EmbodiedBERT: Cognitively Informed Metaphor Detection Incorporating Sensorimotor Information

Anonymous ACL submission

Abstract

The identification of metaphor is a crucial prerequisite for many downstream language tasks, such as sentiment analysis, opinion mining, and textual entailment. State-of-the-art systems of metaphor detection require training data annotated based on heuristic principles such as Metaphor Identification Procedure (MIP) (Pragglejaz Group, 2007) and Selection Preference Violation (SPV) (Wilks, 1975; Wilson, 2002). We propose an innovative approach that leverages the cognitive information of embodiment that can be derived from word embeddings, and explicitly models the process of *sensorimotor shedding* that has been demonstrated as essential for human metaphor processing. We showed that this cognitively motivated module is more effective and can improve the prediction of metaphoricity compared with the heuristic MIP that has been applied previously.

1 Introduction

Metaphor is a common type of figurative language that allows communicators to express novel construal (Shelley, 1890) and convey a myriad of implicit meanings (Gibbs, 2023). Effective metaphor processing is essential for natural language understanding tasks (Rai and Chakraverty, 2020), such as sentiment analysis, machine translation, and textual entailment. (Bahdanau et al., 2014; Wu et al., 2018; Poria et al., 2016). As a result, NLP researchers have focused on the computational modeling of metaphor, which typically starts with the identification of metaphors.

The state-of-the-art systems of metaphor identification typically rely on two heuristic principles: the Metaphor Identification Procedure (MIP) (Pragglejaz Group, 2007), and Selection Preference Violation (SPV) (Wilks, 1975; Wilson, 2002). MIP identifies metaphors by recognizing that a word's metaphorical meaning differs from its basic, 'more concrete', 'related to bodily action', and 'historically older' meaning. SPV detects metaphors by identifying violations of words' semantic selection preferences in context. The modeling of MIP usually begins with the extraction of basic and contextual representations of target words and then learns their general differences (Li et al., 2023a; Choi et al., 2021), while SPV focuses on the relation between target words and their contexts (Song et al., 2021). Despite their effectiveness, they neglect the cognitive characteristics of metaphor.

Embodied cognition posits that all cognitive acts, including language processing, are rooted in perception and action (Wilks, 1978; King and Gentner, 2022). Psycholinguistic evidence supports that metaphor processing is also embodied (Gibbs et al., 2004; Khatin-Zadeh, 2023), but the contribution of embodiment is dynamic. Specifically, the embodiment level of a metaphorical word often decreases compared to the word's basic meaning due to the structural mapping between the target and source concepts (Jamrozik et al., 2016)[1]. For example, in the metaphorical use of the verb 'drink' in (a), the embodied features of the action 'drink', such as 'consumed by mouth', and 'the object must be liquid' are abstracted away, unlike in its literal use in (b). This abstraction of sensorimotor information is essential for humans to derive a metaphorical sense of 'drink' (to consume a large amount quickly), especially in the early stage of a metaphor (Bowdle and Gentner, 2005).

- (a) The students drink the knowledge.
- (b) The horse drinks the water.

Therefore, we hypothesize that the explicit modeling of embodiment change (sensorimotor shedding) can enhance metaphor detection. To test this, we developed EmbodiedBERT, a metaphor identification system that explicitly models the process of *sensorimotor shedding*. Previous research has

