Predicting the Compositionality of German Compounds using Word Vectors

Jianqiang Ma, Corina Dima, Reinhild Barkey and Erhard Hinrichs
SFB 833 and Department of Linguistics, University of Tübingen
jianqiang.ma, corina.dima, reinhild.barkey, erhard.hinrichs@uni-tuebingen.de

1 Introduction

Compositionality is a key theme in semantics concerned with *whether* the semantics of large units of text can be constructed from the meanings of smaller units of text. For instance, while the meaning of most compounds can be composed from the meaning of their constituent words such as *Apfelsaft* 'apple juice', many compounds like *Eselsbrücke* 'mnemonic', literally 'donkey bridge', are semantically non-compositional.

A recent trend is to model compositionality using empirical data (Reddy et al. 2011; Krcmár et al., 2013; Schulte im Walde et al., 2013). Such approaches adopt word vectors (Section 3.1) from distributional semantics (DS) as the semantic representations of words. Compositionality is measured using the cosine similarity between the *observed* vector of a compound, estimated directly by a DS model from corpora, and its *composed* vector, which is a combination of the observed vectors of the compound's head and modifier.

Rather than relying on similarity scores, this paper proposes a method (Section 3.2) that *predicts* compositionality *directly* from the observed vectors of the compounds and their constituents using an artificial neural network. We show that cosine similarities of vectors used in previous work are a poor indicator for the task of predicting whether a compound is either compositional or non-compositional, i.e. binary classification. In contrast, the proposed method leads to significantly improved performance when evaluated (Section 4) on this binary classification task. The results suggest how far compositionality prediction can go by only using distributional information of words in the corpus.

This work also presents a semi-automatic procedure for generating the compositionality dataset (Section 2), which is used for evaluation in this paper. Compared with pure human annotation, the advantage of our approach is lower cost and larger resulting dataset.

2 Constructing a Compositionality Dataset using a Semi-Automatic Approach

GermaNet (Hamp and Feldweg, 1997) is a WordNet-like lexical semantic resource for the German language. It enlists the synsets of words and more importantly, the semantic relations between synsets, such as *synonymy* and *hypernymy*. GermaNet also contains more than 50,000 compounds together with their splitting information (Henrich and Hinrichs, 2011) e.g. the compound *Teebeutel* is split into its modifier *Tee* ‘tea’ and its head *Beutel* ‘bag’. We used a simple heuristic to determine the compositionality of a compound given the GermaNet hierarchy. A compound is considered *compositional* if a synset of its *head* is an ancestor along the hypernym path of the synset(s) of the compound (e.g. the sysnset for *Baum* ‘tree’ is a hypernym of the synset of *Apfelbaum* ‘apple tree’, therefore *Apfelbaum* is considered compositional). Otherwise, the compound is considered a *candidate non-compositional* compound. These candidates undergo a further manual check which filters out the remaining compositional instances, such as *Kanonenkugel* ‘cannon ball’. In this case the direct hypernym is *Geschoss* ‘projectile’, but the object is clearly also a type of *Kugel* ‘ball’. This type of examples are compositional although our simple heuristic does not consider them as such.

Our investigation revealed that compounds exhibit different types of non-compositionality. For the majority of non-compositional compounds it is only the head that conveys a non-literal meaning: e.g. *Datensalat* ‘data salad’, where the head *Salat* ‘salad’ is used in a *figurative sense*; *Schneemann*, ‘snow man’, where the compound *resembles* a *Mann* ‘man’ but is made out of snow. In some cases the resemblance pattern applies to both the modifier and the head. A point in case is the compound *Schneebesen* ‘whisk’, literally ‘snow broom’, where the head refers to the form of the whisk and the modifier to the egg whites that look like snow when whisked. Another type of non-compositionality refers to compounds where the meaning of the compound is completely different from the meaning of any of its constituents (e.g. *Eselsbrücke* ‘mnemonic’, literally ‘donkey bridge’). In our annotations we consider all the above cases as non-compositional compounds, without distinguishing among them.

The advantage of our GermaNet-based approach is the minimal manual check, as the clearly compositional compounds, which are the majority, are automatically selected beforehand. The manual check is still on-going. At the time this paper was written, the constructed dataset consisted of 5927 compositional and 724 non-compositional compounds.

3 Compositionality Prediction

3.1 Word Vectors

We use a set of word vectors for German described in Dima (2015), which has a vocabulary of 1 million words and where each word is represented by a 300-dimensional vector. These word vectors were trained with the GloVe package (Pennington et al., 2014), a state-of-the-art DS model, using a 10 billion token raw-text corpus extracted from the DECOW14AX corpus (Schäfer, 2015). The context of a word consists of 20 words, namely the 10 words to its left and to its right.

As compounds (and their constituent words) need to be frequent enough to yield high quality vectors, we select the compounds in the dataset in Section 2 where the compound and its constituents appear more than 500 times in the corpus. This lead to a final dataset with 3740 compounds, containing 3388 compositional and 352 non-compositional compounds. We use 70% of the data as the *training set* and the rest of 30% as the *test set*.

