	220 (11.00%)	**blazing fire**
NONCOMP	212 (10.60%)	**[real tennis] club**

Table 1: Sample class frequencies

In choosing an appropriate level of granularity, I wished to avoid positing a large number of detailed but rare categories. Levi's (1978) set of nine semantic relations was used as a starting point. The development process involved a series of revisions over six months, aimed at satisfying the six criteria above and maximising interannotator agreement in annotation trials. The nature of the decisions which had to be made is exemplified by the compound **car factory**, whose standard referent seems to qualify as FOR, CAUSE, FROM and IN in Levi's scheme (and causes similar problems for the other schemes I am aware of). Likewise there seems to be no principled way to choose between a locative or purposive label for **dining room**. Such examples led to both redefinition of category boundaries and changes in the category set; for example, FOR was replaced by INST and AGENT, which are independent of purposivity. This resulted in the class inventory shown in Table 1 and a detailed set of annotation guidelines.[3] The scheme's development is described at length in Ó Séaghdha (2007b).

Many of the labels are self-explanatory. INST(rument) applies to non-sentient participants in an event and AGENT applies to sentient participants, with ties (e.g., **stamp collector**) being broken by a hierarchy of coarse semantic roles. REL is an OTHER-style category for compounds encoding non-specific association. LEX(icalised)

[2] One relevant work is Wilson and Thomas (1997).

[3] The guidelines are publicly available at `http://www.cl.cam.ac.uk/~do242/guidelines.pdf`.

applies to compounds which are semantically opaque without prior knowledge of their meanings. MISTAG and NONCOMP(ound) labels are required to deal with sequences that are not valid two-noun compounds but have been identified as such due to tagging errors and the simple data extraction heuristic described in Section 3.1. Coverage is good, as 92% of valid compounds in the dataset described below were assigned one of the six main semantic relations.

3 Annotation Experiment

3.1 Data

A simple heuristic was used to extract noun sequences from the 90 million word written part of the British National Corpus.[4] The corpus was parsed using the RASP parser (Briscoe et al., 2006) and all sequences of two common nouns were extracted except those adjacent to another noun and those containing non-alphabetic characters. This yielded almost 1.6 million tokens with 430,555 types. 2,000 unique tokens were randomly drawn for use in annotation and classification experiments.

3.2 Method

Two annotators were used: the current author and an annotator experienced in lexicography but without any special knowledge of compounds or any role in the development of the annotation scheme. In all the trials described here, each compound was presented alongside the sentence in which it was found in the BNC. The annotators had to assign one of the labels in Table 1 and the rule that licensed that label in the annotation guidelines. For example, the compound **forest hut** in its usual sense would be annotated `IN,2,2.1.3.1` to indicate the semantic relation, the direction of the relation (it is a *hut in a forest*, not a *forest in a hut*) and that the label is licensed by rule 2.1.3.1 in the guidelines (*N1/N2 is an object spatially located in or near N2/N1*).[5] Two trial batches of 100 compounds were annotated to familiarise the second annotator with the guidelines and to confirm that the guidelines were indeed usable for others. The first trial resulted in agreement of 52% and the second in agreement of 73%. The result of the second trial, corresponding to a Kappa beyond-chance agreement estimate (Cohen, 1960) of $\hat{\kappa} = 0.693$, was very impressive and it was decided to proceed to a larger-scale task. 500 compounds not used in the trial runs were drawn from the 2,000-item set and annotated.

3.3 Results and Analysis

Agreement on the test set was 66.2% with $\hat{\kappa} = 0.62$. This is less than the score achieved in the second trial run, but may be a more accurate estimator of the true population κ due to the larger sample size. On the other hand, the larger dataset may have caused annotator fatigue. Pearson standardised residuals (Haberman, 1973) were calculated to identify the main sources of disagreement.[6] In the context of inter-annotator agreement one expects these residuals to have large positive values on the agreement diagonal and negative values in all other cells. Among the six main relations listed at the top of Table 1, a small positive association was observed between INST and ABOUT, indicating that borderline topics such as **assessment task** and **gas alarm** were likely to be annotated as INST by the first annotator and ABOUT by the second. It seems that the guidelines might need to clarify this category boundary.

