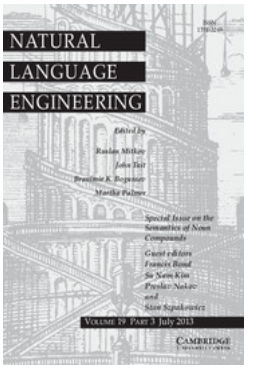
General and specific paraphrases of semantic relations between nouns



General and specific paraphrases of semantic relations between nouns

PAUL NULTY and FINTAN COSTELLO

Natural Language Engineering / Volume 19 / Special Issue 03 / July 2013, pp 357 - 384
DOI: 10.1017/S1351324913000089, Published online: 20 May 2013

Natural Language Engineering **19** (3): 357–384. © Cambridge University Press 2013
doi:10.1017/S1351324913000089

General and specific paraphrases of semantic relations between nouns

PAUL NULTY[1] and FINTAN COSTELLO[2]
[1]*Department of Methodology, London School of Economics, London, UK*
e-mail: paul.nulty@gmail.com
[2]School of Computer Science and Informatics, University College Dublin, Dublin, Ireland.
e-mail: fintan.costello@ucd.ie

(*Received 1 October 2011; revised 27 November 2012; accepted 21 March 2013; first published online 20 May 2013*)

Abstract

Many English noun pairs suggest an almost limitless array of semantic interpretation. A *fruit bowl* might be described as a *bowl for fruit*, a *bowl that contains fruit*, a *bowl for holding fruit*, or even (perhaps in a modern sculpture class), a *bowl made out of fruit*. These interpretations vary in syntax, semantic denotation, plausibility, and level of semantic detail. For example, a *headache pill* is usually a *pill for preventing headaches*, but might, perhaps in the context of a list of side effects, be a *pill that can cause headaches* (Levi, J. N. 1978. *The Syntax and Semantics of Complex Nominals.* New York: Academic Press.). In addition to lexical ambiguity, both relational ambiguity and relational vagueness make automatic semantic interpretation of these combinations difficult. While humans parse these possibilities with ease, computational systems are only recently gaining the ability to deal with the complexity of lexical expressions of semantic relations. In this paper, we describe techniques for paraphrasing the semantic relations that can hold between nouns in a noun compound, using a semi-supervised probabilistic method to rank candidate paraphrases of semantic relations, and describing a new method for selecting plausible relational paraphrases at arbitrary levels of semantic specification. These methods are motivated by the observation that existing semantic relation classification schemes often exhibit a highly skewed class distribution, and that lexical paraphrases of semantic relations vary widely in semantic precision.

1 Introduction

The term *semantic relation* is used throughout research in theoretical linguistics, cognitive science, and artificial intelligence. These relations underpin type theory in generative grammar, form the skeletons of lexical and semantic ontologies, and are the basis for many successful applications in data mining, information retrieval, and natural language processing. Each research area takes a different approach, ranging from a handful of simple structural relations to a carefully enumerated list of minimally distinct relating expressions.

In the surface realization of natural language, semantic relations are undoubtedly an open class. In addition to using verbs to instantiate relations, we find that

prepositions, prepositional verbs, phrasal verbs, and copular constructions are used in various ways to express a predicating relation between two concepts (Baker 2003). However, to a large extent, both applied and theoretical research in this area has preferred to abstract away from the concrete realisation of semantic relations, and instead define categories of relation.

These relation classes are intended to capture syntactic or semantic similarities in the way that pairs of nouns are associated, or are created on an *ad hoc* basis for a particular application. Many authors who devise such abstract classes of relations note the proliferation of edge cases, and the difficulty in obtaining a high agreement among annotators regarding which relation class a noun pair belongs to (Girju *et al.* 2005; Jackendoff 2010). In addition, the class distribution of these taxonomies is often highly skewed, resulting in a high majority-class baseline for classification tasks.

Representing semantic relations with surface words like verbs (Kim and Baldwin 2006; Nakov and Hearst 2006) and prepositions (Lauer 1995) is a promising approach, allowing fine-grained, versatile interpretations of relations, which are easy to integrate into applications.

In this paper we describe methods for ranking paraphrases of semantic relations between constituent nouns of English noun compounds, using surface lexical expressions. Based on recent research into asymmetrical semantic relation association measures, and distributional methods for detecting semantic inclusion, we show that conditional probability and mutual information measures can be balanced to model the sub-typing of relating expressions and reliably predict plausible paraphrases of semantic relations. This allows for a control over the granularity of the relations returned, meaning that a balance can be struck between semantic precision and recall, which has been shown to be useful in other information retrieval tasks, such as detecting verb inferences (Pantel *et al.* 2007).

1.1 Abstract semantic relation classes

One very common approach to the problem of semantically disambiguating noun compounds is to define a set of semantic relations which capture the interaction between the modifier and the head noun, and then attempt to assign one of these semantic relations to each modifier-noun pair. For example, the phrase *flu virus* could be assigned to the semantic relation class *causal* (the virus causes the flu); the relation for *desert storm* could be *location*. There is no consensus as to which set of semantic relations best captures the differences in meaning of various noun phrases. Work in theoretical linguistics has suggested that noun–noun compounds may be formed by the deletion of a predicate verb or preposition (Levi 1978), or an underlying primitive conceptual function (Jackendoff 2010).

In applied research on semantic relations between the constituent nouns of noun compounds, one of the most widely used datasets has been a set of 600 modifier-noun compounds produced by Nastase and Szpakowicz (2003). These compounds were annotated with a general set of five semantic relations, and also with thirty more specific relations. A different taxonomy of nineteen semantic relations was

used in Kim and Baldwin (2005), with each of the abstract relations being defined by a more concrete paraphrase. Girju *et al.* (2005) use a set of thirty-five predefined relations, twenty-one of which were covered by noun compounds extracted from a corpus of the *Wall Street Journal*. Ó Séaghdha (2007) presents a detailed treatment of procedures for deciding on an annotation scheme, and desirable criteria for the resulting relation classes. The resulting scheme consists of a balanced set of six semantic relation classes, with four further classes capturing unusual cases or phrases that have been incorrectly tagged as noun compounds. Tratz and Hovy (2010) present a very large number (17,509) of compounds, annotated by Mechanical Turk users with a fine-grained set of forty-three semantic relations.

1.2 Paraphrases of semantic relations between nouns

The approach of Lauer (1995) and Nakov and Hearst (2006) to representing semantic relations is notably different to other systems. Rather than inventing *ad hoc* categories of relations, they represent relations directly by using paraphrasing lexical expressions.

Paraphrases of semantic relations may be verbs, prepositions, or prepositional verbs like *found in* and *caused by*. Vanderwende (1994) describes a method for generating verbal paraphrases of noun compounds from dictionary definitions. Lauer (1995) categorized compounds using only prepositions. Nakov and Hearst (2006) use only verbs and prepositional verbs; however, many of the paraphrases in this dataset are effectively just prepositions with a copula, such as *be in*, *be for*, and *be of*.

If these relational paraphrases can be discovered automatically, there are several advantages to this approach over classification into abstract relations. The output of a paraphrasing system is more transparent – the meaning of the relation can be directly represented by a word or a phrase instead of needing to be defined in annotation guidelines. This transparency makes applying such techniques much easier, because systems that produce prepositions or verbs to link pairs of nouns can be easily integrated into a summarisation, translation, or query-rewriting system. Kim and Nakov (2011) use noun compounds annotated with paraphrases to iteratively bootstrap more compounds by querying the Yahoo web search engine, and augment these newly discovered compounds with more relating paraphrases.

In addition, lexical phrases allow an arbitrary, controllable level of granularity in the disambiguation – vague and semantically general phrases are sometimes more natural and reliable disambiguations, but more specific phrases can be used if there is high confidence in the result and such precision is required. Most sets of predefined semantic relations have only one or maybe two levels of granularity. This can often lead to semantically converse relations falling under the same abstract category, for example, a *headache tablet* is a tablet for preventing headaches, while *headache weather* is weather that induces headaches – but both compounds would be assigned the same relation (perhaps *instrumental* or *causal*) in many taxonomies of semantic relations. Paraphrases of compounds using verbs or verb–preposition

combinations can provide as much or as little detail as is required to adequately disambiguate the compound.

