Measuring the compositionality of noun-noun compounds over time

Prajit Dhar
Leiden University
dharp@liacs.leidenuniv.nl

Janis Pagel
University of Stuttgart
pageljs@ims.uni-stuttgart.de

Lonneke van der Plas
University of Malta
lonneke.vanderplas@um.edu.mt

Abstract

We present work in progress on the temporal progression of compositionality in noun-noun compounds. Previous work has proposed computational methods for determining the compositionality of compounds. These methods try to automatically determine how transparent the meaning of the compound as a whole is with respect to the meaning of its parts. We hypothesize that such a property might change over time. We use the time-stamped Google Books corpus for our diachronic investigations, and first examine whether the vector-based semantic spaces extracted from this corpus are able to predict compositionality ratings, despite their inherent limitations. We find that using temporal information helps predicting the ratings, although correlation with the ratings is lower than reported for other corpora. Finally, we show changes in compositionality over time for a selection of compounds.

1 Introduction

Compositionality is a long debated issue in theoretical linguistics. The principle of compositionality (Partee, 1984) states that the meaning of an expression is a function of the meanings of its parts and of the way they are syntactically combined. It is often used to describe how the meaning of a sentence can be derived from the meaning of single words and phrases, but the principle might also be postulated for *compounding*, i.e. the process of combining two or more lexemes to form a new concept (Bauer, 2017, p. 1 and 4). Compounds can often be directly derived from the meanings of the involved compound constituents (e.g. *graduate student*, *speed limit*), however, we also find compounds whose meanings can only be derived

partially from their components (*night owl*, *hot dog*).

Surprisingly, diachronic perspectives on compositionality[1] are virtually absent from previous work. To the best of our knowledge, we present the first study on the compositionality of compounds over time. We bring two strands of research together. On the one hand we are inspired by the synchronic work on predicting the degree of compositionality of compounds by comparing the vector-based representations of the parts to the vector-based representations of the compound as a whole. On the other hand, we rely on methods designed for detecting semantic change, such as presented in Hamilton et al. (2018), to study compositionality in compounds from a diachronic viewpoint.

2 Related Work

From a synchronic perspective, Reddy et al. (2011), Schulte im Walde et al. (2013) and Schulte im Walde et al. (2016a) are closest to our approach, since they predict the compositionality of compounds using vector space representations. However, Schulte im Walde et al. (2013) use German data and do not investigate diachronic changes. They report a Spearman's ρ of 0.65 for predicting the compositionality of compounds based on the features of their semantic space and find that the modifiers mainly influence the compositionality of the whole compound, contrary to their expectation that the head should be the main source of influence. This is true for both the human annotation and their vector space model.

---

[1]A notable exception is Vincent (2014), although he mainly focuses on syntactic processes in Romance languages and only briefly covers numeral words.

Schulte im Walde et al. (2016a) further investigate the role of heads and modifiers on the prediction of compositionality and report ρ values between 0.35 and 0.61 for their models on German and English data. Reddy et al. (2011) also report Spearman's ρ between their surveyed compositionality values and word vectors. They achieve ρ values of around 0.68, depending on the model.

From a diachronic perspective, we follow the general methodological approach of Hamilton et al. (2018), who use PPMI, SVD and word2vec based vector spaces to investigate a shift in meaning for chosen words with a known semantic change (*gay*, *broadcast*, etc.). They use time series to detect a significant change-point for two words, using cosine similarity and Spearman's ρ. They also compute the displacement for a single word embedding by calculating the cosine similarity between a point in time t and a later point in time $t + \Delta$. We adapt this methodology and make use of the same corpus (Google Books Ngram).

3 Methods and Data

Several studies have been conducted in order to measure compositionality for compounds in different languages (von der Heide and Borgwaldt, 2009; Reddy et al., 2011; Schulte im Walde et al., 2016b). Some of these works have used large corpora to extract vector-based representations of compounds and their parts to automatically determine the compositionality of a given compound. The models were validated on the basis of their correlation with human compositionality ratings for a set of compounds.

Because we are interested in the diachronic perspective on compounds, we use a time-stamped corpus: the Google Books Ngram corpus[2] (Michel et al., 2011) It contains books from the 1500s to the 2000s, from which we retrieve the contextual information of compounds and their constituents per year. We operate on 5-grams, which is the largest unit provided by Google Ngrams and use the words appearing in the 5-grams as both target words and context. We use the Part-of-Speech information already included in the Google Ngram corpus to extract noun-noun patterns. We then regard all other tokens in the 5-gram as context words and from this build up a semantic space representation of noun compounds for each year. We use a sliding window approach, wherein we capture the context of a compound based on its position in the 5-gram. That means that a bigram (say the compound *gold mine*) could occur in four different positions in the 5-grams (1-2, 2-3, 3-4 and finally 4-5). We then capture the contexts for each of these positions, in order to enrich the representation of a compound and its constituents (which similarly have five such positions, as they are unigrams).

