UCD-PN: Selecting General Paraphrases Using Conditional Probability

Paul Nulty
University College Dublin
Dublin, Ireland
paul.nulty@ucd.ie

Fintan Costello
University College Dublin
Dublin, Ireland
fintan.costello@ucd.ie

Abstract

We describe a system which ranks human-provided paraphrases of noun compounds, where the frequency with which a given paraphrase was provided by human volunteers is the gold standard for ranking. Our system assigns a score to a paraphrase of a given compound according to the number of times it has co-occurred with other paraphrases in the rest of the dataset. We use these co-occurrence statistics to compute conditional probabilities to estimate a sub-typing or Is-A relation between paraphrases. This method clusters together paraphrases which have similar meanings and also favours frequent, general paraphrases rather than infrequent paraphrases with more specific meanings.

1 Introduction

SemEval 2010 Task 9, "Noun Compound Interpretation Using Paraphrasing Verbs", requires systems to rank paraphrases of noun compounds according to which paraphrases were most frequently produced for each compound by human annotators (Butnariu et al., 2010). This paper describes a system which ranks a paraphrase for a given compound by computing the probability of the paraphrase occurring given that we have previously observed that paraphrase co-occurring with other paraphrases in the candidate paraphrase list. These co-occurrence statistics can be built using either the compounds from the test set or the training set, with no significant difference in results.

The model is informed by two observations: people tend to use general, semantically light paraphrases more often than detailed, semantically heavy ones, and most paraphrases provided for a specific compound indicate the same interpretation of that compound, varying mainly according to level of semantic detail.

Given these two properties of the data, the objective of our system was to test the theory that conditional probabilities can be used to estimate a sub-typing or Is-A relation between paraphrases. No information about the compounds was used, nor were the frequencies provided in the training set used.

2 Motivation

Most research on the disambiguation of noun compounds involves automatically categorizing the compound into one of a pre-defined list of semantic relations. Paraphrasing compounds is an alternative approach to the disambiguation task which has been explored by (Lauer, 1995) and (Nakov, 2008). Paraphrases of semantic relations may be verbs, prepositions, or "prepositional verbs" like *found in* and *caused by*. (Lauer, 1995) categorized compounds using only prepositions. (Nakov, 2008) and the current task use only verbs and prepositional verbs, however, many of the paraphrases in the task data are effectively just prepositions with a copula, e.g. *be in*, *be for*, *be of*.

The paraphrasing approach may be easier to integrate into applications such as translation, query-expansion and question-answering — its output is a set of natural language phrases rather than an abstract relation category. Also, most sets of pre-defined 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.

Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pages 234–237,
Uppsala, Sweden, 15-16 July 2010.

2.1 General paraphrases are frequent

The object of SemEval 2010 Task 9 is to rank paraphrases for noun compounds given by 50-100 human annotators. When deciding on a model we took into account several observations about the data.

Firstly, 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 are genuinely ambiguous rather than vague. For example a *stone bowl* could *be a bowl for holding stones* or a *bowl made of stone*. 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 the compound *alligator leather* in the training data, the two most popular paraphrases 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 28 and 20 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.

According to the lesser-known of Zipf's laws — the law of meaning (Zipf, 1945) — words that are more frequent overall in a language tend to have more sub-senses. Frequent terms have a shorter lexical access time (Broadbent, 1967), so to minimize the effort required to communicate meaning of a compound, speakers should tend to use the most common words - which tend to be semantically general and have many possible subsenses. This seems to hold for paraphrasing verbs and prepositions; terms that have a high overall frequency in English such as *be in*, *have* and *be of* are vague — there are many more specific paraphrases which could be considered sub-senses of these common terms.

2.2 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 B occurring given that we have observed another paraphrase A in the same context is a measure of the extent to which B is a more general type (parent sense) of A.

3 System Description

The first step in our model is to generate a conditional probability table by going over all the compounds in the data and calculating the probability of each paraphrase occurring given that we observed another given paraphrase co-occurring for the same compound. We 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 note 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 con-

firmed that using only the training data for this step is not detrimental to the system's performance.