1.3 Discovering general and specific paraphrases of semantic relations

In this paper we describe methods for discovering plausible paraphrases of noun compounds. The first experiment describes a system for ranking paraphrases that have already been judged to be acceptable by human annotators. The evaluation measure for this problem, which was a task in SemEval 2010 (Butnariu *et al.* 2010), is the correlation between scores generated for each paraphrase and the frequency with which each of the acceptable paraphrases was produced by the human annotator. In Section 3 we describe an extension of this method that uses seed paraphrases extracted from a corpus to find a list of acceptable paraphrases for a given noun compound without a human annotated list of plausible paraphrases. Finally, in Section 4 we describe a parameterization of our scoring method that allows control over the level of semantic detail preferred by our paraphrase ranking method. We discuss how directional (or asymmetrical) association measures (Weeds and Weir 2005; Kotlerman *et al.* 2010) are important for tasks such as this where the underlying nature of the coverage of relation classes may be hierarchical.

1.4 Motivation and applications

Noun–noun combinations have been the focus of much of the research into semantic relations – two concepts are simply juxtaposed with no obvious predicate, and the hearer must use knowledge about the concepts and the context to deduce the most likely interpretation. This makes the noun compound a perfect test case for theories and methods in the study of semantic relations. The structure is almost endlessly productive, and nearly any pair of English nouns can be juxtaposed to form a plausible combination of concepts (Ó Séaghdha 2008). The prevalence and ambiguity of noun compounds means that disambiguation of these forms is an important component of many natural language processing tasks.

A typical translation application is addressed by Johnston and Busa (1996) – disambiguation of English noun compounds is necessary to select the correct preposition when translating from Italian to English.

General knowledge ontologies are often automatically populated by a text mining system, and such systems will encounter many noun compounds. An application that gathers common-sense knowledge from unstructured text will be greatly enhanced if it can deduce that a *car door* is a part of a car, but a *car space* is a place for parking a car.

As well as discovering semantic relations hidden in noun compounds, many such applications seek to abstract knowledge from surface sentences to a general knowledge representation ontology, which can then be used for question answering (for example, IBM's Watson (Ferrucci *et al.* 2010)), or to automatically improve resources such as Freebase (Bollacker *et al.* 2008).

In the question answering domain, Welty *et al.* (2010) note that logically discrete relations (for the purposes of question answering) are represented by many surface

forms, and that these surface forms exhibit the familiar long-tailed frequency distribution common of lexical patterns. For example, to answer a question about which films an actor has been in, an algorithm might need to parse text from a web corpus to find patterns that instantiate the *ACTOR acted-in FILM* relation. Using training examples from existing knowledge bases, they find that a whole range of surface patterns can indicate this relation; FILM *starring* ACTOR, FILM *with* ACTOR, ACTOR *won an award for* FILM – the central task is to combine the relational meaning of these patterns to decide whether they represent the relation that is included in the question.

Again, in the information extraction domain, discovering semantic relations from unstructured text requires a link between surface forms and abstracted relations. The level of abstraction at which to encode a relation varies between systems, and the correct level probably depends on the application. One system, KNEXT (Schubert 2002), learned to extract common-sense information from a blog corpus, and stores abstract relations with links to the source surface text, resulting in general hypotheses such as *PERSON have-as-part ANKLE* from surface forms like *Bobby Thomson broke his ankle while sliding into second base during a spring training game*. – the possessive surface form *his* being predictive of the abstract relation *have-as-part*.

2 SemEval 2010 task 9 dataset and evaluation method

For the experiments described in this paper, we use the dataset created for SemEval 2010 Task 9: *Noun Compound Interpretation Using Paraphrasing Verbs* (Butnariu *et al*. 2010). This data consists of 638 two-word noun compounds, annotated with paraphrasing expressions by human subjects using Amazon Mechanical Turk. On average, seventy-one participants were recruited for each noun compound. For each noun compound, the user's task was to provide a paraphrase of the noun compound using verbs and prepositions. The noun compounds are drawn from three sources: Levi (1978), Lauer (1995), and Nastase and Szpakowicz (2003). Adjective–noun compounds and compounds containing hyphenated modifiers (e.g. *test-tube baby*) were excluded from the dataset.

The dataset is described further in Butnariu *et al.* (2010), and the materials, instructions, and a discussion of the data-collection process are outlined in Nakov (2008). The complete dataset is freely available for download on the SemEval website under a Creative Commons License.[1]

2.1 Data collection

During the annotation process, Mechanical Turk users were presented with a noun compound, and instructed to complete a paraphrase as follows:

> Given a noun compound '*noun1 noun2*', you are asked to substitute the dots with one or more verbs, optionally followed by a preposition.
> 'a noun1 noun2 is a **noun2 that .... noun1**'

[1] http://semeval2.fbk.eu/semeval2.php?location=data

Table 1. *Portion of training data for lace handkerchief*

```
lace handkerchief be made of 26
lace handkerchief be made from 10
lace handkerchief contain 8
lace handkerchief be 7
...
lace handkerchief come from 2
lace handkerchief include 2
lace handkerchief be adorned with 1
lace handkerchief be attached with 1
```

Each user was asked to try to provide three paraphrases for each noun compound. The dataset uses two-word noun compounds in which the modifier precedes the head, as is normal in English, such as *apple pie* and *malaria mosquito*. On average, seventy-one Mechanical Turk users provided paraphrases for each compound. In total, an average of 79.3 paraphrase types, and 189.1 paraphrase tokens were provided for each compound (Butnariu *et al*. 2010). There are 7,296 unique paraphrase types across the whole dataset.[2]

2.2 Examples from the dataset

The dataset contains, for each noun compound, a list of all the human-proposed paraphrases for that compound. For the SemEval contest, the data was split into training and testing sets. There were 250 compounds in the training set and 388 in the test set. The frequency with which each paraphrase occurred is included in the training data. After the contest was evaluated by the organisers, the frequencies for the test portion of the dataset were released also.

Table 1 shows a portion of the paraphrases collected for a compound in the training data (*lace handkerchief*). The frequency beside a paraphrase represents the number of human participants who provided that paraphrase for the noun compound – for example, of all the users asked to paraphrase *lace handkerchief*, eight provided the paraphrase *handkerchief contains lace*. In total, fifty-five paraphrase types were provided for this compound.

Table 2 shows an example of how the testing data was presented before the SemEval contest was evaluated. The frequencies are not included and the phrases are ordered at random. The objective of the SemEval task was to rank the paraphrases of the compounds in the test data to correlate with their ranking according to the frequency with which they were provided by the human annotators.

Systems taking part in the task provided a score for each paraphrase in each compound of the test set, and the paraphrases were ranked according to these scores. The Spearman rank correlation between this ranking and the human-provided

[2] The task description paper reports 50,562 paraphrasing verb types. It seems that this is based on the sum of the number of unique paraphrases within each compound; however, there are only 7,296 unique relating paraphrases across the whole dataset.

Table 2. *Portion of testing data for ice crystal. The order of the paraphrases is random*

ice crystal appear in
ice crystal form
ice crystal be related to
ice crystal be found in
ice crystal consist of
...
ice crystal be associated with
ice crystal create
ice crystal be formed from
ice crystal be generated from
ice crystal be from

frequency ranking was the evaluation measure for the contest, although the Pearson and cosine correlations were also reported.

2.3 Intuition behind the model

The aim of the task is to rank paraphrases for noun compounds given by fifty to one hundred human annotators. When deciding on a model we took into account several observations about the data.

For this task, the model does not need to produce plausible paraphrases for noun compounds, it simply needs to rank paraphrases that have been provided. Given that all of the paraphrases in the training and test sets have been produced by people, we presume that all of them will have at least some plausible interpretation, and most paraphrases for a given compound will indicate generally the same interpretation of that compound.

This will not always be the case; some compounds seem to be genuinely ambiguous rather than vague. For example, a *newspaper bowl* could *be a bowl for holding newspaper* or a *bowl made of newspaper*. However, the mere fact that a compound has occurred in text is evidence that the speaker who produced the text believed that the compound was unambiguous, at least in the given context.

Given that most of the compounds in the dataset have one clear plausible meaning to readers, when asked to paraphrase a compound people tend to observe the Grician maxim of brevity (Grice 1975) by using simple, frequent terms rather than detailed, semantically weighty paraphrases. For example, for the compound *alligator leather* in the training data, the two most popular relating phrases were *be made from* and *come from*. Also provided as paraphrases for this compound were *hide of* and *be skinned from*. These are more detailed, specific, and more useful than the most popular paraphrases, but they were only produced once each, while *be made from* and *come from* were provided by twenty-eight and twenty annotators respectively. This trend is noticeable in most of the compounds in the training data – the most specific and detailed paraphrases are not the most frequently produced paraphrases. The most frequently produced paraphrases are hypernyms or parent senses – *be made from* is an acceptable paraphrase of more specific subtypes of relation.