[1] See in Appendix for a more detailed explanation of structural mapping theory

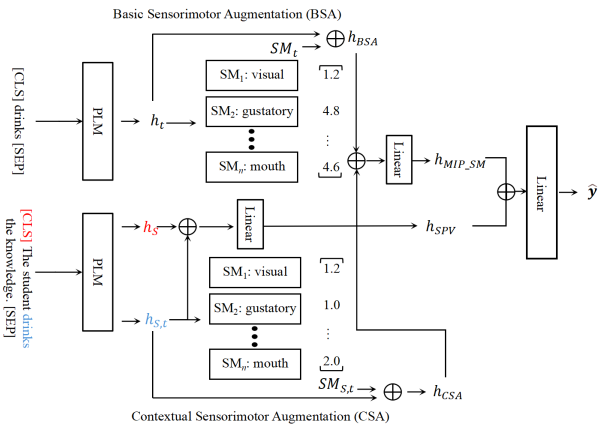

Figure 1: The architecture of EmbodiedBERT includes: two PLM encoders which generates the h_S (representation for `[CLS]`), $h_{S,t}$ (contextual representation of the target word), h_t (basic representation of the target word); ; a suite of sensorimotor regressors (n = 11) which generate $SM_{S,t}$ (sensorimotor representation of the contextual target) and SM_t (the sensorimotor representation of the basic target); a final binary classification module

integrated sensorimotor information for metaphor identification, but most of them merely use it as word-level feature enrichment without considering the change in embodiment(Bulat et al., 2017; Wan et al., 2023). Compared to general semantic change of word in context (MIP), sensorimotor change offers a more cognitively motivated and precise method for predicting metaphoricity. In our study, we show that our cognitive module is indeed effective for predicting metaphoricity.

2 EmbodiedBERT

2.1 Model architecture

EmbodiedBERT has four main components: two basic encoders for representing the target word's contextual and basic meaning; a suite of sensorimotor regressors that *maps distributional embeddings onto sensorimotor-related dimensions*; linear layers learning the function of MIP_SM (sensorimotor shedding) and SPV, and a final metaphoricity predictor.

Meaning Representation We use two `roberta-base` models (Liu et al., 2019) from Hugging Face [2] as the backbone encoder. Given a sentence $S = \{w_1, \ldots, w_n\}$, the first encoder outputs a set of contextualized embeddings $\{h_S, h_{S,1}, \ldots, h_{S,t}, \ldots, h_{S,n}\}$, where h_S stands for the global meaning of S and $h_{S,t}$ stands for the target's contextual meaning. To extract the target's basic meaning, we input the target word with special tokens into another encoder, resulting in the basic meaning embedding h_t.

The meaning representations were input into two linear functions: SPV and MIP_SM. Firstly, **SPV** aims to learn to contrasts a word's contextual meaning with the meaning of its global context. It takes the concatenation of h_S and $h_{S,t}$ and learns their difference through the linear function.

MIP_SM transforms the encoder outputs before the concatenation operation to reflect the specific change in embodiment-related dimensions. It takes an additional step to map distributional word embeddings onto these embodiment-related dimensions. Specifically, we perform such a mapping for both h_t and $h_{S,t}$, to generate SM_t (basic sensorimotor embedding) and $SM_{S,t}$ (contextual sensorimotor embedding). Next, we concatenate the derived SM_t with h_t, and $SM_{S,t}$ with $h_{S,t}$ and input them to another linear function MIP_SM. (See the next section for further details).

Binary classification Finally, the output hidden vectors from SPV and MIP_SM are concatenated together and fed into a linear layer followed by a `sigmoid` function to predict the likelihood of a target being metaphorical (Eq.1). We minimize the binary cross entropy (Eq.2) and update model parameters via back propagation.

$$\hat{y} = \sigma\left(W^T\left(h_{SPV} \bigoplus h_{MIP_{SM}}\right) + b\right) \qquad (1)$$

$$\mathcal{L} = \sum_{i=1}^{N} y_i log\widehat{y_i} + (1 - y_i)log(1 - \widehat{y_i}) \qquad (2)$$

2.2 Sensorimotor regressors

We obtained embodiment-related information as inputs for MIP_SM by mapping distributional embeddings onto sensorimotor-related embeddings (Chersoni et al., 2020). There are 11 sensorimotor dimensions related to humans' embodied experience of the physical world, including: Auditory, Gustatory, Olfactory, Visual, Tactile, Interoceptive, Hand_Arm, Foot_Leg, Head, Mouth, and Torso. A word is assigned a value for each dimension which reflects how strongly the concept embodied by the word is experienced by the respective sensor or affector (Lynott et al., 2020). We trained 11 mapping regressors that can automatically deduct these values for each word from a word's BERT embedding layer (layer 0). Each of the 11 regressors is a neural network mapping a 768-dimension embedding to a single dimension float (two fully connected hidden

[2] https://huggingface.co/FacebookAI/roberta-base

layers of the size of 384 and 192 respectively, both activated by ReLU). The training and evaluation details are in the Appendix.

