Evaluating compositionality function on existing models: 'Adjective + Noun' composition

Sanghee Kim
Computational Linguistics Seminar (Spring 2020)

1 Introduction

Composition is known to be one of the key mechanism that grounds natural language. Composition in natural language is determined by the way that parts are combined. Composition can be defined as a function that combines sub-components of a larger component in natural language. A number of attempts have been made to model such compositional functions. These include learning composition of the 'adjective + noun' phrase (Mitchell and Lapata, 2010; Bride et al., 2015; Asher et al., 2016), the 'noun + noun' compound (Salehi et al., 2015; Cordeiro et al., 2016), the 'verb + noun' phrase (Mitchell and Lapata, 2008, 2010), the 'verb + particle' (Shwartz and Dagan, 2019), and the light verb construction (Shwartz and Dagan, 2019), among the work on composing two linguistic elements.

The goal of this paper is to evaluate compositionality in existing models on compositionality functions. We test the composition of the 'adjective + noun' phrase, in particular. The present study differs from earlier work on composition of the 'adjective + noun' phrase (Mitchell and Lapata, 2010; Baroni and Zamparelli, 2010; Vecchi et al., 2017; Hartung et al., 2017; Shwartz and Dagan, 2019) in that we use a more fine-grained adjective category to evaluate the compositionality functions. Grounded on the well-attested observations on adjectives that not all adjectives are the same (Partee, 2010; Pustejovsky, 2013), we suggest that compositionality functions should be evaluated accordingly to the different inferential patterns of adjectives. While the necessity to evaluate compositionality functions with fine-grained categorization on adjectives has been discussed in some earlier work (e.g. Boleda et al., 2013; Nayak et al., 2014; Pavlick and Callison-Burch, 2016), this is the first study to the best of our knowledge that used four different types of adjectives and tested the existing compositionality functions on the 'adjective + noun' phrase. In specific, in this study, we use the following four adjective categories: INTERSECTIVE, SUBSECTIVE, PLAIN NON-SUBSECTIVE, and PRIVATIVE adjectives (Partee, 2010; Pustejovsky, 2013). Moreover, as opposed to previous work, which used different nouns depending on the type of the adjective, we used the same noun, the *constant* noun, that uniformly combined with all different types of adjectives. This allowed us to specifically test the compositionality of the two elements, avoiding unwanted artifacts other than composition.

2 Types of adjectives

We briefly summarize the well-attested inferential pattern of the four types of adjectives. The four categories we use to test compositionality functions are as, INTERSECTIVE, SUBSECTIVE, PRIVATIVE, and PLAIN NON-SUBSECTIVE (e.g. Partee, 2010; Pustejovsky, 2013), which can be distinguished by set relation with the composing nouns.

Intersective The denotation of the [adjective noun] phrase is an intersection of the denotation of the adjective and the denotation of the noun (1). The intersective type of adjectives include "rectangular" or "American", where a "rectangular map" is a "map" and is "rectangular", and an "American singer" is a "singer" and is "American".

(1) $[\![\text{A N}]\!] = [\![\text{A}]\!] \cap [\![\text{N}]\!]$

Subsective The denotation of the [adjective noun] phrase is a subset of the denotation of the noun (2). Adjectives such as "big", "cold", "skillful", for instance, fall into this category. While a "big ant" is an "ant", it is not necessarily "big"; while a "cold star" is a "star", it is not necessarily "cold'.

(2) $[\![\text{A N}]\!] \subseteq [\![\text{N}]\!]$

Plain non-subsective The denotation of the [adjective noun] phrase may or may not be a subset of the denotation of the noun (3). Adjectives such as "alleged" or "supposed" fall into this category. For instance, it is possible that an "alleged criminal" is a "criminal" or not; a "supposed improvement" may or may not be an "improvement".

(3) $[\![\text{A N}]\!] \cap [\![\text{N}]\!] \neq \emptyset$

Privative The denotation of the [adjective noun] phrase is not a subset of the denotation of the noun (4). In terms of the set relation, there is no overlapping part between the set of all the entities that refer to the phrase and the set of all the entities that refer to the noun. For instance, there is no intersection of the denotation of "fake diamond" and 'diamond'.

(4) $[\![\text{A N}]\!] \cap [\![\text{N}]\!] = \emptyset$

We use these four types of adjectives to evaluate existing models on compositionality function.

