MELODI: A Supervised Distributional Approach for Free Paraphrasing of Noun Compounds

Tim Van de Cruys
IRIT, CNRS
tim.vandecruys@irit.fr

Stergos Afantenos
IRIT, Toulouse University
stergos.afantenos@irit.fr

Philippe Muller
IRIT, Toulouse University
philippe.muller@irit.fr

Abstract

This paper describes the system submitted by the MELODI team for the SemEval-2013 Task 4: Free Paraphrases of Noun Compounds (Hendrickx et al., 2013). Our approach combines the strength of an unsupervised distributional word space model with a supervised maximum-entropy classification model; the distributional model yields a feature representation for a particular compound noun, which is subsequently used by the classifier to induce a number of appropriate paraphrases.

1 Introduction

Interpretation of noun compounds is making explicit the relation between the component nouns, for instance that *running shoes* are shoes used in running activities, while *leather shoes* are made from leather. The relations can have very different meanings, and existing work either postulates a fixed set of relations (Tratz and Hovy, 2010) or relies on appropriate descriptions of the relations, through constrained verbal paraphrases (Butnariu et al., 2010) or unconstrained paraphrases as in the present campaign. The latter is much simpler for annotation purposes, but raises difficult challenges involving not only compound interpretation but also paraphrase evaluation and ranking.

In terms of constrained verbal paraphrases Wubben (2010), for example, uses a supervised memory-based ranker using features from the Google *n*-gram corpus as well as WordNet. Nulty and Costello (2010) rank paraphrases of compounds according to the number of times they co-occurred with other paraphrases for other compounds. They use these co-occurrences to compute conditional probabilities estimating is-a relations between paraphrases. Li et al. (2010) provide a hybrid system which combines a Bayesian algorithm exploiting Google *n*-grams, a score which captures human preferences at the tail distribution of the training data, as well as a metric that captures pairwise paraphrase preferences.

Our methodology consists of two steps. First, an unsupervised distributional word space model is constructed, which yields a feature representation for a particular compound. The feature representation is then used by a maximum entropy classifier to induce a number of appropriate paraphrases.

2 Methodology

2.1 Distributional word space model

In order to induce appropriate feature representations for the various noun compounds, we start by constructing a standard distributional word space model for nouns. We construct a co-occurrence matrix of the 5K most frequent nouns[1] by the 2K most frequent context words[2], which occur in a window of 5 words to the left and right of the target word. The bare frequencies of the word-context matrix are weighted using pointwise mutual information (Church and Hanks, 1990).

Next, we compute a joint, compositional representation of the noun compound, combining the se-

[1] making sure all nouns that appear in the training and test set are included

[2] excluding the 50 most frequent context words as stop words

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 144–147, Atlanta, Georgia, June 14-15, 2013. ©2013 Association for Computational Linguistics

mantics of the head noun with the modifier noun. To do so, we make use of a simple vector-based multiplicative model of compositionality, as proposed by Mitchell and Lapata (2008). In order to compute the compositional representation of a compound noun, this model takes the elementwise multiplication of the vectors for the head noun and the modifier noun, i.e.

$$p_i = u_i v_i$$

for each feature i. The resulting features are used as input to our next classification step.

We compare the performance of the abovementioned compositional model with a simpler model that only takes into account the semantics of the head noun. This model only uses the context features for the head noun as input to our second classification step. This means that the model only takes into account the semantics of the head noun, and ignores the semantics of the modifier noun.

2.2 Maximum entropy classification

The second step of our paraphrasing system consists of a supervised maximum entropy classification approach. Training vectors for each noun compound from the training set are constructed according to the approach described in the previous section. The (non-zero) context features yielded by the first step are used as input for the maximum entropy classifier, together with the appropriate paraphrase labels and the label counts (used to weight the instances), which are extracted from the training set.

We then deploy the model in order to induce a probability distribution over the various paraphrase labels. Every paraphrase label above a threshold ϕ is considered an appropriate paraphrase. Using a portion of held-out training data (20%), we set $\phi = 0.01$ for our official submission. In this paper, we show a number of results using different thresholds.

2.3 Set of paraphrases labels

For our classification approach to work, we need to extract an appropriate set of paraphrase labels from the training data. In order to create this set, we substitute the nouns that appear in the training set's paraphrases by dummy variables. Table 1 gives an example of three different paraphrases and the resulting paraphrase labels after substitution. Note that we did not apply any NLP techniques to properly deal with inflected words.

We apply a frequency threshold of 2 (counted over all the instances), so we discard paraphrase labels that appear only once in the training set. This gives us a total of 285 possible paraphrase labels.

One possible disadvantage of this supervised approach is a loss of recall on unseen paraphrases. A rough estimation shows that our set of training labels accounts for only 25% of the similarly constructed labels extracted from the test set. However, the most frequently used paraphrase labels are present in both training and test set, so this does not prevent our system to come up with a number of suitable paraphrases for the test set.