Ideally, we would validate our diachronic model on diachronic test data. However, as it is not possible to survey compositionality rating for diachronic data, we instead use the synchronic data provided by Reddy et al. (2011) (henceforth referred to as REDDY) for evaluating the quality of the Google Books Ngram data as a source for investigating the compositionality of compounds in general. Reddy et al. (2011) compiled a list of 90 English compounds and asked annotators to rate the compositionality of these compounds on a scale from 0 to 5. They provide three mean values of their ratings for the compounds (*compound-mean*), heads (*head-mean*) and modifiers (*modifier-mean*). We make use of REDDY in order to verify that our methods are capable of capturing compositionality (synchronically) and use the diachronic data of Google Books Ngram to investigate the temporal change of compositionality.

A common challenge in building semantic spaces on a diachronic scale is that when building the spaces for individual spans of time, the spaces need to be aligned later on in order to compare models (see e.g. Kutuzov et al., 2018, Section 3.3). We circumvent this problem by jointly learning the spaces for the target words. To do this, we take the sparse representations of the compounds and their constituents and jointly learn their dense representations using SVD. Similar to Hamilton et al. (2018) we also choose the dimensions of our embeddings to be 300. We carry out row normalization on the embeddings, in order to remove the bias of the frequency of the compounds and their constituents.

We make use of six different semantic features that have been proposed in the literature to capture compositionality (Schulte im Walde et al., 2016a) and plausibility of noun-noun compounds (Günther and Marelli, 2016; Dhar and van der

---

[2]The data is available from `https://commondatastorage.googleapis.com/books/syntactic-ngrams/index.html`

Plas, 2019). Three features are based on the cosine similarity between the embeddings of different compound parts (see Günther and Marelli, 2016):

1. Similarity between compound constituents (*sim-bw-constituents*)

2. Similarity of the compound with its head (*sim-with-head*)

3. Similarity of the compound with its modifier (*sim-with-mod*)

The three information theory based features given below were proposed by Dhar and van der Plas (2019):

4. Log likelihood-ratio (*LLR*)

5. Positive Pointwise Mutual Information (*PPMI*)

6. Local Mutual Information (*LMI*)

Such formulas have been used prior to calculate collocations and associations between words (compare Manning and Schütze, 1999). Each feature will be tested individually for its ability to capture compositionality.

4 Experiments

We ran a total of two experiments[3] (Section 4.2 and 4.3) with different goals.

4.1 Experimental Setup

Hyper-parameters We experiment with certain hyper-parameters, in particular we varied the *time span length*, e.g. single years, decades or a span of 20 years etc. and *frequency cut-off* of compounds and their constituents in a specific time span, i.e. compounds and constituents have to occur above a certain frequency threshold. Choosing a greater time span will increase the observable data per compound and might improve the vector representations. We only consider compounds which retain representations in all time spans starting from the year 1800, which reduces the number of total compounds depending on the specific setup.

---

[3]The code is available at `https://github.com/prajitdhar/Compounding`

Compound-centric setting Dhar and van der Plas (2019) found the compound-centric set up, where the distributional representations of words are based on their usage as constituents in a compound to outperform compound-agnostic setups, for predicting novel compounds in English. They were inspired by research on N-N compounds in Dutch that suggests that constituents such as -molen '-mill' in pepermolen 'peppermill' are separately stored as abstract combinatorial structures rather than understood on the basis of their independent constituents (De Jong et al., 2002). We hence adopt the compound-centric setting.

4.2 Correlation

We first carry out a quantitative experiment, to see if our features bolster the prediction of compositionality in noun-noun compounds. To do so, we calculate correlation scores between our proposed metrics and the annotated compositionality ratings of REDDY. Like Reddy et al. (2011) and Schulte im Walde et al. (2013), we opt for Spearman's ρ.

To find the best configuration of a time span and cut-off for the regression models, we use the R^2 metric. Table 1 presents our findings; we will discuss them in the following Section 5.

4.3 Progression of Compositionality over Time

Based on the results of our correlation experiment, we proceed to analyze the temporal progression of compositionality. Our goals are two-fold: First, investigate if temporal information helps in predicting the contemporary REDDY data and second, use the best feature and setup in order to model the progression of compositionality over time.