3 Experiments

3.1 Dataset

We used VUA-20 (Leong et al., 2020) for model training, and VUA-18 (Leong et al., 2018) and VUA-verb for zero-shot transfer testing. Moreover, to examine our model's generalizability to non-VUA datasets, we also tested our model on MOH (Mohammad et al., 2016) and Trofi (Birke and Sarkar, 2006) in a zero-shot transfer setting [3]. For all the datasets, we adopted the existing split of train, dev, test.

3.2 Baseline models

For a thorough comparison, we selected six baseline models:

MELBERT (Choi et al., 2021) also uses `roberta-base` as basic encoder, and incorporates SPV and MIP for metaphoricity prediction. EmbodiedBERT differs from it by substituting MIP with MIP_SM.

SGNN (Wan et al., 2023) simply incorporates sensorimotor information as word-level feature enrichment. It concatenates words' GloVe embeddings and sensorimotor values from Lancaster Sensorimotor norm as input for a recurrent neural network for metaphoricity prediction.

MrBERT (Song et al., 2021) explores the relations between metaphorical verbs and their various contexts, and predicts whether relations are likely to be metaphorical.

MisNet (Zhang and Liu, 2022) implements MIP and SPV with different encoding and feature concatenation strategies.

BasicBERT (Li et al., 2023b) also proposes a new variant MIP, which can better model the meaning discrepancy between target word in context and its basic meaning. Compared with their model, EmbodiedBERT offers a cognitively motivated measure of contextual meaning change.

FrameBERT (Li et al., 2023a) also attempts to leverage external knowledge base FrameNet. It augments word embedding with self-trained FrameNet embedding for modelling MIP and SPV.

[3] Both MOH and Trofi contain exclusively verb metaphors, with the minor difference that the sentences in Trofi are generally longer than those in MOH.

	Model	Prec	Rec	F1
VUA-18	MrBERT	**82.7**	72.5	77.2
	MelBERT	80.1	76.9	78.5
	MisNet	80.4	78.4	**79.4**
	FrameBERT	**82.7**	75.3	78.8
	BasicBERT	79.5	**78.5**	79.0
	SGNN	76.7	75.5	76.1
	EmbodiedBERT	80.6	76.9	78.7
VUA-20	MrBERT	-	-	-
	MelBERT	75.9	69.0	72.3
	MistNet	-	-	-
	FrameBERT	**79.1**	67.7	73.0
	BasicBERT	73.3	**73.2**	**73.3**
	SGNN	-	-	-
	EmbodiedBERT	73.6	72.2	72.9
VUA-verb	**MrBERT**	80.8	71.5	75.9
	MelBERT	78.7	72.9	75.7
	MisNet	78.3	73.6	75.9
	FrameBERT	-	-	-
	BasicBERT	-	-	-
	SGNN	-	-	-
	EmbodiedBERT	76.3	**76.2**	**76.3**

Table 1: Evaluation of metaphor identification systems on VUA datasets. **Bold** indicates the best, underline indicates the second best

For all the baseline models, we directly obtain the performance of these baselines from the previous publications.

3.3 Implementation

We finetuned the hyperparameters with grid search. We increased our learning rate from 0 to 4e-5 during the first two epochs and gradually decreased it. We used the dropout rate of 0.2. The final model was trained with a batch size of 50 by three epochs, using Adam optimizer. We adopted precision, recall and f1-score as matrix for automatic evaluation. The final model's performance was obtained by averaging the results of five runs with random seeds. The experiments have been run on two NVIDA GeForce RTX 3090 GPUs, with a total of 48GB memory.

3.4 Results and discussion

Table 1 shows the automatic evaluation of our system compared with the baseline systems for metaphor detection in terms of precision, recall and f1 score.