2.4 Using conditional probability to detect subtypes

Our model uses conditional probabilities to detect this sub-typing structure based on the theory that observing a specific, detailed paraphrase is good evidence that a more general parent sense of that paraphrase would be acceptable in the same context. The reverse is not true – observing a frequently occurring, semantically light paraphrase is not strong evidence that any sub-sense of that paraphrase would be acceptable in the same context. For example, consider the spatial and temporal sub-senses of the paraphrase *be in*. A possible spatial sub-sense of this paraphrase is *be located in*, while a possible temporal sub-sense would be *occur during*.

The fact that *occur during* is provided as a paraphrase for a compound almost always means that *be in* is also a plausible paraphrase. However, observing *be in* as a paraphrase does not provide such strong evidence for *occur during* also being plausible, as we do not know which sub-sense of *in* is intended.

If this is correct, then we would expect that the conditional probability of a paraphrase r_1 occurring given that we have observed another paraphrase r_2 in the same context is a measure of the extent to which r_1 is a more general type (parent sense) of r_2.

2.5 System description

The first step in the model is to generate a conditional probability table by iterating over all the compounds in the dataset and counting the co-occurrences of each possible paraphrase with all other paraphrases in the dataset. Using the co-occurrences and the prior probabilities (derived from the the overall frequencies) we can compute the conditional probability of every paraphrase with all other paraphrases individually.

We could use either the training or the test set to collect these co-occurrence statistics, as the frequencies with which the paraphrases are ranked are not used – we simply count how many times each paraphrase co-occurred as a possible paraphrase for the same compound with each other paraphrase. For the submitted system we used the test data, but subsequently we confirmed that using only the training data for this step is not detrimental to the system's performance.

For each paraphrase r_1 in the data, the conditional probability of that paraphrase is computed with respect to all other paraphrases in the data. For any two paraphrases r_1 and r_2, the probability of r_1 given r_2 is the probability of their co-occurrence divided by the prior probability of r_2,

$$P(r_1|r_2) = \frac{P(r_1 \cap r_2)}{P(r_2)}$$

Given a compound in the test set, we score each of its candidate paraphrases by summing the conditional probabilities of it occurring with each other paraphrase provided for the same compound,

$$\text{score}(r_1) = \sum_{r_2 \in R} P(r_1|r_2)$$

Table 3. *Conditional probability table for four candidate-relating paraphrases for soup pot. The table shows the probability of the phrase in the row given the phrase in the column; for example P(hold/enclose) is 0.714*

	enclose	contain	hold	be filled with	Score
enclose	–	0.024	0.100	0.000	0.124
contain	0.857	–	0.880	1.000	2.737
hold	0.714	0.177	–	0.619	1.510
be filled with	0.000	0.084	0.260	–	0.344

For a list of paraphrases R provided for a given compound, we score a paraphrase r_1 in that list by summing its conditional probability individually with every other paraphrase in the list. This method of combining individual conditional probabilities to provide a score for a class given a set of observations is similar to the Naive Bayes algorithm commonly applied in machine learning tasks where the number of classes is too great, or the dimensionality of the data is too complex to apply more sophisticated classification methods such as Support Vector Machines.

The Naive Bayes algorithm estimates a posterior probability for each class by multiplying the prior probability of the class, and the probabilities of the class given each observation, assuming that each piece of evidence observed in the feature vector is conditionally independent of the others. This independence condition is clearly not met in our case – paraphrases with similar meanings are highly covariant. Therefore, rather than combining the probabilities by multiplication and claiming to have obtained a true posterior probability for each paraphrase given the other paraphrases in the list, we combine the probabilities by summing them and use this score for the predictions. Another advantage of this is that since the summed scores are not normalized between 0.1 as a true probability would be, they correlate better with human frequencies using the unscaled cosine similarity correlation measure.

This gives the more general, broad coverage, paraphrases a higher score, and also has a clustering effect whereby paraphrases that have not co-occurred with the other paraphrases in the list very often for other compounds are given a lower score – they are unusual in the context of this paraphrase list.

The system is implemented in a Python script, using a dictionary to store $P(r_1|r_2)$ for all combinations of all paraphrases in the data.[3] The algorithm is computationally intensive, as for each paraphrase in each compound, the co-occurrence scores with each other paraphrase must be updated. Still, as there are only 7,296 unique paraphrases across the dataset, the probabilities can be calculated in a few minutes on an ordinary desktop machine.

2.6 Example

Table 3 shows a worked example of the scoring method for four paraphrases of the compound *soup pot*. The table shows the conditional probability of each phrase

[3] The system implementation is available online at https://github.com/pnulty/semeval9

Table 4. *Results for systems participating in SemEval 2010 Task 9. The baseline scores of each paraphrase according to its overall frequency (prior probability)*

System	Spearman	Pearson	Cosine
UVT	**0.450**	**0.411**	0.635
UCD-PN	0.441	0.361	**0.669**
UCD-GOG-III	0.432	0.395	0.652
UCD-GOG-II	0.418	0.375	0.660
UCD-GOG-I	0.380	0.252	0.659
UCAM	0.267	0.219	0.374
NC-INTERP	0.186	0.070	0.466
Baseline	0.425	0.344	0.524

along the left given the phrase along the top. For example, the probability of *be filled with* given *contain* is 0.084.

This table illustrates the asymmetrical nature of the conditional probability association measure. *be filled with* is a low-frequency, semantically precise relating phrase. It occurs in twenty-one of the 638 compounds in the dataset. *contain* is a high-frequency relatively semantically general paraphrase, occurring in 248 of the 638 compounds. The probability of *contain* given *be filled with* is 1, i.e. every time *be filled with* is an appropriate paraphrase of a compound, *contain* was also provided as an appropriate paraphrase. However, of all the 248 times *contain* occurred, it only co-occurred with *be filled with* twenty-one times.

The intuition that more general terms share more features (or more contexts) than more specific terms (sometimes called distributional inclusion; Geffet and Dagan 2004) has been shown to be predictive of a hyponomy relation for nouns – the effectiveness of conditional probability as an directional (asymmetric) measure of association between relating paraphrases suggests that distributional inclusion is also useful for modelling sub-typing relations between verbs, prepositions, and phrasal verbs.

2.7 Task results

Table 4 shows the performance of all seven participating systems on the task. Our system achieved the second highest correlation according to the official evaluation measure, Spearman's rank correlation coefficient. Results were also provided using Pearson's correlation coefficient and the cosine of the vector of scores for the gold standard and submitted predictions. Our system performed best using the cosine measure, which measures how closely the predicted scores match the gold standard frequencies, rather than the rank correlation. This could be important for tasks which require a scalar measure of acceptability rather than pairwise competition between paraphrases.

The baseline predictions were made by summing the overall frequency for each paraphrase in the training set, and scoring the paraphrases for each compound in

the test set by this frequency. This simple method is similar to a majority class back-off classifier in machine learning tasks or the most-frequent-sense baseline for word sense disambiguation – the unchanged prior probability for each class in the training set is the prediction for every item in the test set.

We collected the co-occurrence statistics for our submitted prediction from the test set of paraphrases alone. Since our model does not use the frequencies provided in the training set, we chose to use the test set as it was larger and had more annotators. This could be perceived as an unfair use of the test data, as we are using all of the test compounds and their paraphrases to calculate the position of a given paraphrase relative to other paraphrases. This is a kind of clustering which would not be possible if only a few test cases were provided. To check that our system did not need to collect co-occurrence probabilities on exactly the same data as it made predictions on, we submitted the second set of predictions for the test based on the probabilities from the training compounds alone. These predictions actually achieved a slightly better score for the official evaluation measure, with a Spearman rho of 0.444, and a cosine of 0.631. This suggests that the model does not need to collect co-occurrence statistics from the same compounds as it makes predictions on as long as sufficient data is available.

2.8 Analysis

The system which achieved the highest Spearman correlation with human scores (Wubben 2010) uses a supervised machine learning method (memory-based learning; Daelemans *et al.* 1999) combining features from an external corpus (Google's Web 1T n-gram dataset; Brants and Franz 2006), WordNet ancestors, and relative frequency in the training data. The UCD-Goggle system (Li, Lopez-Fernandez and Veale 2010) also makes use of the Web 1T dataset, but is unsupervised with respect to the SemEval training data. No other system outperformed the baseline.

