Combining text and vision in compound semantics: Towards a cognitively plausible multimodal model

Abhijeet Gupta[†] **Fritz Günther**[*] **Ingo Plag**[†] **Laura Kallmeyer**[†] **Stefan Conrad**[†]
[†]Heinrich-Heine-Universität Düsseldorf
{abhijeet.gupta,ingo.plag,kallmeyer,stefan.conrad}@uni-duesseldorf.de
[*]University of Tübingen
fritz.guenther@uni-tuebingen.de

Abstract

In the current state-of-the art distributional semantics model of the meaning of noun-noun compounds (such as *chainsaw, butterfly, home phone*), CAOSS (Marelli et al. 2017), the semantic vectors of the individual constituents are combined, and enriched by position-specific information for each constituent in its role as either modifier or head. Most recently there have been attempts to include vision-based embeddings in these models (Günther et al., 2020b), using the linear architecture implemented in the CAOSS model. In the present paper, we extend this line of research and demonstrate that moving to non-linear models improves the results for vision while linear models are a good choice for text. Simply concatenating text and vision vectors does not currently (yet) improve the prediction of human behavioral data over models using text- and vision-based measures separately.

1 Introduction

The meaning and interpretation of noun-noun compounds, i.e. the combination of two words to form a new word (as in *chainsaw, butterfly, home phone*), is a contested area of study. In both theoretical linguistics and psycholinguistic circles one of the central questions is the contribution of the individual constituents in the construction of a compound's meaning (see, e.g., Bauer et al. 2013; Bell and Schäfer 2016; Schmidtke et al. 2018, ch. 20 for recent discussion).

Some psycho-computational approaches use distributional semantic models to produce representations of compound meanings. In the current state-of-the art model CAOSS (Marelli et al. 2017) the semantic vectors of the individual constituents are combined, and enriched by position-specific information for each constituent in its role as either modifier or head (e.g. *chain* as modifier in words like *chainsaw, chain mail, chain reaction, chainsaw, chain-smoking*)[1]. This enrichment is achieved by a linear architecture in which each constituent vector is first multiplied with a position-specific matrix before adding the two constituent representations to derive the compound representation.

Another aspect of compound meaning has only recently begun to attract attention, namely the role of visual information in creating and processing individual concepts and their combination. Research on embodied cognition revealed that concepts are not only based on linguistic experience, but are also grounded in perceptual experience (e.g. Barsalou 1999). In the field of neuro-psychological learning (e.g. Devereux et al. 2018), deep learning networks have been implemented in the learning of word meaning. Similarly, visual information should also play a major role in conceptual combination, at least for concrete concepts. The first study to show the effects of vision-based information in conceptual combination has been (Günther et al., 2020b).

In that study the authors compared two parallel implementations of the CAOSS model: one using text-based embeddings (henceforth *text embeddings*), the other picture-based semantic embeddings (henceforth *vision embeddings*). These embeddings (more specifically, the cosine similarities between the compound embeddings and their constituent embeddings) were then quite successfully used to predict behavioral data from experiments with human participants (i.e. reaction times in different experimental tasks). Importantly, considering information from vision embeddings in addition to text embeddings leads to significantly better predictions of human behavior. This work raises two important questions that merit further exploration. The first is about the modeling architecture, the second about the combination, instead of the

[1]See Mitchell and Lapata (2010) for another approach of dealing with asymmetric models of constituents and, Li et al. (2020); Köper and im Walde (2017) for other interesting and similar work on related phenomena.

comparison, of the two kinds of vector spaces.

Günther et al. (2020b) have used a linear architecture as implemented in the CAOSS model. In the present paper, we will explore whether non-linear architectures are better-suited to construct compound meaning representations. Our second aim is to test whether the combination of vision embeddings and text embeddings is a better basis for predicting human behavior rather than considering text embeddings and vision embeddings separately.

2 Method

2.1 Outline

We started out with pretrained sets of text and vision embeddings for compounds and their single components from (Günther et al., 2020b), which were kindly provided by the authors. We trained different machine learning architectures towards predicting the compound embeddings from their constituents.

2.2 Models

In our approach, we use a supervised learning task with the aim to assess whether the estimation of distributional meaning representations of noun-noun compounds (both, text and vision based) benefits from adding non-linearity to the models.

We compare two generic model architectures: A simple linear regression (LR) model predicting the compound embedding, and a feed-forward neural network (NN) model. Both types of model are built with the Keras toolkit (Chollet, 2015) with a TensorFlow back-end (Géron, 2019).