3 Related work

Work on modeling compositionality can be largely categorized into four types. The first one is a popular approach in the computational linguistics community. This approach models compositionality functions under the assumption that the meaning of a word can be represented with vector space models (VSMs), built on the statistical usage of the word in the corpora. From this distributional semantics model, the meaning of the 'adjective + noun' phrase can be defined by the compositional output of two words, word 1, and word 2, which in our case correspond to the adjective, and the noun, respectively. The compositionality function refers to the way that these two words combine to the get the full meaning of the phrase (Mitchell and Lapata, 2008, 2010).

The second approach is similar to the first approach but differs in that some words are represented in terms of matrices as opposed to vectors (Baroni and Zamparelli, 2010; Guevara, 2010; Bride et al., 2015; Vecchi et al., 2017). When composing the adjective and the noun, for instance, the adjective is defined as a matrix. In this case, the matrix representation of the adjective itself serves as the compositional function. For this reason, this approach has been referred to as the *lexical function* approach. From this approach, the noun is still construed as a vector. The function of the adjective can either be tailored to specific lexical types, or can be treated as a general composition process for combining an adjective and a noun (Bride et al., 2015). Work that used the lexical function approach showed that this approach outperformed the models used in Mitchell and Lapata (2008, 2010), the first type of approach.

The third approach takes into consideration of an ontological property that defines the meaning of the full phrase (Hartung et al., 2017). In this case, the meaning of the 'adjective + noun' phrase is defined by the meaning of the adjective, the noun, and the property (e.g., 'temperature' for the word 'hot'). This is different from the first two models in that this model assumes an additional or an abstract ontological property.

The final approach views the composed structure to be represented through contextualized word embeddings (Shwartz and Dagan, 2019). Using state-of-the-art models (e.g. Peters et al., 2018; Radford et al., 2018; Devlin et al., 2019), this approach models composition by computing the word embeddings given the contextual sentence rather than combining the outputs of each element through VSMs.

The present study takes the first line of approach under the assumption that word meanings can be represented as vectors. We use pre-trained `word2vec` (Mikolov et al., 2013) model for word representation. We summarize in the following section the crucial compositionality functions that were built based on this assumption.

3.1 Mitchell and Lapata (2010)

Mitchell and Lapata (2008, 2010) examined a wide variety of compositional functions, mainly the additive function and the multiplicative function. Mitchell and Lapata (2010) explored variants of additive models, including the weighted additive model, the Kintsch model, and dilation. We summarize the idea of the weighted additive model, which showed the best performance among the variants of the additive models. Given the two elements, the vector representation of the adjective u and the vector representation of the noun v, the weighted additive model derives the composed phrase by summing the vector representation of the two components (Eq. 1).

$$\mathbf{p} = \alpha\mathbf{u} + \beta\mathbf{v} \quad (1)$$

Mitchell and Lapata (2010) also tested a wide variety of multiplicative models, including the simple

multiplicative model, tensor product, and circular convolution. What the multiplicative models do is the capture the contribution of one element to another; the function captures the relevance of one element to the other element. The top two models among the family of the multiplicative models that performed well were the simple multiplicative model (Eq. 2) and the tensor product (Eq. 3). The simple multiplicative model does a point-wise multiplication.

$$\mathbf{p}_i = \mathbf{u}_i \cdot \mathbf{v}_i \tag{2}$$

Tensor product obtains all pairwise product of the components.

$$\mathbf{p}_{i,j} = \mathbf{u}_i \cdot \mathbf{v}_j \tag{3}$$

Mitchell and Lapata (2010) used a human behavioral result as a reference to evaluate the model. In a human behavioral task, the participants were given two pairs of 'adjective + noun' phrases (e.g., 'social worker', 'wide range') and were instructed to rate the similarity of the two phrases from a scale of 1 to 7. The authors also obtained the cosine similarity between the two phrases. They then calculated the correlation between the human rating score and the cosine similarity and used it as a metric of evaluating the performance of the model.

All the tested models showed meaningful correlation with the human rating results. The simple multiplicative model significantly outperformed any other multiplicative models, and the weighted additive model and the dilation model performed better than any other additive models. A similar trend was found with the Latent Dirichlet Allocation (LDA) topic model. The weighted additive model and the dilation model outperformed any other models; in this case, the tensor product performed better than other multiplicative models.