The strength of the baseline again indicates the high coverage of generally frequent paraphrases across compounds in the training and test set. Our system takes advantage of this because the score is not normalized to reduce the effect of frequent paraphrases – the co-occurrences are only divided by the prior probability of the observed paraphrase, not a combination of the observed and the target phrases, as is the case with man word similarity measures. This asymmetric property of the equation will be discussed further in the next section.

The most significant drawback of this system is that it cannot discover paraphrases for noun compounds – it is designed to rank paraphrases that have already been provided. Using the conditional probability to rank paraphrases has two effects. First, there is a clustering effect which favours paraphrases that are more similar to the other paraphrases in a list for a given compound. Second, paraphrases which are more frequent overall receive a higher score, as frequent verbs and prepositions may co-occur with a wide variety of more specific terms.

These effects lead to two possible drawbacks. First, the system would not perform well if detailed, specific paraphrases of compounds were needed. Although less

frequent, more specific paraphrases may be more useful for some applications, these are not the kind of paraphrases that people seem to produce spontaneously.

Second, because of the clustering effect, this system does not work well for compounds that are genuinely ambiguous, e.g. *stone bowl* (*bowl made of stone* or *bowl contains stones*). Most examples are not this ambiguous, and therefore almost all of the provided paraphrases for a given compound are plausible, and indicate the same relation. They vary mainly in how specific/detailed their explanation of the relation is.

The three compounds which our system produced the worst rank correlation for were *diesel engine*, *midnight train*, and *bathing suit*. Examining the list of possible paraphrases for the first two of these suggests that the annotators identified two distinct senses for each: *diesel engine* is paraphrased by verbs of containment (e.g. *be in*) and verbs of function (e.g. *runs on*), while midnight train is paraphrased by verbs of location (e.g. *be found in*, *be located in*) and verbs of movement (e.g. *run in*, *arrive at*).

Our model works by separating paraphrases according to granularity and cannot disambiguate these distinct senses. The list of possible paraphrases for *bathing suit* suggests that our model is not robust if implausible paraphrases are in the candidate list – the model ranked *be in*, *be found in* and *emerge from* among the top eight paraphrases for this compound, even though they are barely comprehensible as plausible paraphrases.

The difficulty here is that even if only one annotator suggests a paraphrase, it is deemed to have co-occurred with other paraphrases in that list, since we do not use the frequencies from the training set. In the next section we will describe the use of a threshold to adjust the reliability of co-occurrences using a minimum frequency score to exclude paraphrases provided by only a small number of annotators.

The compounds for which the highest correlations were achieved were *wilderness areas*, *consonant systems*, and *fiber optics*. The candidate paraphrases for the first two of these seem to be fairly homogeneous in semantic intent. *Fiber optics* is probably a lexicalised compound which hardly needs paraphrasing. This would lead people to use short and semantically general paraphrases, since no further semantic information is needed to understand a lexicalised form.

3 Discovering relational paraphrases of noun compounds

In the previous section we showed that ordering paraphrases of semantic relations using a simple conditional probability maximization method is effective at reproducing the order of frequency with which such paraphrases were produced by human volunteers.

This method is only useful in situations where a list of possible paraphrases is available, but the frequency with which each is produced is not available. For practical applications, it is more likely that a list of acceptable paraphrases is not available – the task then is to automatically provide an appropriate relating paraphrase without a list of suitable candidates to rank.

For automatic relation extraction, a common approach (Pantel and Pennacchiotti 2006; Turney 2006; Banko *et al.* 2007) has been to use a large corpus to extract relations by finding sentences in the corpus where the relation between the two words is explicit. For example, Nakov (2008) obtains a medium correlation between web-extracted and human-provided lists of noun compound paraphrases. Even with very large corpora, however, coverage can still be a problem – the number of sentences containing a useful explicit lexical relation between two specific nouns is low. Nakov (2008) notes that no paraphrasing verbs could be extracted from the web for fourteen of the 250 noun compounds taken from Levi (1978).

In this section we apply the sum of conditional probability method described in the previous section as a general association measure between relating phrases, using this measure to rank all possible relating paraphrases provided by annotators, rather than just ranking the within-compound paraphrases. We test this by using a small number of seed paraphrases, either extracted from a corpus or drawn from the gold-standard examples, and using these seed paraphrases to predict a longer list of plausible relating paraphrases for each noun compound.

In some respects, we are using the conditional probability as a phrase similarity metric, but rather than finding the most semantically similar relating paraphrases, we want to find paraphrases that are reliably acceptable when substituted for the seed paraphrase. The distinction between semantic similarity and acceptable 'substitutability' is highlighted in Weeds and Weir (2005) and Kotlerman *et al.* (2010), and is relevant to any task where acceptability in context is used for evaluation, especially if the frequency distribution of types are very uneven, as is often the case in word sense disambiguation tasks.

We use the same algorithm for scoring each paraphrase as in the previous section, but, rather than ranking a small list of (around seventy) paraphrases that have been specifically chosen for a given compound, the algorithm is applied to the large (7,296) list of all paraphrase types that have been provided for any compound in the dataset.

We evaluate the system by counting what fraction of the top m paraphrases ranked by this scoring method (excluding the chosen seed paraphrases) has been provided by human annotators. This accuracy is evaluated by comparison with a random choice baseline, and also with a stronger baseline which always predicts the most frequent paraphrase. In order to evaluate the system at different levels of coverage and precision, we experiment with different values of several thresholds relating to the number of paraphrases predicted and the reliability of the annotators' judgements.

3.1 Paraphrase distribution and thresholds

In order to judge the effectiveness of a system that produces paraphrases of the noun compounds in the dataset, it is necessary to decide which of the human-provided paraphrases are 'acceptable' or 'correct.' It would perhaps seem natural that any paraphrase produced by a human annotator should be considered an acceptable paraphrase. However, because many annotators were used per compound

Table 5. *The proportion of noun compounds for which each phrase has been provided as an acceptable paraphrase by at least one (left) or two (right) annotators. For example, the paraphrase* 'come from' *has been provided by at least one annotator for 91 percent of noun compounds in the dataset*

$n = 1$		$n = 2$	
come from	0.91	come from	0.76
be related to	0.91	be in	0.75
be in	0.89	be related to	0.74
be of	0.82	be found in	0.68
be found in	0.81	be of	0.61
emerge from	0.80	deal with	0.57
deal with	0.78	involve	0.54
involve	0.77	be for	0.53
be for	0.77	emerge from	0.53
be associated with	0.74	be associated with	0.45
be located in	0.66	contain	0.39
relate to	0.60	have	0.34
contain	0.57	relate to	0.34
be from	0.55	include	0.33
use	0.54	use	0.32
be concerned with	0.53	consist of	0.27
have	0.51	make	0.25
include	0.51	be located in	0.24
be connected to	0.49	be used for	0.24
make	0.46	be concerned with	0.24

(on average 79.2), each compound has a large number of paraphrases that have been provided as acceptable by at least one annotator.

If the threshold for acceptability of a paraphrase is that at least one annotator has provided the paraphrase, then the most frequent paraphrases overall have a very high coverage. For example, the most frequent paraphrase, *come from*, is provided by at least one annotator for 582 of the 638 paraphrases (91 percent). Therefore, a system that simply provides *come from* as a paraphrase for each compound will achieve an accuracy of 91 percent. With almost eighty people producing interpretations for each compound on average, and given that Amazon Mechanical Turk workers are not always motivated to produce gold-standard data, it seems appropriate to place the threshold for judging acceptability higher than one annotator.

Tables 5 and 6 show the top twenty paraphrases and their coverage at different values of n, where n is the number of annotators who are required to have provided the paraphrase in order for it to be judged as valid.

With $n = 5$, the top paraphrase no longer covers more than half of the compounds.

Another threshold that can be adjusted is the number of paraphrases that the system predicts for each compound. Since the system ranks the entire list of all paraphrases according to their score for a given noun compound, we can choose the top m paraphrases as the system's predictions.