5 Results

We find the best predictors for the compositionality ratings of REDDY to be *LMI* and *LLR* (compare Table 2). The overall highest correlation occurs between *compound-mean* and *LMI* with ρ of 0.54. We also see that *sim-bw-constituents* and *sim-with-heads* are generally good predictors as well. Contrary to Schulte im Walde et al. (2013) we do not find a strong correlation between modifiers and the REDDY ratings. Interestingly, *PPMI* is always weakly negatively correlated with the ratings. This could be due to *PPMI*'s property of inflating scores for rare events. As can also be seen

236

Time span	Cut-off	$R^2 \pm$ sd
	20	0.343 ± 0.028
NA (Non-temporal)	50	0.344 ± 0.026
	100	0.337 ± 0.035
	20	0.350 ± 0.029
1 (Year)	50	0.171 ± 0.039
	100	0.326 ± 0.030
	20	0.332 ± 0.024
10 (Decade)	50	0.328 ± 0.034
	100	0.360 ± 0.062
	20	0.341 ± 0.039
20 (Score)	50	0.331 ± 0.031
	100	$\mathbf{0.370 \pm 0.012}$
	20	0.352 ± 0.038
50 (Half-century)	50	0.360 ± 0.029
	100	0.364 ± 0.034
	20	0.351 ± 0.037
100 (Century)	50	0.343 ± 0.033
	100	0.344 ± 0.034

Table 1: R^2 values and standard deviation for the different configurations.

from Table 2, our correlation values are considerably lower than that of Reddy et al. (2011), but on par with a replication study by Schulte im Walde et al. (2016a) for *compound-mean*. We speculate that these differences are potentially due to the use of different data sets, the fact that we use a considerably smaller context window for constructing the word vectors (5 due to the restrictions of Google Ngram corpus vs. 100 in Reddy et al. (2011) and 40 in Schulte im Walde et al. (2016b)) and the use of a compound-centric setting (as described in 4.1).

From Table 1 we see that our best reported R^2 value occurs when observing time in stretches of 20 years (scores) and compounds having a frequency cut-off of at least 100. A few other observations could be made: In general the cut-off seems to improve the R^2 metric and the time spans of 10 and 20 years appear to be the most informative and stable for the cut-off values. Also, using temporal information almost always outperforms the setup that ignores all temporal information.

For our following experiment, we choose to use the configuration with the highest R^2 value: a time span of 20 years and a cut-off of 100. Since LMI

achieved the highest ρ values, we also choose LMI over the other features. We group the compounds of REDDY into three groups based on the human ratings they obtained: *low* (0-1), *med* (2-3) and *high* (4-5). Each group contains around 30 compounds. We then plot the LMI values of these three groups with their confidence interval across the time step of 20 years, shown in Figure 1. We can observe that there is a separation between the groups towards the later years, and that the period between 1940s and 1960s caused a noticeable change in the compositionality of the REDDY compounds. We find the same trends for all three information theory based features. Although care should be taken given the small data sets (especially for the earlier decades) on which the models were build and tested, the slope of the lines for the three groups of compounds seems to suggest that less compositional compounds go through a more pronounced change in compositionality than compositional compounds, as expected.

We also show the graphs for *sim-with-head* and *sim-with-mod* (Figures 2 and 3) for the different groups of compounds across time, as these underperformed in our previous experiment. Both figures based on cosine based features largely confound the three groups of compounds across time, reinforcing our previous findings.
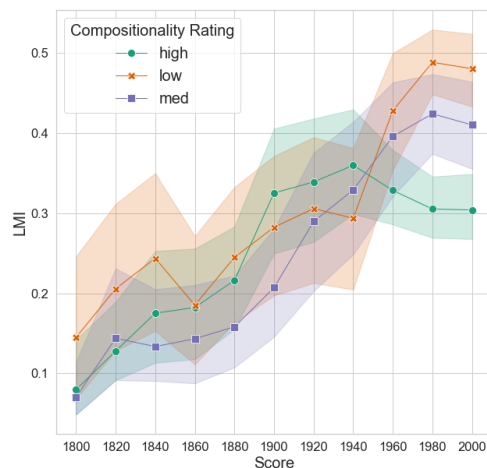

Figure 1: LMI of a compound in time point t and $t + 1$, with a time span of 20 years and a frequency cut-off of 100. Compounds are grouped according to their rating in REDDY.

	modifier-mean	head-mean	compound-mean
sim-bw-constituents	0.35	0.41	0.48
sim-with-head	0.26	0.43	0.43
sim-with-mod	0.1	0.18	0.2
LLR	0.36	0.44	0.52
PPMI	-0.12	-0.1	-0.14
LMI	**0.38**	**0.45**	**0.54**