VUA datasets For VUA-18, EmbodiedBERT achieves the forth best f1 score, while outperforming MelBERT, MrBERT and SGNN. For

VUA-20, our system still outperforms MelBERT, but lags behind FrameBERT and BasicBERT. In terms of VUA-verb, our system achieves 0.79%, 0.79% and 0.53% performance gains compared with MelBERT, MisNet and MrBERT. The consistent improvements over MelBERT in all three datasets show that modelling sensorimotor change (MIP_SM) is indeed effective for predicting metaphoricity, considering the major difference between EmbodiedBERT and MelBERT is the substitution of MIP by MIP_SM. Also, our system achieves the best performance in verb metaphor detection (VUA-verb), which validates that *sensorimotor shedding* is indeed an important aspect of verb metaphor processing, during which verbs are become semantically mutable to derive a metaphorical meaning (King and Gentner, 2022; Jamrozik et al., 2016).

Break-down analysis by POS When breaking down the VUA-18 dataset by Part-of-speech (POS), we find that EmbodiedBERT surpasses all other systems except MisNet in verb metaphors (f1 = 76.9, 1.3% gain over MelBERT)[4], while achieving the second best in adjective (f1 = 68.0) and third best in adverb categories (f1 = 72.5). Our system does not work well for noun metaphors (f1 = 69.5) (see details in the Appendix). It remains unclear why the modelling of sensorimotor change does not help improve the detection of noun metaphor, a category that should have also demonstrated obvious sensorimotor changes as predicted by the Structural Mapping Theory.

Break-down analysis by genre Meanwhile, when dividing the VUA-18 dataset by genre, our system outperforms all other systems in academic writings (f1 = 84.2, 0.4% gain over MelBERT) and achieves the second best in news (f1 = 78.6, 1.8% gain over MelBERT), which are both formal genres. Meanwhile, our system ranks the third best in terms of conversation (f1 = 69.9) and fiction (f1 = 75.3) (see details in the Appendix). This is relatively surprising, for we originally hypothesized that modelling sensorimotor shedding will help detect more novel metaphors which often appear in conversation and fiction, for conventional metaphors (close to literal use) are less likely to show contextual change in embodiment (Bowdle and Gentner, 2005).

Transfer to non-VUA datasets We also tested our system's transferability to non-VUA datasets, like TroFi and MOH-X, and the overall results are shown in the following table. In general, our system is relatively inferior compared with other systems in terms of zero-shot transfer ability towards non-VUA datasets (MOH-X: f1 = 78.2; TroFi: f1 = 61.5) (see details in the Appendix).

Case analysis We reveal how the integration of sensorimotor shedding can help the model reduce both false positives and false negatives. Specifically, we compared our model's predictions for VUA20-test with the predictions of MelBERT. For the reduction of false positives, EmbodiedBERT does not identify literal phrases with a minimal sensorimotor change as metaphor. For example, in the phrase 'MODERN trams, as most continental Europeans know, neither shake nor rattle, nor do they roll.', 'shake' and 'rattle' are supposed to be literal description of the tram's movement, but MELBERT predicts them to be metaphor. For the reduction of false negatives, our system is more skilled at identifying embodiment-based metaphors. For example, it can successfully identify visual metaphors like 'hazy' in 'a poet's sense of other people's very hazy', which represents cognitive incapacity by visual haziness (see more examples in the Appendix).

4 Conclusion

In this study, we contribute a novel system for metaphor detection EmbodiedBERT, which explicitly models the change in sensorimtor information of metaphorical words. The performance improvements over systems using MIP (general meaning change in context) shows that the cognitive-informed MIP_SM is indeed a promising predictor of metaphoricity. Based on our results, we envision that the incorporation of embodiment information cannot only benefit metaphor detection, but also many other language understanding tasks that require embodied experience. Therefore, a promising direction is to distill embodiment knowledge from large language models trained on multimodal inputs and apply the distilled knowledge to downstream architecture designs.