Table 6. *The proportion of noun compounds for which each phrase has been provided as an acceptable paraphrase by at least three (left) or five (right) annotators. For example, the paraphrase* 'come from' *has been provided by at least one annotator for 41 percent of noun compounds in the dataset*

$n = 3$		$n = 5$	
come from	0.62	come from	0.41
be in	0.61	be in	0.36
be related to	0.59	be found in	0.33
be found in	0.56	be related to	0.30
deal with	0.41	contain	0.24
be for	0.40	involve	0.23
involve	0.40	be for	0.23
be of	0.39	deal with	0.22
contain	0.33	be of	0.18
emerge from	0.31	have	0.14
be associated with	0.25	use	0.14
have	0.24	be made of	0.14
use	0.24	consist of	0.13
include	0.23	include	0.12
consist of	0.21	be used for	0.12
be made of	0.18	emerge from	0.31
relate to	0.17	be associated with	0.25
be	0.17	be	0.10
be used for	0.17	emerge from	0.09
be about	0.16	be made from	0.09

Therefore, when evaluating the system, the accuracy is computed by counting how many of the system's top m predictions (excluding the corpus-derived seeds) have been provided by at least n annotators in the gold standard data. The baseline ranks all paraphrases by their overall frequency, and counts how many of the top m most frequent paraphrases (again, excluding the seed paraphrases) were provided by at least n annotators for a given compound. The overall accuracy is the mean of the accuracy over all 638 compounds.

As raised in the previous section, building the conditional probability table on the same set of noun compounds as the system is tested on could be regarded as an unfair advantage, since, for each compound during testing, the co-occurrence statistics include the co-occurrences for the noun compound being evaluated. To ensure that this was not a factor, cross-validation was used when evaluating the accuracy of the system in this section.

The dataset is split into k folds and the training and testing process is repeated k times. Each time, the fold used for testing and calculating accuracy is excluded during the training process, while the rest of the dataset is used to build the conditional probability table. This ensures that none of the items used to train the algorithm are also used to evaluate it. The evaluation measure, accuracy, is calculated for each

iteration, and the final accuracy value is the average of the k runs. The SemEval 9 dataset contains 638 noun compounds. The results reported here use a value of $k = 22$, which means that each accuracy and baseline reported is the result of the average of twenty-two runs of the algorithm; during each run the conditional probability table is built using co-occurrence statistics from 589 noun compounds, and the accuracy is tested using the remaining thirty-nine compounds.

This value of k chosen as a compromise between having a high number of compounds to estimate the conditional probability, while needing only twenty-two runs to estimate accuracy, rather than 687 runs that would be necessary if leave-one-out cross-validation was used.

3.2 Paraphrase scoring

Using the conditional probability association measure, we score each paraphrase in the entire dataset as a candidate paraphrase for a particular noun compound, given a small set of seed paraphrases. In the SemEval task, the seed paraphrases were not necessary because a list of correct paraphrases was already available, annotated for each compound – the scoring method used this list to predict the frequency with which each paraphrase had been annotated for the compound.

The score for a given paraphrase for a particular compound is computed by summing the conditional probability of it occurring with each of the small sets of seed paraphrases. The task is to use the seed paraphrases to predict other paraphrases in the list of acceptable paraphrases that the annotators provided.

Where S is a list of seed paraphrases, and r is a paraphrase from the large list of all possible paraphrases,

$$\text{score}(r) = \sum_{s \in S} P(r|s)$$

The list of all paraphrases is then sorted by score. This scoring system favours paraphrases that tend to occur in the same context (i.e. as paraphrases for the same compound) as the seed paraphrases.

The problem of choosing paraphrases from a list of all unique relating paraphrases (7,296 types) is a challenging classification problem. The highly skewed distribution, and the very large number of classes make the simple independent probability method effective because once the initial conditional probabilities are computed, classification is very computationally efficient.

3.3 Evaluation

We first extracted seed paraphrases from the Google Web 1T n-gram dataset (Brants and Franz 2006) by finding n-grams that began with the head word from the noun compounds and ended with the modifying noun. This corpus consists of n-grams collected from web data, and is available to researchers in its entirety, rather than through a web search interface. This means that there is no limit to the amount of searches that may be performed, and an arbitrarily complex query syntax is

possible. Hawker (2006) provides an example of using the corpus for word sense documentation, and describes a method for efficient searching.

The Web 1T corpus consists of n-grams taken from approximately one trillion words of English text taken from web pages in Google's index of web pages. The data includes all 2, 3, 4, and 5-grams that occur more than forty times in these pages. The data comes in the form of approximately 110 compressed files for each of the window sizes. Each of these files consists of exactly 10 million n-grams, with their frequency counts. Below is an example of the 3-gram data:

```
ceramics collection and 43
ceramics collection at 52
ceramics collection is 68
ceramics collection | 59
ceramics collections , 66
ceramics collections . 60
```

To reduce noise in the data, we excluded n-grams that contained any punctuation or non-alphanumeric characters. Also excluded were n-grams that contained any upper case letters, except for the case where the first letter of the string is capitalized.

The data was indexed using a custom python script that created an inverted index based on both first word and last word of the n-gram. Only n-grams with a frequency of 40 or higher are included in the dataset, which means that an average query returns fewer results than a web search.

We retrieved relating paraphrases by searching the corpus for all morphological variations of the component nouns of the noun compound, and extracting strings of verbs and prepositions that occurred between the constituent nouns. These strings were then lemmatized and string-matched to lemmatized paraphrases in the human-generated dataset.

If no paraphrases were found in the corpus for each compound, the system backs off and uses the most frequent overall paraphrases. If seed paraphrases are found, the system uses the sum of conditional probabilities method to return new relating phrases from the large list of all human-generated paraphrases. An accuracy score is generated by counting how many of the new paraphrases predicted by this method are among those that were provided by the human annotators in the SemEval data. These results are shown in Table 7.

The baseline is obtained by always predicting the *m* paraphrases from the large paraphrase list that were provided most frequently by the human annotators, excluding the seed paraphrases. A similar baseline is commonly used to evaluate word sense disambiguation algorithms, the most-frequent-sense baseline. In the word sense disambiguation domain, this baseline is very strong, with only recent systems outperforming it by more than 5 percent accuracy.

We also evaluated the model using seed paraphrases randomly sampled from the gold-standard data. The seeds were again excluded from the accuracy calculation. These results are shown in Table 8. As might be expected, the system performs better when the human-produced seed paraphrases are used.

Table 7. *Accuracy using conditional probability and seed paraphrases retrieved from the Web 1T corpus. For each compound, all paraphrases retrieved from the corpus for that compound were used as seed paraphrases. If no seed paraphrases were found, the algorithm backs off to the most frequent overall paraphrases (the same method is used by the baseline)*

Acceptability threshold	Number of predictions	Baseline	Accuracy
3	1	0.602	0.687
3	3	0.611	0.626
3	5	0.546	0.569
5	1	0.410	0.525
5	3	0.400	0.506

Table 8. *Accuracy using conditional probability and randomly chosen seed paraphrases.* 'n' *is the threshold frequency required for a paraphrase in the human-annotated set to be judged correctly,* 'm' *is the number of new paraphrases predicted by the algorithm, and NumSeeds is the number of randomly chosen seed paraphrases used to make the prediction. A random guessing algorithm achieves an accuracy of 0.001 under all conditions*

n	*m*	NumSeeds	Baseline	Accuracy
3	1	1	0.534	0.699
3	1	2	0.521	0.738
3	1	3	0.519	0.748
n	*m*	NumSeeds	Baseline	Accuracy
3	2	3	0.560	0.689
3	5	3	0.482	0.599
3	9	3	0.417	0.503
n	*m*	NumSeeds	Baseline	Accuracy
3	3	3	0.557	0.674
5	3	3	0.259	0.456

4 An association measure for general and specific relating phrases

In the above sections, we have described the scoring algorithm used to rank noun compound paraphrases as a sum of conditional probabilities. The conditional probabilities represent a particular measure of association between candidate paraphrases – given that we have observed one phrase, the conditional probability measures the probability that the second phrase is acceptable in for the same noun compound. Using co-occurrences to estimate associations between words is one technique often used to judge the semantic association of words in keeping with the idea that a word's meaning can be inferred from its distribution across contexts.

When calculating associations between words, the alternative to distributional similarity is to use similarity based on a hand-built hierarchy or taxonomy that links related terms according to a couple of basic semantic relations (usually hyponymy – the 'is a type of' relation). Although comprehensive taxonomies exist for nouns and verbs, there is no hand-built taxonomy of relating phrases, such as those that are the focus of these experiments, that include prepositions, verbs, and phrasal verbs. Distributional similarity methods have been shown to be effective when using lexical similarity of the constituent nouns to disambiguate noun compounds (Nulty and Costello 2010).