It is clear from analysis of the data that the REL, LEX and UNKNOWN categories show very low agreement. They all have low residuals on the agreement diagonal (that for UNKNOWN is negative) and numerous positive entries off it. REL and LEX are also the categories for which it is most difficult to provide clear guidelines. On the other hand, the MISTAG and NONCOMP categories showed good agreement, with slightly higher agreement residuals than the other categories. To get a rough idea of agreement on the six categories used in the classification experiments described below, agreement was calculated for all items which neither annotator annotated with any of REL, LEX, UNKNOWN, MISTAG and NONCOMP. This left 343 items with agreement of 73.6% and $\hat{\kappa} = 0.683$.

[4] `http://www.natcorp.ox.ac.uk/`

[5] The additional information provided by the direction and rule annotations could be used to give a richer classification scheme but has not yet been used in this way in my experiments.

[6] The standardised residual of cell ij is calculated as

$$e_{ij} = \frac{n_{ij} - \hat{p}_{i+}\hat{p}_{+j}}{\sqrt{\hat{p}_{i+}\hat{p}_{+j}(1-\hat{p}_{i+})(1-\hat{p}_{+j})}}$$

where n_{ij} is the observed value of cell ij and $\hat{p}_{i+}$, $\hat{p}_{+j}$ are row and column marginal probabilities estimated from the data.

3.4 Discussion

This is the first work I am aware of where compounds were annotated in their sentential context. This aspect is significant, as compound meaning is often context dependent (compare *school management decided...* and *principles of school management*) and in-context interpretation is closer to the dynamic of real-world language use. Context can both help and hinder agreement, and it is not clear whether in- or out-of-context annotation is easier.

Previous work has given out-of-context agreement figures for corpus data. Kim and Baldwin (2005) report an experiment using 2,169 compounds taken from newspaper text and the categories of Nastase and Szpakowicz (2003). Their annotators could assign multiple labels in case of doubt and were judged to agree on an item if their annotations had any label in common. This less stringent measure yielded agreement of 52.31%. Girju et al. (2005) report agreement for annotation using both Lauer's 8 prepositional labels ($\hat{\kappa} = 0.8$) and their own 35 semantic relations ($\hat{\kappa} = 0.58$). These figures are difficult to interpret as annotators were again allowed assign multiple labels (for the prepositions this occurred in "almost all" cases) and the multiply-labelled items were excluded from the calculation of Kappa. This entails discarding the items which are hardest to classify and thus most likely to cause disagreement.

Girju (2006) has recently published impressive agreement results on a related task. This involved annotating 2,200 compounds extracted from an online dictionary, each presented in five languages, and resulted in a Kappa score of 0.67. This task may have been facilitated by the data source and its multilingual nature. It seems plausible that dictionary entries are more likely to refer to familiar concepts than compounds extracted from a balanced corpus, which are frequently context-dependent coinages or rare specialist terms. Furthermore, the translations of compounds in Romance languages often provide information that disambiguates the compound meaning (this aspect was the main motivation for the work) and translations from a dictionary are likely to correspond to an item's most frequent meaning. A qualitative analysis of the experiment described above suggests that about 30% of the disagreements can confidently be attributed to disagreement about the semantics of a given compound (as opposed to how a given meaning should be annotated).[7]

4 SVM Learning with Co-occurrence Data

4.1 Method

The data used for classification was taken from the 2,000 items used for the annotation experiment, annotated by a single annotator. Due to time constraints, this annotation was done before the second annotator had been used and was not changed afterwards. All compounds annotated as BE, HAVE, IN, INST, AGENT and ABOUT were used, giving a dataset of 1,443 items. All experiments were run using Support Vector Machine classifiers implemented in LIBSVM.[8] Performance was measured via 5-fold cross-validation. Best performance was achieved with a linear kernel and one-against-all classification.[9] The single parameter C was estimated for each fold by cross-validating on the training set. Due to the efficiency of the linear kernel the optimisation, training and testing steps for each fold could be performed in under an hour.

I investigated what level of performance could be achieved using only corpus information. Feature vectors were extracted from the written BNC for each modifier and head in the dataset under the following conditions:

w5, w10: Each word within a window of 5 or 10 words on either side of the item is a feature.