Limitations

There are some limitations to be addressed in the future research. First, the modelling of sensorimotor change highly depends on the representations of basic meaning and contextual meaning of the target

[4]Note that the verb metaphors in the break-down analysis only come from VUA-18, while VUA-verb is the mixture of data from both VUA-18 and VUA-20.

word. We currently used the output by feeding single word into encoder to represent basic meaning, but a more precise of basic meaning representation will be beneficial, which has begun to be investigated by many researchers (e.g. Li et al. (2023a), Zhang and Liu (2022)).

Second, we currently used a relatively simple method to derive contextual and basic sensorimotor representation. We envision that a more sophisticated way of integrating sensorimotor change will not only improve the performance on existing datasets, but could also be beneficial for increasing the system's transfer ability to detect novel metaphors in new datasets.

Finally, compared with BERT, recent large language models presumably contain more embodiment knowledge due to more sufficient training and more diverse inputs, which could be a more ideal source for deriving embodiment representation.

Acknowledgments

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Brian F. Bowdle and Dedre Gentner. 2005. The career of metaphor. *Psychological Review*, 112(1):193.

Ludmila Bulat, Stephen Clark, and Ekaterina Shutova. 2017. Modelling metaphor with attribute-based semantics. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 523–528, Valencia, Spain. Association for Computational Linguistics.

Emmanuele Chersoni, Rui Lin Xiang, and Chu-Ren Huang. 2020. Automatic learning of modality exclusivity norms with crosslingual word embeddings. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 32–38.

M. Choi, S. Lee, E. Choi, H. Park, J. Lee, D. Lee, and J. Lee. 2021. Melbert: Metaphor detection via contextualized late interaction using metaphorical identification theories. *arXiv preprint arXiv:2104.13615*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dedre Gentner. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2):155–170.

R. W. Jr. Gibbs, P. L. Costa Lima, and E. Francozo. 2004. Metaphor is grounded in embodied experience. *Journal of Pragmatics*, 36(7):1189–1210.

Raymond W. Jr. Gibbs. 2023. Pragmatic complexity in metaphor interpretation. *Cognition*, 237:Article 105455.

Anja Jamrozik, Marguerite McQuire, Eileen R. Cardillo, and Anjan Chatterjee. 2016. Metaphor: Bridging embodiment to abstraction. *Psychonomic Bulletin & Review*, 23(4):1080–1089.

Omid Khatin-Zadeh. 2023. Embodied metaphor processing: A study of the priming impact of congruent and opposite gestural representations of metaphor schema on metaphor comprehension. *Metaphor and Symbol*, 38(1):70–80.

David King and Dedre Gentner. 2022. Verb metaphoric extension under semantic strain. *Cognitive Science*, 46(5):e13141.

Chee Wee Leong, Beata Beigman Klebanov, Chris Hamill, Egon Stemle, Rutuja Ubale, and Xianyang Chen. 2020. A report on the 2020 vua and toefl metaphor detection shared task. In *The Second Workshop on Figurative Language Processing (FigLang@ACL 2020)*, pages 18–29.

Chee Wee Leong, Beata Beigman Klebanov, and Ekaterina Shutova. 2018. A report on the 2018 vua metaphor detection shared task. In *The Workshop on Figurative Language Processing (FigLang@NAACL-HLT)*, pages 56–66.

Y. Li, S. Wang, C. Lin, and F. Guerin. 2023a. Metaphor detection via explicit basic meanings modelling. In *The 61st Annual Meeting of the Association for Computational Linguistics*.

Y. Li, S. Wang, C. Lin, F. Guerin, and L. Barrault. 2023b. Framebert: Conceptual metaphor detection with frame embedding learning. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1558–1563, Dubrovnik, Croatia. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Dermot Lynott, Louise Connell, Marc Brysbaert, James Brand, and James Carney. 2020. The lancaster sensorimotor norms: multidimensional measures of perceptual and action strength for 40,000 english words. *Behavior Research Methods*, 52:1271–1291.

Saif Mohammad, Ekaterina Shutova, and Peter D. Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Joint Conference on Lexical and Computational Semantics (*SEM@ACL)*.

Soujanya Poria, Erik Cambria, and Alexander Gelbukh. 2016. Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems*, 108:42–49.