One advantage of using conditional probability to indicate similarity is that it is not a symmetrical relation. This is important when judging similarity between phrases that have very different frequencies. Weeds, Weir and McCarthy (2004) present an overview of distributional similarity measures focusing on the differences in characteristics between algorithms depending on the relative frequency of the words they are comparing. Some similarity measures tend to return words in a similar frequency band to the query word, while others tend to return high-frequency words regardless of the frequency of the query word.

The sum of conditional probabilities method used in this section tends to return high-frequency paraphrases regardless of the frequency of the seed paraphrases selected. The reason that this method is successful is that in order to be judged as correct, a paraphrase need only be acceptable when substituted between the two nouns which make up the compound. This substitution test is not normally among the criteria used to evaluate similarity measures, rather it is an evaluation technique more commonly used to test word sense disambiguation algorithms. The confusion probability metric (Essen and Steinbiss 1992) also tends to give a high score to words which will be acceptable when substituted into a wide range of contexts.

We might expect that words which are distributionally general will tend to be also semantically general. This is a useful property for many applications. Often, for example for translation, text summarisation, or natural language generation, it is more important that a paraphrase is acceptable and makes sense when read in context than that it retains its full semantic weight.

However, for some applications, we might want to find terms that are both acceptable when substituted for the target term and also semantically specific enough to convey the precise meaning encoded in the original phrase. This is more in keeping with how word similarity is traditionally perceived, especially for judgements of noun similarity. For example, the noun *bus* may be judged to be more similar to *truck* than to a more general term like *vehicle*, even though *vehicle* is acceptable in a wider range of contexts. Weeds *et al.* (2004) show that relative frequency of nouns can be used to predict a hypernymy relationship with some success, finding correlations between semantic generality, distributional generality, and relative frequency. We examine these correlations further in the context of the lexical relating phrases used to paraphrase noun compounds.

To see how this applies to relating paraphrases, consider the noun phrase *apple cake*. If a paraphrase for this phrase is required, a human annotator, or a corpus retrieval system, might return the phrase *cake baked with apples*. If we want to

find other acceptable relating paraphrases, the most likely phrases to be acceptable are those that are general, for example, *cake of apples* or *cake with apples*, partly because *of* and *with* are acceptable in a wide range of contexts. However, if we want to retain the more precise meaning, we need to find paraphrases that are more semantically precise such as *cake cooked with apples* or *cake made using apples*. This trade-off between sensitivity and specificity has previously been modelled using Mutual Information and Conditional Probability for the task of filtering inferences (Pantel *et al.* 2007).

As already discussed, if we observe that a relational paraphrase r_2 occurs in a given context, the conditional probability of another relational paraphrase r_1 occurring in the same context is the number of contexts in which r_1 and r_2 have occurred together, divided by the number of contexts in which r_2 has occurred overall,

$$P(r_1|r_2) = \frac{P(r_1 \cap r_2)}{P(r_2)}$$

Due to the way the conditional probability uses co-occurrence counts of both phrases divided only by the marginal probability of one, it is not a symmetric similarity measure – phrase x may be highly probable given that phrase y occurs, but phrase y may still be improbable given the occurrence of phrase x.

Most similarity measures are symmetrical, and divide the shared features (such as co-occurrence in context) of two terms by some combination of their individual features. One such method is Pointwise Mutual Information (PMI):

$$\text{PMI}(x, y) = log \frac{P(x \cap y)}{P(x).P(y)}$$

This is a straightforward ratio relating the joint probability (the numerator) with the independent probability (the denominator). If the two variables are independent, their co-occurrence probability is expected to be equal to the product of their independent probabilities. Ignoring the logarithm, the only difference between this and the expression for conditional probability is that the prior probabilities of both variables are combined in the denominator, while in the conditional probability expression only the second variable is used in the denominator. As a result, PMI is symmetric, but conditional probability is asymmetric (or directional).

While PMI finds the most closely related phrase, conditional probability finds the phrase with the maximum posterior probability, given the occurrence of another phrase. As discussed above, this measure gives a high score to general phrases, and thus achieves a high accuracy due to the high coverage of general phrases. To illustrate the difference between the two methods, Tables 9 and 10 show examples of the highest-scoring phrases for an example seed phrase under each method.

The terms returned by the conditional probability association have a higher overall frequency in the data, and are more semantically general. The phrases returned by the mutual information association are more semantically precise.

Both measures use co-occurrence counts to measure association. The key difference is that conditional probability corrects for the prior probability of one of the terms (the 'observed' or 'given' term), while PMI corrects for the prior probability of both the terms. Therefore, the independent prior frequency of the two terms does

Table 9. *Conditional probability association*

be caused by	be made of	be used for	be during	be in
come from	contain	be for	be found in	be found in
be due to	consist of	be made for	be in	come from
result from	be made from	be related to	happen in	be related to
be related to	come from	be used in	occur during	be of
emerge from	be composed of	be found in	occur in	occur in
be created by	be of	come from	happen during	be located in
involve	use	provide	come in	be for
be made by	be made up of	help	begin in	belong to
be associated with	be in	be in	transpire in	involve
be from	have	deal with	be of	deal with

Table 10. *Mutual Information association*

be caused by	be made of	be used for
occur due to	be made out of	be used during
be because of	be constructed from	facilitate
occur after	be formed from	assist
be induced by	be cast from	receive
occur from	be fashioned from	help in
be due to	be created from	help with
result from	be constructed of	be required for
be by	be manufactured from	aid
precede	be composed of	cook
be created by	be formed of	be manufactured for

be during	be in
work in	be built in
transpire in	be situated in
come during	transpire in
begin in	come during
fall in	fall in
happen during	start in
commence in	exist in
be undertaken in	be during
fall during	happen during
occur during	be held in

not effect their PMI score, but the prior frequency of the first term will affect the conditional probability score.

We can parameterize the influence of the prior frequency of the second term to adjust the extent to which it influences the association measure. This gives a measure that can be adjusted to give any desired level of generality (and the associated good coverage and accuracy) or specificity (returning terms closer to the semantically

Table 11. *Paraphrases predicted for example compounds*

Compound: summer months	Seeds: be of, be by	
alpha = 0	alpha = 0.5	alpha = 1
come from	be found in	occur during
be found in	occur in	happen in
involve	occur during	happen during
occur in	happen in	begin in
be related to	happen during	transpire in
Compound: oil well	**Seeds: be in, produce**	
alpha = 0	alpha = 0.5	alpha = 1
be in	supply	be made for
be found in	have	be formed by
be related to	contain	involve
be used in	come from	be related to
belong to	be associated with	spew
Compound: sea breeze	**Seeds: come from, blow from, from**	
alpha = 0	alpha = 0.5	alpha = 1
emerge from	emerge from	emerge from
be in	be in	be in
contain	be made from	waft from
be made of	be made up of	originate from
be made from	be created from	blow over

detailed meaning of the target phrase),

$$\text{Score}(r_1|r_2) = \frac{P(r_1 \cap r_2)}{P(r_2)P(r_1)^{\alpha}}$$

With a value for α of 0, the formula reduces to the conditional probability association. With a value for α of 1, we get a measure like PMI. The higher the value of α, the more specific the terms returned tend to be.

Table 11 shows paraphrases extracted using this method for three compounds with different seed paraphrases retrieved from the corpus. In general, the higher the values of the parameter (and thus the closer the formula to PMI), the more unusual and semantically fine-grained the returned paraphrases.

5 Discussion

The uneven distribution of words in natural language has been extensively studied. Zipf (1935) first observed that the relationship between the rank of a word in the frequency list and its frequency followed a power-law distribution; the frequency of a word is inversely proportional to its rank in an ordered word-frequency list. While the Zipfian distribution might not be an indisputable hallmark of human language

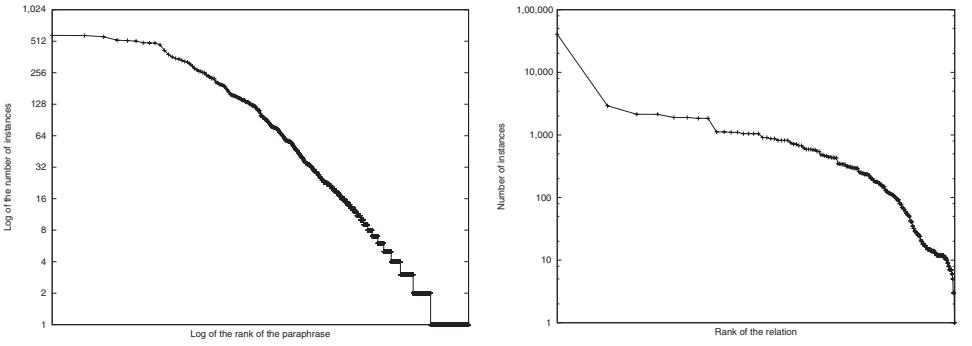


Fig. 1. Log (rank)–log (frequency) graphs of relational paraphrases from Butnariu and Veale (2008) (left) and Mohamed *et al.* (2011) (right).