The LR model is inspired by Günther et al. (2020b), but does not use the position matrices of the CAOSS model. It has no hidden layers, thus treating all features as independent. In our experiments, we use the LR model as the baseline instead of the CAOSS model for two reasons: 1) In terms of architecture, the two models are analogous; however, 2) CAOSS does not train and test on distinct datasets, which potentially inflates the evaluation results (due to model memorization, Levy et al. 2015) [2]. The NN model, on the other hand, has 1 or more hidden layers that model non-linear relationships between the input and output, and facilitate interactive behavior between the input features. We experimented with 1-4 hidden layers, and report results up to 3 due to a decline in model performance beyond 3 hidden layers.

For both text and vision compound estimations, we employ the same set of model architectures, using text-based embeddings for the former and picture-based embeddings for the latter (Section 2.3). For each datapoint, the input is a function of the embeddings $\overrightarrow{c_1}$, $\overrightarrow{c_2}$ of the constituents of the compound, $f(\overrightarrow{c_1}, \overrightarrow{c_2})$, and the output is the embedding of the compound. f can be any operation; we experiment with concatenation, addition and multiplication.

Hyperparameters. The number of units in each hidden layer of the NN models is optimized for each model separately. We consider a step-size of 50 between a range of 250 to 750 hidden units in a hidden layer. All hidden layers use *tanh* as activation function and *tanh* or *sigmoid* as the activation function for the final output layer. To avoid over-fitting, we add a dropout layer in front of each hidden layer with a standard dropout value of 0.5 (Baldi and Sandowski, 2013). We use *mean-squared-error* as the loss function and an additional L_2 weight regularization in the range $[10^1, 10^{-3}]$ at the time of loss computation to further optimize over any parameters that might be outliers. For model optimization we experimented with SGD and Adadelta (Zeiler, 2012).

2.3 Datasets for the compound embeddings

Semantic Spaces. The 400-dimensional text and 300-dimensional vision pretrained embeddings were obtained *as-is* from Baroni et al. (2014) and Günther et al. (2020b) respectively.

Datasets [3]. The training datasets are obtained from Günther et al. (2020b). The dataset for the text models contains 5988 datapoints with 2387 unique constituents and 5988 compounds, the dataset for the vision models 1577 datapoints with 942 constituents and 578 compounds. Since we evaluate model performance on both text and vision data against human behavioural measures (Section 3), we create a test dataset where: 1) for each datapoint, the constituents have an overlap in the text and vision semantic spaces[4]; and, 2) the datapoints in the test set do not overlap with the training datasets. This dataset contains 352 datapoints with

[2] Our datasets are designed towards minimizing memorization.

[3] The datasets are publicly available at https://doi.org/10.17026/dans-xdp-3qhj.

[4] It is not necessary to also have text and vision embeddings for the compounds in the test sets since these are not required by the current evaluation, see below.

321 unique constituents and 352 compounds.

We introduce three different ways to combine the two input constituents – the modifier (M) and head (H): 1) Concatenation (Con) ($\vec{M} \oplus \vec{H}$) – allows the model to freely combine the information of the two embeddings; 2) Addition (Add) ($\vec{M} + \vec{H}$); and, 3) Multiplication (Mul) ($\vec{M} \odot \vec{H}$) – both variants make the dimension-wise correspondence between two embeddings comparatively explicit. In addition to the above datasets, we generate in parallel another set of datasets (identical to the above) where the semantic spaces have been normalized via L_2 normalization. We choose this overhead to ensure that the compound prediction models are not confounded by outlier values.

3 Evaluation

The empirical performance of all models was assessed with five behavioral data sets, consisting of participant ratings from Gagné et al. (2019), and reaction times as used by Günther et al. (2020b): 1) **rC1:** ratings as to what extent the meaning of the *first* constituent (modifier) is retained in the compound meaning; 2) **rC2:** to what extent the meaning of the *second* constituent (head) is retained in the compound meaning; and, 3) **rcmp:** to what extent the meaning of the compound is predictable from *both* constituents (i.e., compositionality ratings). 4) **TS:** timed sensibility task, in which participants have to judge whether a given compound has a meaningful interpretation (Günther et al., 2020b); and, 5) **LDT:** lexical decision task, in which participants have to judge whether a given word is a real English words or not (Balota et al., 2007).

For each of these data sets, we initially identified an optimal linear mixed-effects regression model predicting these behavioral measures from a set of control variables (constituent frequencies and family sizes, compound length and frequency) using step-wise backwards model selection. We then added to each model as additional predictors the cosine similarities between the compound embeddings produced by the model and their respective constituent embeddings. These similarities have been identified as the main predictors of human behavioral data in previous empirical studies (Günther and Marelli, 2019; Günther et al., 2020a). In a semantically transparent compound we expect the embeddings of a constituent (or of both constituents) to be more similar to the embedding of the compound than in a semantically opaque compound. For instance, we expect a low cosine similarity between *lady* and *ladybug* since meaningwise there is little of 'lady' in *ladybug*. As shown in numerous empirical studies, more compositionally-transparent compounds receive higher compositionality ratings (e.g. Gagné et al. 2019) and are processed faster (e.g. Günther et al. 2020b).