Rbasic, Rmod, Rverb, Rconj: These feature sets use the grammatical relation output of the RASP parser run over the written BNC. The **Rbasic** feature set conflates information about 25 grammatical relations; **Rmod** counts only prepositional, nominal and adjectival noun modification; **Rverb** counts only relations

[7] For example, one annotator thought **peat boy** referred to a *boy who sells peat* (AGENT) while the other thought it referred to a *boy buried in peat* (IN).

[8] `http://www.csie.ntu.edu.tw/~cjlin/libsvm`

[9] Keerthi and Lin (2003) prove that in the scenario of binary classification, the Gaussian kernel will never perform worse than the linear kernel given adequate parameter optimisation. However, Gaussian kernel classifiers consistently performed worse in my experiments; this seems to result from the standard methodology of using a single set of parameters for all of the binary classifiers used in multiclass classification.

among subjects, objects and verbs; **Rconj** counts only conjunctions of nouns. In each case, each word entering into one of the target relations with the item is a feature and only the target relations contribute to the feature values.

Each feature vector counts the target word's co-occurrences with the 10,000 words that most frequently appear in the context of interest over the entire corpus. Each compound in the dataset is represented by the concatenation of the feature vectors for its head and modifier. To model aspects of co-occurrence association that might be obscured by raw frequency, the log-likelihood ratio G^2 was used to transform the feature space.[10]

4.2 Results and Analysis

Results for these feature sets are given in Table 2. The simple word-counting conditions **w5** and **w10** perform relatively well, but the highest accuracy is achieved by **Rconj**. The general effect of the log-likelihood transformation cannot be stated categorically, as it causes some conditions to improve and others to worsen, but the G^2-transformed **Rconj** features give the best results of all with 54.95% accuracy (53.42% macro-averaged). Analysis of performance across categories shows that in all cases accuracy is lower (usually below 30%) on the BE and HAVE relations than on the others (often above 50%). These two relations are least common in the dataset, which is why the macro-averaged figures are slightly lower than the micro-averaged accuracy.

4.3 Discussion

It is interesting that the conjunction-based features give the best performance, as these features are also the most sparse. This may be explained by the fact that words appearing in conjunctions are often taxonomically similar (Roark and Charniak, 1998) and that taxonomic information is particularly useful for compound interpretation, as evidenced by the success of WordNet-based methods (see Section 5).

Comparing reported classification results is more problematic than comparing annotation results, as it is impossible to disentangle the effects of different data, annotation schemes and classification methods. The results described here should be taken to demonstrate the feasibility of learning using a well-motivated annotation scheme and to provide a baseline for future work on the same data. In terms of methodology, Turney's (2006) Vector Space Model experiments are most similar. Using feature vectors derived from lexical patterns and frequencies returned by a Web search engine, a nearest-neighbour classifier achieved 45.7% accuracy on compounds annotated with 5 semantic classes. Turney improves accuracy to 58% with a combination of query expansion and linear dimensionality reduction. This method trades off efficiency for accuracy, requiring many times more resources in terms of time, storage and corpus size than that described here. Lapata and Keller (2004) obtain accuracy of 55.71% on Lauer's (1995) prepositionally annotated data using simple search engine queries. Their method has the advantage of not requiring supervision, but it cannot be used with deep semantic relations.

	Raw		G^2	
	Accuracy	Macro	Accuracy	Macro
w5	52.60%	51.07%	51.35%	49.93%
w10	51.84%	50.32%	50.10%	48.60%
Rbasic	51.28%	49.92%	51.83%	50.26%
Rmod	51.35%	50.06%	48.51%	47.03%
Rverb	48.79%	47.13%	48.58%	47.07%
Rconj	**54.12%**	**52.44%**	**54.95%**	**53.42%**

Table 2: Performance of BNC co-occurrence data

5 SVM Classification with WordNet

5.1 Method

The experiments reported in this section make a basic use of the WordNet[11] hierarchy. Binary feature vectors are used whereby a vector entry is 1 if the item belongs to or is a hyponym of the synset corresponding to that feature, and 0 otherwise. Each compound is represented by the concatenation of two such vectors, for the head and modifier. The same classification method was used as in Section 4.

5.2 Results and Discussion

This method achieves accuracy of 56.76% and macro-averaged accuracy of 54.59%, slightly higher

[10] This measure is relatively robust where frequency counts are low and consistently outperformed other association measures in the empirical evaluation of Evert (2004).