4 Method

The goal of this paper is to evaluate existing models how well they reflect compositionality. I test two compositional functions that performed well in Mitchell and Lapata (2010): the weighted additive model and the (simple) multiplicative model. We used the pre-trained `word2vec` model (Mikolov et al., 2013) provided in spaCy, a free open-source library for Natural Language Processing in Python.

4.1 Weighted additive model

Based on the weighted additive model, the meaning of the phrase is a compositional result of the meaning of the adjective and the meaning of the noun. Hence the vector representation of the phrase (p) can be understood as the sum of (a) the vector representation of the adjective (u) with a certain weight α, and (b) the vector representation of the noun (v) with a certain weight β (Eq. 4).

$$\mathbf{p} = \alpha\mathbf{u} + \beta\mathbf{v} : \begin{cases} \alpha = 0.88, & \beta = 0.12 \\ \alpha = 0.6, & \beta = 0.4 \\ \alpha = 0.3, & \beta = 0.7 \end{cases} \tag{4}$$

The hyper-parameters have been optimized differently in previous work. The parameters have been tuned to $\alpha = 0.88$ and $\alpha = 0.12$ in Mitchell and Lapata (2010), $\alpha = 0.3$ and $\alpha = 0.7$ in Vecchi et al. (2017), and $\alpha = 0.6$ and $\alpha = 0.4$ in Bride et al. (2015). We use these three combinations of the weight for the weighted additive model. As we are using a pre-trained vector representation, we do not fine-tune the weights but use the existing values instead.

4.2 Multiplicative model

We also use the simple multiplicative model, also used in(Mitchell and Lapata, 2010). This is a compositional function that uses a point-wise multiplication (Eq. 5).

$$\mathbf{p}_i = \mathbf{u}_i \cdot \mathbf{v}_i \tag{5}$$

We do not use the tensor product in this study, given the pre-trained size of the vector representation.

5 Data and experiment

5.1 Target words and phrase

Instead of using different nouns for each adjective, we set what we call as a "constant noun". This combines with any types of adjectives that we test. We chose the noun, "map', as the constant noun.

The target adjectives were collected from theoretical linguistics papers that contain examples of different kinds of adjectives (Partee, 2009, 2010), and from computational linguistics papers that categorized (Nayak et al., 2014) and tested (Boleda et al., 2013) these adjectives for model evaluation. We also referred to the list of adjectives in Aparicio et al. (2016) for subsective adjectives.

Category	Phrases
Intersective	acrylic map, elliptical map, Nordic map, rectangular map, scarlet map, Hungarian map, porcelain map, metallic map
Subsective	crumpled map, bent map, soaked map, bumpy map, spotted map, fluffy map, striped map, curved map
Plain non-subsective	assumed map, debatable map, disputed map, predicted map, doubtful map, probable map, plausible map, questionable map
Privative	spurious map, forged map, counterfeit map, fictitious map, mythical map, phony map, hypothetical map, imaginary map

Table 1: The selected phrases ('adjective + *map*') (8 adjectives for each category)

Category	Mean	Std
Intersective	3.96	1.40
Subsective	3.65	1.47
Plain non-subsective	3.68	1.05
Privative	4.18	1.01

Table 2: Human rating result to the question: "How much property of "NOUN" do you think "ADJECTIVE NOUN" has?" Data collected from 18 native English speakers. Std = standard deviation.

We made sure that the adjectives we use meet the following two criteria. First, the adjective should be compatible with the constant "map" and the phrase 'adjective + *map*'. By "compatible', the 'adjective + *map*' combination (a) has a definable meaning (e.g., types of phrases to exclude: "shy table", "believed keyboard"), and (a) does not have an idiomatic meaning (e.g., types of phrases to exclude: "big brother", "long run", "hot potato"). We allowed phrases that do not exist in the real world but have meanings (e.g., "white strawberry", "purple whale"). We chose 32 adjectives (= 8 adjectives * 4 categories) from the existing words and combined them with "map". We informally asked four native speakers of English to evaluate whether the candidate phrases meet this criterion. As some adjectives were incompatible with "map", we used new adjectives that were not used in earlier studies (see Appendix for the source of the adjectives.)