2.4 Implementational details

All frequency co-occurrence information has been extracted from the ukWaC corpus (Baroni et al., 2009). The corpus has been part of speech tagged and lemmatized with Stanford Part-Of-Speech Tagger (Toutanova and Manning, 2000; Toutanova et al., 2003). Distributional word space algorithms have been implemented in Python. The maximum entropy classifier was implemented using the Maximum Entropy Modeling Toolkit for Python and C++ (Le, 2004).

3 Results

Table 2 shows the results of the different systems in terms of the isomorphic and non-isomorphic evaluation measures defined by the task organizers (Hendrickx et al., 2013). For comparison, we include a number of baselines. The first baseline assigns the two most frequent paraphrase labels (*Y of X*, *Y for X*) to each test instance; the second baseline assigns the four most frequent paraphrase labels (*Y of X*, *Y for X*, *Y on X*, *Y in X*); and the third baseline assigns all of the possible 285 paraphrase labels as correct answer for each test instance.

For both our primary system (the multiplicative model) and our contrastive system (the head noun model), we vary the threshold used to select the final set of paraphrases. A threshold $\phi = 0.01$ results in a smaller set of paraphrases, whereas a threshold of $\phi = 0.001$ results in a broad set of paraphrases. Our official submission uses the former threshold.

compound	**paraphrase**	**paraphrase label**
textile company	company that makes textiles	Y that makes Xs
textile company	company that produces textiles	Y that produces Xs
textile company	company in textile industry	Y in X industry

Table 1: Example of induced paraphrase labels

model	ϕ	isomorphic	non-isomorphic
baseline (2)	–	.058	.808
baseline (4)	–	.090	.633
baseline (all)	–	.332	.200
multiplicative	.01	.130	.548
	.001	.270	.259
head noun	.01	.136	.536
	.001	.277	.302

Table 2: Results

First of all, we note that the different baseline models are able to obtain substantial scores for the different evaluation measures. The first two baselines, which use a limited number of paraphrase labels, perform very well in terms of the non-isomorphic evaluation measure. The third baseline, which uses a very large number of candidate paraphrase labels, gets more balanced results in terms of both the isomorphic and non-isomorphic measure.

Considering our different thresholds, the results of our models are in line with the baseline results. A larger threshold, which results in a smaller number of paraphrase labels, reaches a higher non-isomorphic score. A smaller threshold, which results in a larger number of paraphrase labels, gives more balanced results for the isomorphic and non-isomorphic measure.

There does not seem to be a significant difference between our primary system (multiplicative) and our contrastive system (head noun). For $\phi = 0.01$, the results of both models are very similar; for $\phi = 0.001$, the head noun model reaches slightly better results, in particular for the non-isomorphic score.

Finally, we note that our models do not seem to improve significantly on the baseline scores. For $\phi = 0.001$, the results of our models seem somewhat more balanced compared to the *all* baseline, but the differences are not very large. In general, our systems (in line with the other systems participating in the task) seem to have a hard time beating a number of simple baselines, in terms of the evaluation measures defined by the task.

4 Conclusion

We have presented a system for producing free paraphrases of noun compounds. Our methodology consists of two steps. First, an unsupervised distributional word space model is constructed, which is used to compute a feature representation for a particular compound. The feature representation is then used by a maximum entropy classifier to induce a number of appropriate paraphrases.

Although our models do seem to yield slightly more balanced scores than the baseline models, the differences are not very large. Moreover, there is no substantial difference between our primary multiplicative model, which takes into account the semantics of both head and modifier noun, and our contrastive model, which only uses the semantics of the head noun.

References

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Cristina Butnariu, Su Nam Kim, Preslav Nakov, Diarmuid Ó Séaghdha, Stan Szpakowicz, and Tony Veale. 2010. Semeval-2 task 9: The interpretation of noun compounds using paraphrasing verbs and prepositions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 39–44, Uppsala, Sweden, July. Association for Computational Linguistics.

Kenneth W. Church and Patrick Hanks. 1990. Word association norms, mutual information & lexicography. *Computational Linguistics*, 16(1):22–29.

Iris Hendrickx, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Stan Szpakowicz, and Tony Veale. 2013. SemEval-2013 task 4: Free paraphrases of noun compounds. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '13, June.

Zhang Le. 2004. Maximum entropy modeling toolkit for python and c++. `http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html`.

Guofu Li, Alejandra Lopez-Fernandez, and Tony Veale. 2010. Ucd-goggle: A hybrid system for noun compound paraphrasing. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 230–233, Uppsala, Sweden, July. Association for Computational Linguistics.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. *proceedings of ACL-08: HLT*, pages 236–244.

Paul Nulty and Fintan Costello. 2010. Ucd-pn: Selecting general paraphrases using conditional probability. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 234–237, Uppsala, Sweden, July. Association for Computational Linguistics.

Kristina Toutanova and Christopher D. Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*, pages 63–70.

Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL 2003*, pages 252–259.

Stephen Tratz and Eduard Hovy. 2010. A taxonomy, dataset, and classifier for automatic noun compound interpretation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 678–687, Uppsala, Sweden, July. Association for Computational Linguistics.

Sander Wubben. 2010. Uvt: Memory-based pairwise ranking of paraphrasing verbs. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 260–263, Uppsala, Sweden, July. Association for Computational Linguistics.