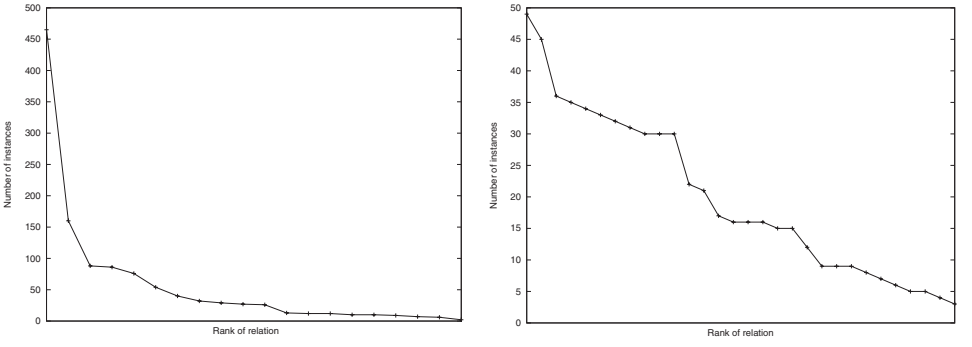


Fig. 2. Rank–frequency graphs of relation classes from Kim and Baldwin (2005) (left) and Nastase and Szpakowicz (2003) (right).

(Li 1992), it does demonstrate clearly that tokens in a language are distributed very unevenly among types.

Long-tailed distributions are evident in the log–log graphs of rank-frequency of paraphrasing semantic relations as shown in Figure 1. The ontology used in Mohamed, Hruschka and Mitchell (2011) explicitly represents a hierarchy among its semantic relations.

Similar patterns are also seen in the relationship between rank and frequency of abstract semantic relations between nouns. One of the most widely used datasets has been a set of 600 modifier noun compounds produced by Nastase and Szpakowicz (2003). These compounds were annotated with a general set of five abstract semantic relations, and also with thirty more specific relations. Another taxonomy of semantic relations was used in Kim and Baldwin (2005). The class distributions of these two datasets are shown in Figure 2. This uneven class distribution is a property of many semantic relation taxonomies, although some (e.g. O Séaghdha 2007) explicitly design the annotation requirements to strive for a relatively even distribution.

Manin (2008) suggests that Zipfian distributions may be a result of the hierarchical nature of the semantic space. This idea builds on an idea referred to by Zipf (1935) as the 'Principle of Least Effort', and later by Grice (1975) as the 'Maxim of Brevity'.

In an information theoretic view of language, both the speaker and the hearer would like to minimize the effort required to achieve effective communication. It has been shown that more frequent words are shorter and have faster lexical access times (Balota and Chumbley 1984). Given this, it makes sense for the speaker to use the most frequent word that is sufficient to communicate the intended meaning. Zipf (1945) demonstrates a 'meaning-frequency relationship of words', showing that the more frequent a word is, the more sub-senses it is likely to have in a dictionary. Simply put, frequent words are easier to access, but they are more ambiguous. Given these observations, it is unsurprising that the most frequent paraphrases are those that are semantically broader, and are super-senses of semantically narrower, less frequent paraphrases.

5.1 Conclusion and future work

In this paper, we have described methods for ranking and discovering relational paraphrases of noun compounds. We have also discussed how a hypernym relation between relating paraphrases can be predicted by their distributional generality, and describe how our algorithm can be adjusted to return semantically specific, precise paraphrases, or semantically general, broad-coverage paraphrase. We introduced a simple parameterized model that can be tuned to produce a measure of association on a continuum between conditional probability, which favours high recall, semantically general phrases, and pointwise mutual information, which features precise, semantically specific phrases.

The paraphrasing approach to noun compound disambiguation is a practical avenue for future research, but results from the SemEval 2010 competition indicate that current methods have only achieved 'moderate success', in particular when compared to the strong most-frequent-paraphrase baseline. Relation-extraction systems such as *Textrunner* (Banko *et al.* 2007) and Pantel and Pennacchiotti (2006) can retrieve relational paraphrases similar to those found by our system, and noun compound disambiguation might be considered to be a special case of the more general task of extracting semantic relations from text. Whether it is appropriate to disambiguate compounds with paraphrases or with abstract relations depends on the application and the level of granularity required, although we believe that the issues we have highlighted in this paper – such as the large coverage of a small number of general relations – apply also to the more general task of relation extraction. An avenue for future work would be to investigate the performance of paraphrases retrieved from the Web 1T corpus or information retrieval systems like Textrunner when used as features to predict abstract semantic relations.

The challenges presented by the task of discovering lexical expressions of semantic relations seem to have much in common with the task of word-sense-disambiguation: human agreement is low, substitution acceptability is a somewhat flawed method of evaluation, backoff baseline performance is strong, and the correlation between word frequency and semantic generality is at the core of the task (Stokoe, Oakes and Tait 2003; McCarthy *et al.* 2004). A promising avenue for evaluation of relational paraphrasing systems is to incorporate some of the methods that have been developed

to evaluate word sense disambiguation systems, such as lexical substitutability tasks which control for skewed frequency distributions (e.g. McCarthy and Navigli 2007; Sinha, McCarthy and Mihalcea 2010).

It seems intuitive that the power-law frequency distribution of word senses and relational paraphrases arises from an underlying hierarchical structure in the manner that is outlined in Manin (2008). Distributional tensor models (Baroni and Lenci 2010; Turney and Pantel 2010) capture this hierarchy implicitly, and can apply this knowledge to a wide range of semantic engineering tasks. Such models are to a certain extent 'black-box' representations, and from a theoretical linguistics perspective, an analysis along the lines of Kotlerman *et al.* (2010), which explicitly models the asymmetric semantic relations resulting from asymmetric distributions, is one of the most promising avenues.

Acknowledgments

This research was conducted while the first author was a research student at University College Dublin. We thank the School of Computer Science at UCD for their support. We are grateful also to the reviewers of this paper for their valuable comments and insights.

References

Baker, M. C. 2003. *Lexical Categories: Verbs, Nouns, and Adjectives*. Cambridge Studies in Linguistics. Cambridge, UK: Cambridge University Press.

Balota, D. A., and Chumbley, J. I. 1984. Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Science Psychology: Human Perception and Performance* **10**(3): 340–57.

Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., and Etzioni, O. 2007. Open information extraction from the web. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, Hyderabad, India, pp. 2670–76.

Baroni, M., and Lenci, A. 2010. Distributional memory: a general framework for corpus-based semantics. *Computational Linguistics* **36**(4): 673–721.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., and Taylor, J. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data (SIGMOD '08)*, pp. 1247–50. New York, NY: ACM.

Brants, T., and Franz, A. 2006. *Web 1T 5-gram Version 1*. California: Google Research, Google Inc.

Butnariu, C., Kim, S. N., Nakov, P., Séaghdha, Diarmuid Ó., Szpakowicz, S., and Veale, T. 2010. Semeval-2010 task 9: the interpretation of noun compounds using paraphrasing verbs and prepositions. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval '10)*, pp. 39–44. Uppsala, Sweden: Association for Computational Linguistics.

Butnariu, C., and Veale, T. 2008. A concept-centered approach to noun-compound interpretation. In *Proceedings of the 22nd International Conference on Computational Linguistics* – vol. 1 *(COLING '08)*, pp. 81–8. Stroudsburg, PA: Association for Computational Linguistics.

Daelemans, W., Zavrel, J., van der Sloot, K., and van den Bosch, A. 1999. *TiMBL: Tilburg Memory Based Learner – Version 2.0 – Reference Guide*. Tilburg, Netherlands: Tilburg University. Available at: http://ilk.uvt.nl/downloads/pub/papers/ilk.0707.pdf.