Secondly, we tried to balance the word frequency (intersective: *mean* = 2575.75, *SD* = 874.20; subsective: *mean* = 2576.75, *SD* = 845.01; plain non-subsective: *mean* = 4360, *SD* = 2918.79; privative: *mean* = 3266.5, *SD* = 2426.25). The frequency information comes from the Corpus of Contemporary American English (COCA) corpus. The apparent large discrepancy between the categories was inevitable as we had to use adjectives that are compatible with the noun "map". The final list of adjectives used in our experiment is shown in Table 1.

5.2 Behavioral data

We collected human judgments on the semantic property rating task on the 'adjective + noun' phrase. The judgment task was conducted on 9 volunteers, who are self-reported native speakers of English. The participants were instructed to rate how much of the 'noun' property of the given 'adjective + noun' phrase has: "How much property of 'noun' do you think 'adjective + noun' has?" From a range of 1 to 5, score 1 indicated "does not have the property at all" and 5 indicated "has all the property".

Table 2 summarizes the result of the human rating task. The result shows that the privative category has the highest mean with the smallest standard deviation. The subsective category has the lowest mean and has the largest standard deviation. As shown in Figure 1, the subsective category has the largest variation in the range. In an informal survey after the rating task, some participants responded that they gave low score when the composed phrase ('adjective + noun') no longer had the function of the 'noun'. For example, when they were to rate the phrase, 'spotted map' or 'striped map', where both of the adjectives are subsective adjectives, they considered the phrases to no longer have the function of a map as they are 'spotted' and 'striped'. These participants gave low score in such cases. This may explain the comparatively large deviation of the rating in the subsective category compared to other types of category.

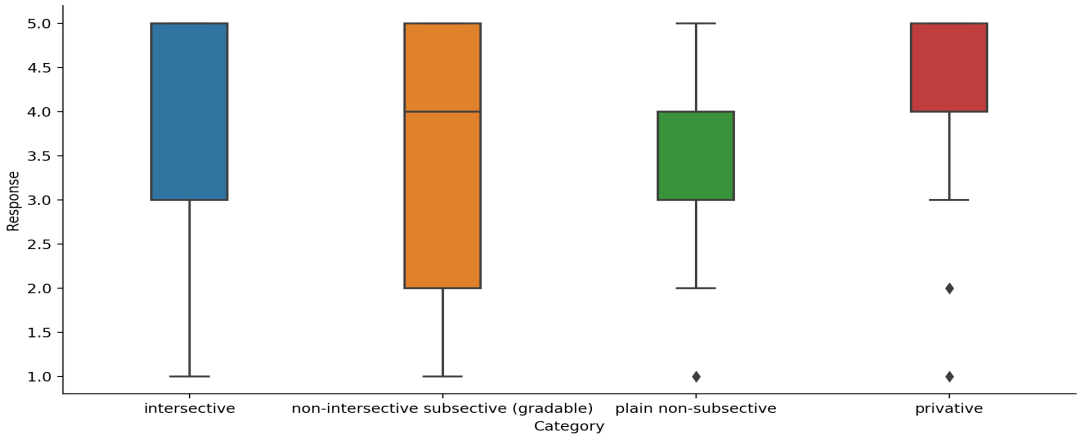

Figure 1: Semantic property rating result

5.3 Results

In order to evaluate the compositionality function of the existing model, we compute (a) the cosine similarity of the composed phrase and the noun, and (b) the correlation of the cosine similarity and the human rating result. First, we calculate the cosine similarity of the two vectors (Eq. 6). The two vectors in our case correspond to (a) the composed phrase ('adjective + noun') according to the compositional function, and (b) the noun, "map".

$$cos(\theta) = \frac{A \cdot B}{\| A \| \| B \|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}} \quad (6)$$

Next, following previous study on model evaluation (Mitchell and Lapata, 2008, 2010; Fyshe et al., 2015, a.o) we obtain the correlation of the calculated cosine similarity and the mean of human rating response. Table 3 presents the correlation coefficients of the calculated cosine similarity and the human rating score between different models. [1]

We summarize the key observations. First, the performance of the models were comparatively poor on the intersective (ρ = 0.10) and the subsective categories (ρ = 0.22) than on the plain non-subsective (ρ = 0.92) and the privative (ρ = 0.73) ones. Even the same model had different performance depending on the type of the adjective which combines with the noun (e.g. Pavlick and Callison-Burch, 2016). And our results show that there is a similar trend between the intersective and the subsective category; there is a similar trend between the plain non-subsective and the privative category.