Table 2: Spearman's ρ of our measures and the compositionality ratings of REDDY.
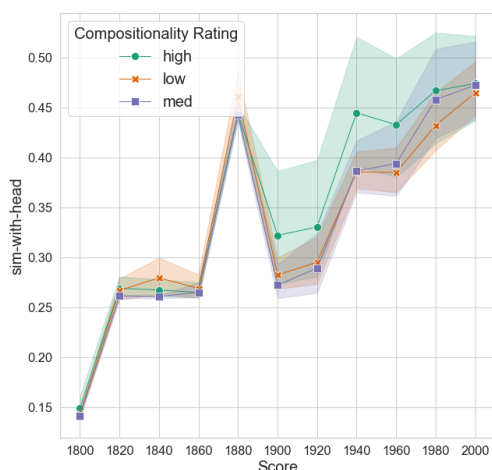

Figure 2: *sim-with-head* of a compound in time point t and $t+1$, with a time span of 20 years and a frequency cut-off of 100. Compounds are grouped according to their rating in REDDY.
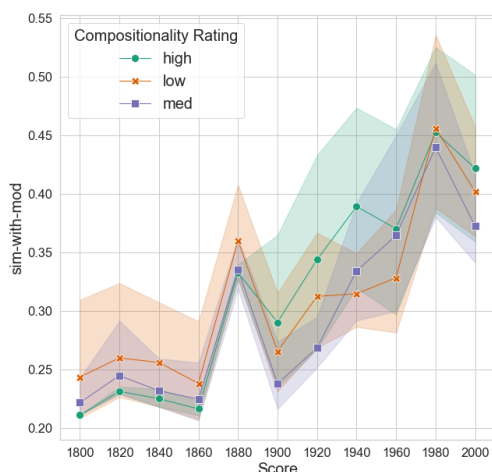

Figure 3: *sim-with-mod* of a compound in time point t and $t+1$, with a time span of 20 years and a frequency cut-off of 100. Compounds are grouped according to their rating in REDDY.

6 Future Work

Our current work was limited to English compounds from Reddy et al. (2011). We plan to expand our models to other languages for which compositionality ratings are available, such as German. We would also like to further investigate the fact that the information theory based measures outperform the ones based on cosine similarity. We intend to do so by incorporating more compounds and their compositionality ratings, as well as by using larger corpora.

Lastly, we will seek to find ways to harvest proxies for compositionality ratings of compounds over time. A possible avenue could be to use the information available in dictionaries.

7 Conclusion

We have shown work in progress on determining the compositionality of compounds over time. We showed that for our current setup, information theory based measures seem to capture compositionality better. Furthermore, we showed that adding temporal information increases the predictive power of these features to prognosticate synchronic compositionality. Finally, we showed how our best performing models trace the compositionality of compounds over time, delineating the behavior of compounds of varying levels of compositionality.

Acknowledgements

References

Laurie Bauer. 2017. *Compounds and Compounding*, volume 155 of *Cambridge Studies in Linguistics*. Cambridge University Press, Cambridge.

Nivja H. De Jong, Laurie B. Feldman, Robert Schreuder, Matthew Pastizzo, and R. Harald Baayen. 2002. The processing and representation of dutch and english compounds: peripheral morphological and central orthographic effects. *Brain and Language*, 81:555–67.

Prajit Dhar and Lonneke van der Plas. 2019. Learning to predict novel noun-noun compounds. In *Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*.

Fritz Günther and Marco Marelli. 2016. Understanding Karma Police: The Perceived Plausibility of Noun Compounds as Predicted by Distributional Models of Semantic Representation. *PLoS ONE*, 11(10):1–36.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2018. Diachronic word embeddings reveal statistical laws of semantic change. *CoRR*, abs/1605.09096.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.

Barbara H. Partee. 1984. Compositionality. In *Varieties of Formal Semantics: Proceedings of the 4th Amsterdam Colloquium, Sept. 1982*, pages 281–311. Foris Publications.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 210–218, Chiang Mai, Thailand. AFNLP.

Sabine Schulte im Walde, Anna Hätty, and Stefan Bott. 2016a. The role of modifier and head properties in predicting the compositionality of English and German noun-noun compounds: A vector-space perspective. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics (*SEM 2016)*, pages 148–158.

Sabine Schulte im Walde, Anna Hätty, Stefan Bott, and Nana Khvtisavrishvili. 2016b. $G_h$ost-NN: A Representative Gold Standard of German Noun-Noun Compounds. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 2285–2292, Portoroz, Slovenia.

Sabine Schulte im Walde, Stefan Müller, and Stephen Roller. 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*, pages 255–265, Atlanta, GA, USA.

Nigel Vincent. 2014. Compositionality and change. In Claire Bowern and Bethwyn Evans, editors, *The Routledge Handbook of Historical Linguistics*, pages 103–123. Routledge, United Kingdom.

Claudia von der Heide and Susanne Borgwaldt. 2009. Assoziationen zu Ober-, Basis- und Unterbegriffen. Eine explorative Studie. In *Proceedings of the 9th Norddeutsches Linguistisches Kolloquium*, pages 51–74, Bielefeld, Germany.