[11] `http://wordnet.princeton.edu/`

than that achieved by the co-occurrence features. Combining WordNet and co-occurrence vectors by simply concatenating the G^2-transformed **Rconj** vector and WordNet feature vector for each compound gives a further boost to 58.35% accuracy (56.70% macro-averaged).

These results are higher than those reported for similar approaches on open-text data (Kim and Baldwin, 2005; Girju et al., 2005), though the same caveat applies about comparison. The best results (over 70%) reported so far for compound interpretation use a combination of multiple lexical resources and detailed additional annotation (Girju et al., 2005; Girju, 2006).

6 Conclusion and Future Directions

The annotation scheme described above has been tested on a rigorous multiple-annotator task and achieved superior agreement to comparable results in the literature. Further refinement should be possible but would most likely yield diminishing returns. In the classification experiments, my goal was to see what level of performance could be gained by using straightforward techniques so as to provide a meaningful baseline for future research. Good results were achieved with methods that rely neither on massive corpora or broad-coverage lexical resources, though slightly better performance was achieved using WordNet. An advantage of resource-poor methods is that they can be used for the many languages where compounding is common but such resources are limited.

The learning approach described here only captures the lexical semantics of the individual consituents. It seems intuitive that other kinds of corpus information would be useful; in particular, contexts in which the head and modifier of a compound both occur may make explicit the relations that typically hold between their referents. Kernel methods for using such relational information are investigated in Ó Séaghdha (2007a) with promising results, and I am continuing my research in this area.

References

Collin Baker, Charles Fillmore, and John Lowe. 1998. The Berkeley FrameNet project. In *Proc. ACL-98*.

Ted Briscoe, John Carroll, and Rebecca Watson. 2006. The second release of the RASP system. In *Proc. of the ACL-06 Interactive Presentation Sessions*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Stefan Evert. 2004. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, Universität Stuttgart.

Roxana Girju, Dan Moldovan, Marta Tatu, and Daniel Antohe. 2005. On the semantics of noun compounds. *Computer Speech and Language*, 19(4):479–496.

Roxana Girju. 2006. Out-of-context noun phrase semantic interpretation with cross-linguistic evidence. In *Proc. CIKM-06*.

Shelby J. Haberman. 1973. The analysis of residuals in cross-classified tables. *Biometrics*, 29(1):205–220.

S. Sathiya Keerthi and Chih-Jen Lin. 2003. Asymptotic behaviors of support vector machines with Gaussian kernel. *Neural Computation*, 15:1667–1689.

Su Nam Kim and Timothy Baldwin. 2005. Automatic interpretation of noun compounds using WordNet similarity. In *Proc. IJCNLP-05*.

Mirella Lapata and Frank Keller. 2004. The Web as a baseline: Evaluating the performance of unsupervised Web-based models for a range of NLP tasks. In *Proc. HLT-NAACL-04*.

Mark Lauer. 1995. *Designing Statistical Language Learners: Experiments on Compound Nouns*. Ph.D. thesis, Macquarie University.

Judith N. Levi. 1978. *The Syntax and Semantics of Complex Nominals*. Academic Press, New York.

Vivi Nastase and Stan Szpakowicz. 2003. Exploring noun-modifier semantic relations. In *Proc. IWCS-5*.

Brian Roark and Eugene Charniak. 1998. Noun-phrase co-occurrence statistics for semi-automatic semantic lexicon construction. In *Proc. ACL-COLING-98*.

Diarmuid Ó Séaghdha. 2007a. Co-occurrence contexts for corpus-based noun compound interpretation. In *Proc. of the ACL Workshop A Broader Perspective on Multiword Expressions*.

Diarmuid Ó Séaghdha. 2007b. Designing and evaluating a semantic annotation scheme for compound nouns. In *Proc. Corpus Linguistics 2007*.

Peter D. Turney. 2006. Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416.

Andrew Wilson and Jenny Thomas. 1997. Semantic annotation. In R. Garside, G. Leech, and A. McEnery, editors, *Corpus Annotation*. Longman, London.

Tong Zhang and Frank J. Oles. 2001. Text categorization based on regularized linear classification methods. *Information Retrieval*, 4(1):5–31.