The second observation pertains to the (out-)performance of the weighted additive models compared to the multiplicative model. This is similar to the results in Bride et al. (2015) and Vecchi et al. (2017), and in Mitchell and Lapata (2010) with the LDA topic model, where the weighted additive model showed better performance than the multiplicative model.

Thirdly, the weighted additive model produced different results depending on the weight assigned on the adjective and the noun. In specific, having more weight either on the adjective or on the noun ($\alpha = 0.3$, $\beta = 0.7$; $\alpha = 0.88$; $\beta = 0.12$) showed better performance than having equal ($\alpha = 0.5$; $\beta = 0.5$) or similar weight ($\alpha = 0.6$; $\beta = 0.4$) on the adjective or the noun. This contrasts with some of the previous studies as they (a) did not consider the different types of the adjective category, and (b) used a fixed fine-tuned value for the weight on the adjective and the noun across the adjectives (Bride et al. 2015; Vecchi et al. 2017; Mitchell and Lapata 2010; but see Guevara 2010; Baroni and Zamparelli 2010; Pavlick and Callison-Burch 2016).

Furthermore, the type of the adjective category showed an interaction with the type of the model. While the model with greater weight on the noun resulted in a better performance on the intersective and the subsective categories, the reverse–greater weight on the adjective–had better performance on the plain non-subsective and the privative categories. This shows that model performance differ depending on the type of the adjective.

[1] The coefficient efficient of the previous model (Winter, 2020) can be found in Appendix.

Category	Model	Correlation coeff.
Intersective	Simple additive ($\alpha = 0.5$, $\beta = 0.5$)	-0.17
	Weighted additive ($\alpha = 0.3$, $\beta = 0.7$; Vecchi et al. 2017)	**0.10**
	Weighted additive ($\alpha = 0.6$, $\beta = 0.4$; Bride et al. 2015)	-0.31
	Weighted additive ($\alpha = 0.88$, $\beta = 0.12$; Mitchell & Lapata 2010)	-0.60
	Multiplicative	-0.14
Subsective	Simple additive ($\alpha = 0.5$, $\beta = 0.5$)	0.20
	Weighted additive ($\alpha = 0.3$, $\beta = 0.7$; Vecchi et al. 2017)	**0.22**
	Weighted additive ($\alpha = 0.6$, $\beta = 0.4$; Bride et al. 2015)	0.01
	Weighted additive ($\alpha = 0.88$, $\beta = 0.12$; Mitchell & Lapata 2010)	-0.19
	Multiplicative	0.20
Plain non-subsective	Simple additive ($\alpha = 0.5$, $\beta = 0.5$)	0.71
	Weighted additive ($\alpha = 0.3$, $\beta = 0.7$; Vecchi et al. 2017)	0.53
	Weighted additive ($\alpha = 0.6$, $\beta = 0.4$; Bride et al. 2015)	0.77
	Weighted additive ($\alpha = 0.88$, $\beta = 0.12$; Mitchell & Lapata 2010)	**0.92**
	Multiplicative	-0.22
Privative	Simple additive ($\alpha = 0.5$, $\beta = 0.5$)	0.49
	Weighted additive ($\alpha = 0.3$, $\beta = 0.7$; Vecchi et al. 2017)	0.18
	Weighted additive ($\alpha = 0.6$, $\beta = 0.4$; Bride et al. 2015)	0.57
	Weighted additive ($\alpha = 0.88$, $\beta = 0.12$; Mitchell & Lapata 2010)	**0.73**
	Multiplicative	0.26

Table 3: Model performance evaluation by comparing cosine similarity to human rating score. Spearman's correlation of the cosine similarity to the human rating score data.

5.4 Discussion

In this paper we evaluated compositionality in the existing models, particularly including the (baseline) simple additive model, three variations of weighted additive models, and the multiplicative model. One of the observations to note is that the multiplicative model did not perform well than some weighted additive models. This may be due to the characteristics of the multiplicative model, where composition of two elements only affect the magnitude of the phrase p. As Mitchell and Lapata (2010) also speculate, the weighted additive models reflect "the relative magnitude of u [adjective] and v [noun]" (p. 1404), by which the effect of "both the magnitude and direction of p" (p. 1404) can be taken into consideration. The fact that weighted additive models reflect direction as well as magnitude, contrary to the multiplicative model, matters especially when we use the cosine similarity as the measurement. As direction rather than magnitude is a meaningful factor in calculating the cosine similarity, it is understandable that the weighted additive model, when defined with the "right" weight, shows better performance than the multiplicative model. Yet, we do note that it is of further question whether the performance of the model would change when a different method of measurement is used, as the cosine similarity was the only metric we used for the measurement.