Essen, U., and Steinbiss, V. 1992. Cooccurrence smoothing for stochastic language modeling. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 1992 (ICASSP-92)*, vol. 1, pp. 161–4. New York: IEEE.

Ferrucci, D. A., Brown, E. W., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A., Lally, A., Murdock, J. W., Nyberg, E., Prager, J. M., Schlaefer, N., and Welty, C. A. 2010. Building Watson: an overview of the DeepQA project. *AI Magazine* **31**(3): 59–79.

Geffet, M., and Dagan, I. 2004. The distributional inclusion hypothesis and lexical entailment. In *Proceedings of the 42rd Annual Meeting of the ACL*, pp. 107–14. Ann Arbor, MI: Association for Computational Linguistics.

Girju, R., Moldovan, D., Tatu, M., and Antohe, D. 2005. On the semantics of noun compounds. *Computer Speech and Language* **19**(4): 479–96 (Special issue on Multiword Expression).

Grice, H. P. 1975. Logic and conversation. In P. Cole and J. L. Morgan (eds.), *Syntax and Semantics, vol. 3, Speech Acts,* pp. 41–58. San Diego, CA: Academic Press.

Hawker, T. (2006). Using Contexts of One Trillion Words for WSD. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, pp. 85–93.

Jackendoff, R. 2010. *Meaning and the Lexicon: The Parallel Architecture, 1975-2010*. Oxford, UK: Oxford University Press.

Johnston, M., and Busa, F. 1996. Qualia structure and the compositional interpretation of compounds. In *Proceedings of the ACL SIGLEX Workshop on Breadth and Depth of Semantic Lexicons*, Santa Cruz, CA, pp. 77–88. Stroudsburg, PA: ACL.

Kim, S. N., and Baldwin, T. 2005. Automatic interpretation of compound nouns using wordnet similarity. In *Proceedings of 2nd International Joint Conference on Natural Language Processing*, Jeju Island, Korea, pp. 945–56. Berlin, Germany: Springer.

Kim, S. N., and Baldwin, T. 2006. Interpreting semantic relations in noun compounds via verb semantics. In *Proceedings of the COLING/ACL Main Conference Poster Sessions (COLING-ACL '06)*, pp. 491–8. Stroudsburg, PA: Association for Computational Linguistics.

Kim, S. N., and Nakov, P. 2011. Large-scale noun compound interpretation using bootstrapping and the web as a corpus. In *EMNLP*, pp. 648–58. Stroudsburg, PA: ACL.

Kotlerman, L., Dagan, I., Szpektor, I., and Zhitomirsky-Geffet, M. 2010. Directional distributional similarity for lexical inference. *Natural Language Engineering*, **16**(4): 359–89.

Lauer, M. 1995. *Designing Statistical Language Learners: Experiments on Compound Nouns.* PhD thesis, Macquarie University, NSW, Australia .

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

Li, W. 1992. Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory* **38**(6): 1842–5.

Li, G., Lopez-Fernandez, A., and Veale, T. 2010. UCD-Goggle: a hybrid system for noun compound paraphrasing. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval '10)*, pp. 230–33. Stroudsburg, PA: Association for Computational Linguistics.

Manin, D. Y. 2008. Zipf's law and avoidance of excessive synonymy. *Cognitive Science* **32**(7): 1075–98.

McCarthy, D., Koeling, R., Weeds, J., and Carroll, J. 2004. Finding predominant word senses in untagged text. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pp 279–87. Stroudsburg, PA: Association for Computational Linguistics.

McCarthy, D., and Navigli, R. 2007. SemEval-2007 task 10: English lexical substitution task. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval '07)*, pp. 48–53. Stroudsburg, PA: Association for Computational Linguistics.

Mohamed, T., Hruschka, E., and Mitchell, T. 2011. Discovering relations between noun categories. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 1447–55. Edinburgh, Scotland: Association for Computational Linguistics.

Nakov, P. 2008. Noun compound interpretation using paraphrasing verbs: feasibility study. In D. Dochev, M. Pistore, and P. Traverso (eds.), *Artificial Intelligence: Methodology, Systems, and Applications*, pp. 103–17. Lecture Notes in Computer Science, vol. 5253. Berlin, Germany: Springer.

Nakov, P., and Hearst, M. 2006. Using verbs to characterize noun-noun relations. In J. Euzenat and J. Domingue (eds.), *Artificial Intelligence: Methodology, Systems, and Applications*, pp. 233–244. Lecture Notes in Computer Science, vol. 4183. Berlin, Germany: Springer.

Nastase, V., and Szpakowicz, S. 2003. Exploring noun-modifier semantic relations. In *Fifth International Workshop on Computational Semantics (IWCS-5)*, Tilburg, Netherlands, pp. 285–301. Berlin, Germany: Springer/Kluwer.

Nulty, P., and Costello, F. 2010. A comparison of word similarity measures for noun compound disambiguation. In L. Coyle and J. Freyne (eds), *Artificial Intelligence and Cognitive Science*, pp 231–40. Lecture Notes in Computer Science, vol. 6206. Berlin, Germany: Springer.

Ó Séaghdha, D. 2007. Designing and evaluating a semantic annotation scheme for compound nouns. In *Proceedings of the 4th Corpus Linguistics Conference*, July 27–30, University of Birmingham, UK.

Ó Séaghdha, D. 2008. *Learning Compound Noun Semantics*. PhD thesis, Computer Laboratory, University of Cambridge, Cambridge, UK. Published as Computer Laboratory Technical Report 735, University of Cambridge.

Pantel, P., Bhagat, R., Coppola, B., Chklovski, T., and Hovy, E. H. 2007. ISP: learning inferential selectional preferences. In C. L. Sidner, T. Schultz, M. Stone, and Cheng Xiang Zhai (eds.), *HLT-NAACL*, pp. 564–71. Edinburgh, Scotland: Association for Computational Linguistics.

Pantel, P., and Pennacchiotti, M. 2006. Espresso: leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pp. 113–20. Stroudsburg, PA: Association for Computational Linguistics.

Schubert, L. 2002. Can we derive general world knowledge from texts? In *Proceedings of the Second International Conference on Human Language Technology Research*, pp. 94–7. San Francisco, CA: Morgan Kaufmann.

Sinha, R., McCarthy, D., and Mihalcea, R. 2010. Semeval-2010 task 2: crosslingual lexical substitution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, July 15–16, Uppsala, Sweden.

Stokoe, C., Oakes, M. P., and Tait, J. 2003. Word sense disambiguation in information retrieval revisited. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval (SIGIR '03)*, pp. 159–66. New York, NY: ACM.

Tratz, S., and Hovy, E. 2010. A taxonomy, dataset, and classifier for automatic noun compound interpretation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL '10)*, pp. 678–87. Stroudsburg, PA: Association for Computational Linguistics.

Turney, P. D. 2006. Similarity of semantic relations. *Compututational Linguistics* **32**(September): 379–416.

Turney, P. D., and Pantel, P. 2010. From frequency to meaning: vector space models of semantics. *Journal of Artificial Intelligence Research (JAIR)* **37**: 141–88.

Vanderwende, L. 1994. Algorithm for automatic interpretation of noun sequences. In *Proceedings of the 15th Conference on Computational Linguistics, vol. 2 (COLING '94)*, pp. 782–8. Stroudsburg, PA: Association for Computational Linguistics.

Weeds, J., and Weir, D. 2005. Co-occurrence retrieval: a flexible framework for lexical distributional similarity. *Computational Linguistics* **31**(4): 439–75.

Weeds, J., Weir, D., and McCarthy, D. 2004. Characterising measures of lexical distributional similarity. In *Proceedings of COLING 2004*, pp. 1015–21. Geneva, Switzerland: COLING.

Welty, C., Fan, J., Gondek, D., and Schlaikjer, A. 2010. Large-scale relation detection. In *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading, (FAM-LbR '10)*, pp. 24–33. Stroudsburg, PA: Association for Computational Linguistics.

Wubben, S. 2010. UvT: memory-based pairwise ranking of paraphrasing verbs. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval '10)*, pp. 260–63. Stroudsburg, PA: Association for Computational Linguistics.

Zipf, G. K. 1935. *The Psychobiology of Language*. New York, NY: Houghton-Mifflin.

Zipf, G. K. 1945. The meaning-frequency relationship of words. *Journal of General Psychology* **1945**(33): 251–6.