We obtained the conditional variance explained (r^2) of the mixed-effects regression models as our index of goodness-of-fit (using the R package *MuMIn*; Barton 2018). For each of the five data sets, we determined the rank order of these r^2 values for all models under evaluation, and calculated as an overall measure of a model's performance its mean rank across all five data sets.

4 Results & Discussion

Table 1 gives our main results. We start by predicting the text and the vision compound embeddings independently (columns 1-3, and 4-6, resp.). For each model: *Norm* – indicates whether the semantic space has been L_2 normalized (or not), *Input* – the type of input representation (Sec. 2.3) and *Arch* – the model architecture along with the number of hidden layers and units, if applicable (Section 2.2). For evaluation, we combine the text and vision model outputs for each datapoint in our test set in two different ways (column 7): a) **Mono** – we compute the cosine similarities between the predicted compound embedding and the constituent embeddings separately for the text embeddings and the vision embeddings (in all, 4 values); and, b) **Multi** – we compute the cosine similarities between the concatenation of the two predicted compound embeddings (text and vision) and the concatenations of the respective constituent embeddings (text and vision), i.e., we operate on multi-modal representations of compounds and constituents (in all, 2 values). Columns 8-12 give the evaluation scores as described in Sec. 3. Column 13 gives the order of the mean ranks for each text-vision model as computed on the basis of the r^2 values. Table 1 shows our top 5 *Mono* and *Multi* models from a rank-ordered list. The last line is the baseline model i.e., the LR model nearest to CAOSS (Sec. 2.2).

Two important points emerge from Table 1. First, we see that the best text models are all LR models (column 3), and that the vision models are all NN models (column 6). It appears that, in the case of a picture-based semantic space, predicting compounds effectively is a non-linear problem and

Table 1: Rank ordered list of top 5 Mono and Multi (NN) models along with the baseline model (BL). Best r^2 scores for each evaluation metric for both *Mono* and *Multi* in bold.

Text Models			Vision Models			Type	r^2					Rank
1	2	3	4	5	6	7	8	9	10	11	12	13
Norm	Input	Arch	Norm	Input	Arch	Type	TS	LDT	rC1	rC2	rcmp	
-	Add	LR	L_2	Add	NN 450-350	Mono	0.357	0.479	0.652	0.489	0.444	1
-	Add	LR	L_2	Con	NN 650-550	Mono	**0.358**	0.477	**0.654**	0.490	**0.446**	2
-	Add	LR	-	Add	NN 450	Mono	0.355	**0.483**	0.650	**0.492**	0.428	3
-	Add	LR	L_2	Con	NN 350-250	Mono	**0.358**	0.479	0.652	0.474	0.438	4
-	Con	LR	L_2	Add	NN 450-350	Mono	0.356	0.479	0.651	0.482	0.441	5
L_2	Con	LR	L_2	Con	NN 550-450-400	Multi	0.341	0.481	0.651	0.475	0.456	1704
L_2	Con	LR	L_2	Con	NN 450-350	Multi	**0.342**	0.478	0.652	0.475	0.457	1807
L_2	Con	LR	L_2	Add	NN 700-600-500	Multi	0.340	**0.485**	0.650	**0.489**	0.460	4993
L_2	Con	LR	L_2	Con	NN 350-250	Multi	0.341	0.478	0.657	0.477	0.468	6436
L_2	Con	LR	L_2	Add	NN 450-350	Multi	0.340	0.475	**0.663**	0.483	**0.477**	9302
-	Add	LR	-	Add	LR	BL	0.352	0.463	0.621	0.467	0.401	415874

should be treated as such. The vision-based space is a comparatively richer space (than text) in terms of features (Deng et al., 2009), and requires a more complex architecture for an effective treatment of compounds and constituents. The text semantic space (normalized or otherwise), on the other hand, is known to work well with straightforward inputs (Baroni et al., 2012) and to that effect our results are in line with the previous works.

Second, we see that the *Mono* models outperform the *Multi* models (column 7). In an ideal scenario, the multi-modal representations should resonate better with cognitive data as compared to those generated from individual semantic spaces. This is because language users do not primarily learn word meanings from reading texts, but by encountering new words in situations that involve and necessitate the integration of various kinds of information present. Combining vision embeddings and text embedding is thus an important step towards a more realistic model of meaning construction by language users. The worse performance of our combined embeddings does not bear this out. This may mean that the simple concatenation of text and vision features is not optimal and seems to blur information contained in the single text and vision embeddings. A more promising way to combine text and vision semantic spaces might be to encode the two into one and use the resultant multi-modal space as input for the compound prediction. Given the data we currently have, this is however difficult since the number of compounds for which we have text and vision embeddings both for constituents and compound is rather low.