Another crucial observation is the interaction of the adjective category and the models. In specific, why does a model with more weight on the adjective have a better performance with the plain non-subsective and the privative category? And why does a model with more weight on the noun have a better performance with the intersective and the subsective category? We explain that the interaction comes from the feature of the type of the adjective (e.g., Pavlick and Callison-Burch, 2016; Boleda et al., 2013). For example, Pavlick and Callison-Burch (2016) show that the insertion or the deletion of the adjective in a phrase significantly affects the entailment judgment. In their human judgment task, the participants responded that if something is a "rectangular map" (intersective adjective) then that is a "map". Meanwhile, the participants responded that if something is a "fake diamond" (privative adjective) then it is mostly unknown or it entails that that is a "diamond". This suggests that the property of a noun in a phrase is highly dependent on the type of the adjective that the noun is composed with. it further implies that

property of the phrase is dependent on the adjective, especially with the privative category than with the intersective category.

This can partly explain why a model with higher weight on the adjective than on the noun rendered better performance with the privative and the plain non-subsective category. In other words, as for the privative and the plain non-subsective category, the meaning of the adjective is crucial than the noun in determining the property of the phrase. We assume that it would be the reverse case with the intersective and the subsective category.

6 Conclusion

We tested how well the existing compositionality functions capture compositionlaity in natural language, particularly with the 'adjective + noun' phrases. Different from previous work, we used four types of existing categorization of adjectives, and used a constant noun ('map') that uniformly combines with different types of adjectives. We explored mainly two compositionality functions, the weighted additive model and the multiplicative model. The results showed that the weighted additive model performs better than the multiplicative model in general. More importantly, the results showed that the model perform differently depending on the type of the adjectives. This possibly serves as evidence that the compositionality functions may be tuned differently by the types of adjectives, or each adjective. Following this finding, we plan to explore the lexical function approach (Guevara, 2010; Baroni and Zamparelli, 2010; Bride et al., 2015; Vecchi et al., 2017) in the future, as this method assumes different matrices for each adjective. We can also extend the work of exploring the compositionality of the four types of adjectives by using contextualized word embeddings (e.g. Shwartz and Dagan, 2019).

References

Helena Aparicio, Ming Xiang, and Chris Kennedy. 2016. Processing gradable adjectives in context: A visual world study. In *Semantics and Linguistic Theory*, volume 25, pages 413–432.

Nicholas Asher, Tim Van de Cruys, Antoine Bride, and Márta Abrusán. 2016. Integrating type theory and distributional semantics: A case study on adjective–noun compositions. *Computational Linguistics*, 42(4):703–725.

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1193. Association for Computational Linguistics.

Gemma Boleda, Marco Baroni, Louise McNally, et al. 2013. Intensionality was only alleged: On adjective-noun composition in distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013): long papers; 2013 Mar 20-22; Postdam, Germany. Stroudsburg (USA): Association for Computational Linguistics (ACL); 2013. p. 35-46.* ACL (Association for Computational Linguistics).

Antoine Bride, Tim Van de Cruys, and Nicholas Asher. 2015. A generalisation of lexical functions for composition in distributional semantics.

Silvio Cordeiro, Carlos Ramisch, Marco Idiart, and Aline Villavicencio. 2016. Predicting the compositionality of nominal compounds: Giving word embeddings a hard time. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1986–1997.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alona Fyshe, Leila Wehbe, Partha Talukdar, Brian Murphy, and Tom Mitchell. 2015. A compositional and interpretable semantic space. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 32–41.

Emiliano Guevara. 2010. A regression model of adjective-noun compositionality in distributional semantics. In *Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics*, pages 33–37. Association for Computational Linguistics.

Matthias Hartung, Fabian Kaupmann, Soufian Jebbara, and Philipp Cimiano. 2017. Learning compositionality functions on word embeddings for modelling attribute meaning in adjective-noun phrases. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 54–64.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

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

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1429.