Pragglejaz Group. 2007. Mip: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, 22(1):1–39.

Sunny Rai and Shampa Chakraverty. 2020. A survey on computational metaphor processing. *ACM Computing Surveys*, 53(2).

Percy Shelley. 1890. *A Defense of Poetry*. Ginn, Boston.

W. Song, S. Zhou, R. Fu, T. Liu, and L. Liu. 2021. Verb metaphor detection via contextual relation learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4240–4251.

Chuandong Su, Fumiyo Fukumoto, Xiaoxi Huang, Jiyi Li, Rongbo Wang, and Zhiqun Chen. 2020. Deepmet: A reading comprehension paradigm for token-level metaphor detection. In *The Second Workshop on Figurative Language Processing (FigLang@ACL)*, pages 30–39.

Mingyu Wan, Qi Su, Kathleen Ahrens, and Chu-Ren Huang. 2023. Perceptional and actional enrichment for metaphor detection with sensorimotor norms. *Natural Language Engineering*, page 1–29.

Yorick Wilks. 1975. A preferential, pattern-seeking, semantics for natural language inference. *Artificial Intelligence*, 6(1):53–74.

Yorick Wilks. 1978. Making preferences more active. *Artificial Intelligence*, 11(3):197–223.

Margaret Wilson. 2002. Six views of embodied cognition. *Psychonomic Bulletin Review*, 9:625–636.

C. Wu, F. Wu, Y. Chen, S. Wu, Z. Yuan, and Y. Huang. 2018. Neural metaphor detecting with cnn-lstm model. In *Proceedings of the Workshop on Figurative Language Processing*, pages 110–114.

Shenglong Zhang and Ying Liu. 2022. Metaphor detection via linguistics enhanced Siamese network. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4149–4159, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

A Appendix

A.1 Structural Mapping Theory

Structure mapping theory (Gentner, 1983) aims to offer a general way of accounting for conceptual analogy, of which metaphor is a specific category. It proposes that any kind of analogy involves two processing stages: structural alignment and projection. To process analogies, human begin to take two entities in an analogy into comparison and structurally align their corresponding properties. The alignment process observes three principles: one-to-one mapping, parallel connectivity, and systematicity. Sensorimotor features of the source concept which fail to connect to the aligned system due to the violation of the principles will be shed away from source representation, and thus cannot be projected to the target representation.

A.2 Text representation

We used the byte-pair encoding (BPE) to tokenize S. Following Choi et al. (2021), we used the position embedding to distinguish target word and its local context. Also, following Su et al. (2020), after adding special tokens `[CLS]` and `[SEP]` to the beginning and the end of S, we utilized the part of speech (POS) information of the target word by appending its POS after `[SEP]`. Finally, we fed the element-wise addition of BPE token embedding, position embedding and segment embedding of S as input into the first encoder.

A.3 Training and evaluation of sensorimotor regressors

To train the regressors, we adopted Lancaster Sensorimotor Norm (Lynott et al., 2020), which contains contains 11-dimension sensorimotor information for 2,9000 English words. Meanwhile, we used word embedding from BERT embedding layer as input (Devlin et al., 2019). The size of overlapping vocabulary of Lancaster Sensorimotor Norm and BERT vocabulary is 11,402, and we split it into training and testing with the ratio of 8:2. We used mean squared error as criterion for calculating loss and adopted Adam optimizer for parameter updating. Our initial learning rate was 0.001 and was gradually decreased by the factor of 0.1 with the patience of 10. We performed 5-fold cross-validation and used early stopping to save the best model based on the loss on validation set. For evaluation, we used Pearson correlations of models' predicted values with human rating. Overall, the relatively high correlations suggest that our regressors can reliably deduct sensorimotor information from word embeddings.

Dimension	BERT
auditory	0.76
gustatory	0.78
haptic	0.79
interoceptive	0.81
olfactory	0.75
visual	0.72
foot_leg	0.74
hand_arm	0.73
head	0.61
mouth	0.73
torso	0.69
by-word	0.88

Table 2: Correlations of sensorimotor prediction with human judgement. **Bold** indicates the best, underline indicates the second best.