Looking at the r^2 scores between *Mono* and *Multi*, none of the models outperforms the others in all criteria. However, except for (Multi - TS) all our models score considerably better than our LR baseline analogous to (Günther et al., 2020b). We see an improvement that is between the range of 0.6 to 7.6 percentage points, which is substantial for this kind of behavioral data: In the mixed-effect models for our TS and LDT data sets, most frequency effects (the most robust predictors of response times) explain between 1 and 15 percent of variance, and in the rating studies these values range between 1 and 5 percent.

5 Conclusion

Our results confirm that the modelling of compound semantics that is aimed at emulating human cognition, does indeed benefit from the use of non-linear models. While in this work the vision semantic space was the main benefactor from non-linearity, it remains to be seen if hyperparameter tuning over a broader range might also improve the contribution put forth by the text models. The natural next step in further developing such models is to give combined text and vision information at input rather than at output level and to allow the models to freely select the best features from both semantic spaces for compound prediction. This would presumably also be a step closer towards human cognition. We aim to achieve this in our ongoing experiments by either utilizing an existing multi-modal space for such modelling tasks or by encoding spaces of different modality into one.

References

Pierre Baldi and Peter Sandowski. 2013. Understanding dropout. In *Proceedings of Advances in Neural Information Processing Systems*, pages 2814–2822.

David A Balota, Melvin J Yap, Keith A Hutchison, Michael J Cortese, Brett Kessler, Bjorn Loftis, James H Neely, Douglas L Nelson, Greg B Simpson, and Rebecca Treiman. 2007. The English Lexicon Project. *Behavior Research Methods*, 39:445–459.

Marco Baroni, Raffaella Bernardi, Ngoc-Quynh Do, and Chung-chieh Shan. 2012. Entailment above the word level in distributional semantics. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 23–32. Association for Computational Linguistics.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247.

Lawrence W Barsalou. 1999. Perceptual symbol systems. *Behavioral and brain sciences*, 22(4):577–660.

Kamil Barton. 2018. *MuMIn: Multi-Model Inference*. R package version 1.40.4.

Laurie Bauer, Rochelle Lieber, and Ingo Plag. 2013. *The Oxford reference guide to English morphology*. Oxford University Press, Oxford.

Melanie J. Bell and Martin Schäfer. 2016. Modelling semantic transparency. *Morphology*, 26(2):157–199.

François Chollet. 2015. keras. `https://github.com/fchollet/keras`.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Barry J Devereux, Alex Clarke, and Lorraine K Tyler. 2018. Integrated deep visual and semantic attractor neural networks predict fmri pattern-information along the ventral object processing pathway. *Scientific reports*, 8(1):1–12.

Christina L Gagné, Thomas L Spalding, and Daniel Schmidtke. 2019. Ladec: the large database of english compounds. *Behavior Research Methods*, 51(5):2152–2179.

Aurélien Géron. 2019. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media.

Fritz Günther and Marco Marelli. 2019. Enter sandman: Compound processing and semantic transparency in a compositional perspective. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 45:1872–1882.

Fritz Günther, Marco Marelli, and Jens Bölte. 2020a. Semantic transparency effects in german compounds: A large dataset and multiple-task investigation. *Behavior Research Methods*, 52:1208—-1224.

Fritz Günther, Marco Alessandro Petilli, and Marco Marelli. 2020b. Semantic transparency is not invisibility: A computational model of perceptually-grounded conceptual combination in word processing. *Journal of Memory and Language*, 112:104104.

Maximilian Köper and Sabine Schulte im Walde. 2017. Complex verbs are different: Exploring the visual modality in multi-modal models to predict compositionality. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 200–206.

Omer Levy, Steffen Remus, Chris Biemann, and Ido Dagan. 2015. Do supervised distributional methods really learn lexical inference relations? In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 970–976.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11336–11344.

Marco Marelli, Christina L. Gagné, and Thomas L. Spalding. 2017. Compounding as Abstract Operation in Semantic Space: Investigating relational effects through a large-scale, data-driven computational model. *Cognition*, 166:207–224.

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

Daniel Schmidtke, Christina L Gagné, Victor Kuperman, Thomas L Spalding, and Benjamin V Tucker. 2018. Conceptual relations compete during auditory and visual compound word recognition. *Language, cognition and neuroscience*, 33(7):923–942.

Matthew D. Zeiler. 2012. Adadelta: An adaptive learning rate method. In *CoRR, abs/1212.5701*.