Neha Nayak, Mark Kowarsky, Gabor Angeli, and Christopher D Manning. 2014. A dictionary of non-subsective adjectives. Technical report, Technical Report CSTR 2014-04, Department of Computer Science, Stanford University.

Barbara Partee. 2009. Formal semantics, lexical semantics and compositionality: The puzzle of privative adjectives. *Philologia*, 7:11–24.

Barbara H Partee. 2010. 10: Privative adjectives: Subsective plus coercion. In *Presuppositions and discourse: Essays offered to Hans Kamp*, pages 273–285. Brill.

Ellie Pavlick and Chris Callison-Burch. 2016. So-called non-subsective adjectives. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 114–119.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

James Pustejovsky. 2013. Inference patterns with intensional adjectives. In *Proceedings of the 9th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, pages 85–89.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2015. A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 977–983.

Vered Shwartz and Ido Dagan. 2019. Still a pain in the neck: Evaluating text representations on lexical composition. *Transactions of the Association for Computational Linguistics*, 7:403–419.

Eva M Vecchi, Marco Marelli, Roberto Zamparelli, and Marco Baroni. 2017. Spicy adjectives and nominal donkeys: Capturing semantic deviance using compositionality in distributional spaces. *Cognitive Science*, 41(1):102–136.

Category	Word	Source	Frequency
Intersective	acrylic	SK	1479
Intersective	elliptical	SK	1605
Intersective	Nordic	SK	1673
Intersective	rectangular	P10	2800
Intersective	scarlet	SK	2896
Intersective	Hungarian	SK	3069
Intersective	porcelain	SK	3291
Intersective	metallic	SK	3793
Subsective	crumpled	SK	1465
Subsective	bent	A16	1927
Subsective	soaked	SK	2029
Subsective	bumpy	A16	2074
Subsective	spotted	A16	2530
Subsective	fluffy	SK	3331
Subsective	striped	A16	3541
Subsective	curved	A16	3711
Plain non-subsective	assumed	NM14	1428
Plain non-subsective	debatable	NM14	1510
Plain non-subsective	disputed	P10	2237
Plain non-subsective	predicted	P10	2513
Plain non-subsective	doubtful	P09	3932
Plain non-subsective	probable	B13	7500
Plain non-subsective	plausible	NM14	7619
Plain non-subsective	questionable	P10	8146
Privative	spurious	P10	1138
Privative	forged	SK	1173
Privative	counterfeit	P10	1490
Privative	fictitious	P10	1592
Privative	mythical	P10	2684
Privative	phony	NM14	4843
Privative	hypothetical	B13	5830
Privative	imaginary	P10	7382

Table 4: Sources and frequency of the target adjectives. Word frequency from COCA. Source abbreviation: 'SK' for the current present; 'P10' for Partee (2010); 'A16' for Aparicio et al. (2016); 'NM14' for Nayak et al. (2014), 'P09' for Partee (2009), 'B13' for Boleda et al. (2013)

Category	Model	Correlation coeff.
Intersective	Weighted additive (α = 0.3, β = 0.7)	-0.12
	Simple additive (α = 0.5, β = 0.5)	-0.20
	Weighted additive (α = 0.7, β = 0.3)	-0.29
	Multiplicative	**0.12**
Subsective	Weighted additive (α = 0.3, β = 0.7)	0.12
	Simple additive (α = 0.5, β = 0.5)	0.07
	Weighted additive (α = 0.7, β = 0.3)	0.01
	Multiplicative	**0.54**
Plain non-subsective	Weighted additive (α = 0.3, β = 0.7)	0.18
	Simple additive (α = 0.5, β = 0.5)	0.31
	Weighted additive (α = 0.7, β = 0.3)	**0.40**
	Multiplicative	-0.67
Privative	Weighted additive (α = 0.3, β = 0.7)	0.24
	Simple additive (α = 0.5, β = 0.5)	0.14
	Weighted additive (α = 0.7, β = 0.3)	0.02
	Multiplicative	**0.45**

Table 5: The result in the previous project (Winter, 2020).Model performance evaluation by comparing cosine similarity to human rating score. Pearson's correlation of the cosine similarity to the human rating score data. Different nouns were used for each adjective. 12 adjectives for each category were used.