Dataset	Model	F1	Prec	Rec
TroFi	MelBERT	62.0	53.4	74.1
	MrBERT	72.9	**73.9**	72.1
	MisNet	-	-	-
	FrameBERT	**74.2**	70.7	**78.2**
	EmbodiedBERT	61.5	52.5	74.6
MOH-X	MelBERT	79.2	79.3	79.7
	MrBERT	**84.2**	84.1	**85.6**
	MisNet	83.4	**84.2**	84.0
	FrameBERT	83.8	83.2	84.4
	EmbodiedBERT	78.2	76.4	81.1

Table 3: Zero-shot transfer to non-VUA datasets. **Bold** indicates the best, underline indicates the second best.

POS	Model	F1	Prec	Rec
ADJ	MelBERT	64.4	69.4	60.1
	MisNet	67.0	68.8	65.2
	SGNN	**70.8**	-	-
	EmbodiedBERT	68.0	**70.9**	**65.3**
ADV	**MelBERT**	**74.6**	**80.2**	69.7
	MisNet	73.3	76.4	**70.5**
	SGNN	65.4	-	-
	EmbodiedBERT	72.5	79.5	66.6
NOUN	**MelBERT**	70.7	75.4	66.5
	MisNet	70.6	74.4	**67.2**
	SGNN	**73.2**	-	-
	EmbodiedBERT	69.5	**76.0**	64.0
VERB	MelBERT	75.1	74.2	75.9
	MisNet	**77.6**	**77.5**	**77.6**
	SGNN	76.2	-	-
	EmbodiedBERT	76.5	76.1	76.9

Table 4: POS-specific evaluation of different systems. **Bold** indicates the best, underline indicates the second best.

Genre	Model	F1	Prec	Rec
Acad	MelBERT	83.9	85.3	**82.5**
	MisNet	83.8	85.1	**82.5**
	SGNN	76.5	-	-
	EmbodiedBERT	**84.2**	**86.8**	81.7
Conv	MelBERT	70.9	70.1	71.7
	MisNet	**71.9**	**71.8**	**72.0**
	SGNN	65.5	-	-
	EmbodiedBERT	69.9	70.3	69.5
Fict	MelBERT	75.4	74.0	76.8
	MisNet	**76.0**	**74.5**	**77.5**
	SGNN	69.0	-	-
	EmbodiedBERT	75.3	73.7	77.0
News	MelBERT	77.2	81	73.7
	MisNet	**79.7**	82.6	**77.0**
	SGNN	74.4	-	-
	EmbodiedX	78.6	**83.0**	74.7

Table 5: Genre-specific evaluation of different systems. Acad: academic; Conv: conversation; Fict: fiction. **Bold** indicates the best, underline indicates the second best.

Sentence	**True**	**EB**	**MB**
This violent event, described at **length** in hysterically colourful terms, is the only piece of history to be woven convincingly into the plot.	0	0	1
Hardly a page goes by without the hapless Francis noticing something which **reminds** him, improbably, of something else.	0	0	1
There are strict time **limits**: generally, six years from when damage first occurred...	0	0	1
A solicitor **fails** to draw up a will within a reasonable time for a client who subsequently dies.	0	0	1
Children still would not have full political **status**.	0	0	1
That, says Mr Tyson, has been their only **blessing**.	1	1	0
But '**posturing** and pretending' went far beyond the unions.	1	1	0
But the chief result of all this farming was to produce huge food **mountains** which we could then refuse to give to the Third World	1	1	0
Nowadays, we all **swoon** with pleasure at the sight of a cow.	1	1	0
Though individuals are nailed, the greatest **villain** of all is the system.	1	1	0
Berry's songs are plausible **emblems** of rock'n'roll rebellion or, at any rate, youthful hedonism.	1	1	0

Table 6: Case analysis: reduction of false positives and negatives by EmbodiedBERT (EB) compared with MelBERT (MB). **Bold** indicates the target word.