3.2 Compositionality Prediction with Neural Networks

The model proposed in this paper uses a neural network classifier (Bishop, 1995) to tackle the binary classification task of labeling compounds as compositional or non-compositional. The classifier is trained with examples from the *training set* of the form (*compound, modifier, head, class*), where *class* can only be *compositional* or *non-compositional*. Each word is represented as a 300-dimensional word vector (see Section 3.1). At each training step, the network is given a 900-dimensional vector as input (3 words x 300 dimensions) and outputs one of the two possible classes. If the prediction of the network was incorrect, the weights in the network are adjusted so that next time the network sees an input with similar values it is able to make a correct prediction. Otherwise, the training continues with the next example. The training ends when the predictions of the network stop improving.

We use a simple multilayer perceptron architecture consisting of a 900-dimensional input layer, a 900-dimensional hidden layer, a nonlinearity and a 2-dimensional output layer. To avoid over-fitting, we use a 0.75 dropout layer (Srivastava et al., 2014. The model is optimized with mini-batch gradient descent.

4 Evaluation

4.1 Task, Baseline and Metrics

This paper focuses on the binary classification of compounds as compositional or non-compositional. The simplest baseline is a 50/50 random guess of the compositionality, which would be correct in about 50% of the cases. However, a strong baseline is to predict *each* compound to be compositional (*all-compositional*), as most compounds are. This would lead to a correct prediction in 90.46% of the cases (equal to the percentage of compositional compounds in our test set) but would not distinguish between compositional and non-compositional compounds at all. Therefore, the overall number of correct predictions, that is, the *accuracy*, gives limited insight into performance of a model. A more detailed account is given by the *per-class* F1-score (Manning et al., 2008) for the two classes individually, which we adopt as a complementary measure. A good method should yield a high overall accuracy and high F-score on *non-compositional* compounds at the same time.

Previous methods in the literature judge the compositionality of a compound using the similarities between its observed and composed vectors. For a fair comparison, we trained neural network models in the same manner as our method, but used similarity scores instead of the vectors as input, denoted henceforth as *sim-based*.

4.2 Results

Table 1 shows the results of our model with three different input configurations, using: (i) the word vectors of the compounds (*predict([c])*); (ii) the vectors of the heads and modifiers of compounds (*predict([m,h])*) and (iii) all three vectors (*predict([c,m,h]*)). We also report the results of the *all-compositional* baseline and the *sim-based* method using similarities between the compound and (i) the head (*sim-based(h,c)*); and (ii) the vector addition of the head and the modifier (*sim-based(m+h,c)*). The difficulty of the task is reflected by the fact that both *sim-based* models fail to beat the *all-compositional* baseline. In contrast, the proposed model significantly outperforms the baseline in all three configurations, and works best when using all the three vectors (*predict([c, m, h])*). All the results are on the *test set*, while the training of the models only uses the non-overlapping *training set*.

model	non-compositional	compositional	overall accuracy
all-compositional	0	0.95	90.46%
sim-based(h,c)	0	0.95	90.46%
sim-based(m+h,c)	0	0.95	90.46%
predict([c])	0.29	0.95	91.18%
predict([m,h])	0.29	**0.96**	91.89%
predict([c,m,h])	**0.47**	**0.96**	**92.69%**

Table 1: Predicting the compositionality of German compounds

5 Conclusion

This paper presented a neural network model to predict compound compositionality, the input of which are the observed word vectors of compounds and their constituents. The evaluation shows that the semantic features in word vectors can better capture compound compositionality than similarity-based methods. We have also described a semi-automatic method of generating a large dataset of compound compositionality for German with minimal human effort. In summary, this work offers an alternative way of using distributional semantics in the study of compound compositionality.

References

Bishop. C. M. (1995) *Neural Networks for Pattern Recognition*. Oxford University Press.

Dima, C. (2015) Reverse-engineering language: a study on the semantic compositionality of German compounds. In *Proceedings of EMNLP 2015*.

Hamp, B., & Feldweg, H. (1997) Germanet - a lexical-semantic net for german. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*.

Henrich, V., & Hinrichs, E. W. (2011) Determining Immediate Constituents of Compounds in GermaNet. In *Proceedings of RANLP* 2011.

Krcmár, L., Jezek, K., & Pecina, P. (2013) Determining compositionality of expresssions using various word space models and methods. In *Proceedings of the ACL Workshop on Continuous Vector Space Models and their Compositionality*.

Manning, C. D., Raghavan, P., & Schütze, H. (2008) *Introduction to information retrieval* (pp. 142). Cambridge University Press.

Pennington, J., Socher, R., & Manning, C. D. (2014) Glove: Global vectors for word representation. In *Proceedings of EMNLP 2014*.

Reddy, S., McCarthy, D., & Manandhar, S. (2011) An empirical study on compositionality in compound nouns. In *Proceedings of IJCNLP 2011*.

Schäfer, R. (2015) Processing and querying large web corpora with the COW14 architecture. In *Challenges in the Management of Large Corpora (CMLC-3)*.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014) Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*(1), 1929-1958.

Im Walde, S. S., Müller, S., & Roller, S. (2013) Exploring vector space models to predict the compositionality of German noun-noun compounds. In *Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics*.