For each paraphrase in the data, the conditional probability of that paraphrase is computed with respect to all other paraphrases in the data. For any two paraphrases B and A:

$$P(B|A) = \frac{P(A \wedge B)}{P(A)}$$

As described in the previous section, we anticipate that more general, less specific paraphrases will be produced more often than their more detailed sub-senses. Therefore, we score each paraphrase by summing its conditional probability with each other paraphrase provided for the same compound.

For a list of paraphrases A provided for a given compound, we score a paraphrase *b* in that list by summing its conditional probability individually with every other paraphrase in the list.

$$score(b) = \sum_{a \in A} P(b|a)$$

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.

4 Results and Analysis

4.1 Task results

Table 1 shows the results of the top 3 systems in 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 helpful as the scores provide a scale of acceptability.

As mentioned in the system description, 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.

System	Spearman	Pearson	Cosine
UVT	**.450**	**.411**	.635
UCD-PN	.441	.361	**.669**
UCD-GOG	.432	.395	.652
baseline	.425	.344	.524

Table 1: Results for the top three systems.

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 a second set of predictions for the test based on the probabilities from the training compounds alone. [1]

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.

4.2 Error Analysis

The most significant drawback of this system is that it cannot generate 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. Firstly there is a clustering effect which favours paraphrases that are more similar to the other paraphrases in a list for a given compound. Secondly, 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. Firstly, 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.

[1] Thanks to Diarmuid Ó Séaghdha for pointing this out and scoring the second set of predictions

Also, because of the clustering effect, this system would not work well for compounds that are genuinely ambiguous e.g. *stone bowl* (*bowl made of stone* vs *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*. Without access to the gold-standard scores for these compounds it is difficult to explain the poor performance, but 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 8 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.

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.

5 Conclusion

We have described a system which uses a simple statistical method, conditional probability, to estimate a sub-typing relationship between possible paraphrases of noun compounds. From a list of candidate paraphrases for each noun compound, those which were judged by this method to be good "parent senses" of other paraphrases in the list were scored highly in the rankings.

The system does require a large dataset of compounds with associated plausible paraphrases, but it does not require a training set of human provided rankings and does not use any information about the noun compound itself, aside from the list of plausible paraphrases that were provided by the human annotators.

Given the simplicity of our model and its performance compared to other systems which used more intensive approaches, we believe that our initial observations on the data are valid: people tend to produce general, semantically light paraphrases more often than specific or detailed paraphrases, and most of the paraphrases provided for a given compound indicate a similar interpretation, varying instead mainly in level of semantic weight or detail.

We have also shown that conditional probability is an effective way to compute the sub-typing relation between paraphrases.

Acknowledgement

This research was supported by a grant under the FP6 NEST Programme of the European Commission (ANALOGY: Humans the Analogy-Making Species: STREP Contr. No 029088).

References

Donald E. Broadbent 1967. Word-frequency effect and response bias.. *Psychological Review*, 74,

Cristina Butnariu and Su Nam Kim and Preslav Nakov and Diarmuid Ó Séaghdha and Stan Szpakowicz and Tony Veale. 2010. SemEval-2 Task 9: The Interpretation of Noun Compounds Using Paraphrasing Verbs and Prepositions, *Proceedings of the 5th SIGLEX Workshop on Semantic Evaluation*, Uppsala, Sweden

Paul Grice. 1975. *Studies in the Way of Words*. Harvard University Press, Cambridge, Mass.

Mark Lauer 1995. *Designing statistical language learners: experiments on noun compound*, PhD Thesis Macquarie University, Australia

Preslav Nakov and Marti Hearst 2008. Solving Relational Similarity Problems using the Web as a Corpus. *In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics* (ACL-08), Columbus, OH.

George Kingsley Zipf. 1945. The Meaning-Frequency Relationship of Words. *Journal of General Psychology*, 33,