



ARTICLE

Perceptual and actional enrichment for metaphor detection with sensorimotor norms

Mingyu Wan¹ , Qi Su², Kathleen Ahrens³ and Chu-Ren Huang⁴ 

¹School of Continuing Education, Hong Kong Baptist University, Hong Kong, China, ²School of Foreign Languages, Peking University, Beijing, China, ³Department of English and Communication, The Hong Kong Polytechnic University, Hong Kong, China, and ⁴Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong, China

Corresponding author: Chu-Ren Huang; Email: churen.huang@polyu.edu.hk

(Received 22 November 2021; revised 1 August 2023; accepted 2 August 2023)

Abstract

Understanding the nature of meaning and its extensions (with metaphor as one typical kind) has been one core issue in figurative language study since Aristotle's time. This research takes a computational cognitive perspective to model metaphor based on the assumption that meaning is perceptual, embodied, and encyclopedic. We model word meaning representation for metaphor detection with embodiment information obtained from behavioral experiments. Our work is the first attempt to incorporate sensorimotor knowledge into neural networks for metaphor detection, and demonstrates superiority, consistency, and interpretability compared to peer systems based on two general datasets. In addition, with cross-sectional analysis of different feature schemas, our results suggest that metaphor, as a device of cognitive conceptualization, can be 'learned' from the perceptual and actional information independent of several more explicit levels of linguistic representation. The access to such knowledge allows us to probe further into word meaning mapping tendencies relevant to our conceptualization and reaction to the physical world.

Keywords: Metaphor detection; Deep learning; Knowledge incorporation; Sense modality; Embodiment

1. Introduction

1.1. Metaphor and perception

Metaphor is one of the prominent figurative devices in our cognitive system. Research on metaphor and theory of metaphor have advanced our understanding of the conceptual systems underpinning human language and advanced findings in the fields of lexical semantics, cognitive linguistics and computational linguistics. Lakoff and Johnson (1980) describe metaphor as a cognitive mechanism (a property of language) reflected by our conceptual system for structuring our understanding of the world. Using metaphors, we can relate our known experiences to a multitude of other subjects and contexts that are more complex, implicit or less known. In general, metaphor is widely used in a language for effective communication, which usually involves a domain transfer (Ahrens and Jiang 2020), concreteness contrast (Maudslay *et al.* 2020), or semantic surprise (Zhang and Barnden 2013). The common theoretical premise is that this degree of complexity is driven by the concept of embodiment (Gibbs *et al.* 2004), which involves a mapping from a more embodied and concrete domain to a less embodied one. That is, metaphors use our shared bodily experience to describe more abstract concepts. For instance, the metaphorical expression 'apple of my eye' uses the concrete and embodied apple to describe more abstract concept of 'something to be cherished'. Other than having a solid body, embodiment is also often described by



the perception of our body, and the degree of bodily involvement of the particular perception. For instance, tactile sense requires actual contact with the body and is considered the most concrete sensory domain. Based on this interpretation, metaphorical expressions are highly reliant on bodily experiences, especially perception.

1.2. Metaphor detection

Metaphor detection can refer generally to the identification of metaphorical expressions, or more specifically to the identification of the source domain, target domain, and mapping principles of each metaphorical expression. Past studies on metaphor detection can be broadly classified as theory-oriented or processing-oriented. For the identification of metaphor, the best-known framework is the metaphor identification procedure (MIP) by the Pragglejaz Group (Pragglejaz Group 2007), which focuses on how to differentiate the metaphorical usages of a linguistic expression from other usages, including literal, ironic, metonymic, etc. Earlier theoretical works typically assume that metaphorical expressions are already identified with consensus. This accounts for why a set of criteria for identifying metaphor, that is the MIP (Steen 2010), was proposed only recently despite the long history of metaphor analysis in the literature. Several NLP studies on the identification of metaphorical expressions predates the MIP [e.g., Martin (1990); Fass (1991)]. Fass (1991), for instance, focused on differentiating metaphoric expressions from metonymic expressions, a topic that remains challenging both theoretically and computationally. On the other hand, since the emergence of the Conceptual Metaphor Theory (CMT, Lakoff and Johnson 1980), there has been continuous attention in theoretical studies on the identification of mappings between target and source domains [e.g., Gentner (1988); Gibbs (1996); Kovecses (2000)]. Computational work on the identification of mapping and interpretation soon followed [e.g., Ahrens *et al.* (2003); Veale (2003); Mason (2004)].

More recently, the processing of metaphoric expression has become one of the most important challenges in the processing of non-literal, figurative language (Veale, Shutova, and Klebanov 2016). Metaphor detection can be formalized in several ways, for example sequence-to-sequence labeling (Gao *et al.* 2018; Chen *et al.* 2021; Raval *et al.* 2021), IOB (inside-outside-beginning) tagging or sequence chunking (Bizzoni and Ghanimifard 2018; Tanasescu *et al.* 2018; Rohanian *et al.* 2020), as a paraphrasing task (Shutova 2010), and token-level binary classification (Leong, Klebanov, and Shutova 2018; Leong *et al.* 2020; Su *et al.* 2020), and so on. For replicability and evaluation, the current study adopts the model of two shared tasks on metaphor detection focusing on token-level binary classification. Take Example 1 for illustration. Given a sentence $S = w_1, \dots, w_n$ with n words and a target word $w_t \in S$, the classification task predicts the metaphoricity (i.e., metaphorical or literal) of the target word w_t . We aim at developing a metaphor detection model for a binary classification task.

Example 1.

The	Ahlbergs	have	been	accused	of	not	facing	up	to	the	harsh
w_1	w_2	w_3
realities	of	life	,	of	being	too	cozy	and	sweet	,	
...	w_t		w_n

The model returns a binary output, that is 1 if the target word w_t ('sweet' in Example 1 in S) is metaphorical or 0 otherwise. Here the target word 'sweet' is marked as a metaphorical use following the metaphor identification procedure (MIP) (Steen 2010). The word 'sweet' in its context demonstrates a meaning of pleasant experience which is part of the property of sweetness yet not denoting the gustatory meaning, where a modality shift is observed. Evaluation of the

performance of the model is then calculated by the widely adopted metrics of Precision, Recall and F1-score based on the predictions of the model by reference to the gold labels in the entire dataset.

1.3. The linguistic motivation

Metaphor is an important device for the linguistic representation of embodied cognition [e.g., Gibbs (2006); Lakoff (2012)]. Since the sensory inputs are the main sources of information of the embodied world, sensory information plays an essential role in the conceptualizing and mapping of metaphor. That is, the sense modalities (*touch, hearing, smell, taste, vision, and interoception*) as perceptual domains in our cognitive system can either serve as the source domain of metaphors (Yu 2003; Zhao 2018) or provide crucial information about the source domains. In addition to the perceptual manifestations, the action effectors (*mouth/throat, hand/arm, foot/leg, head, and torso*), serve as another fundamental way to reflect human experience and their cognition in language. The actions denoted in word concepts can affect people's way of using metaphors, a notion noted as an embodied metaphor (Gibbs *et al.* 2004; Casasanto and Gijssels 2015), as exemplified in “head_{high-embodied} of [Pembridge Investments]_{low-embodied}”. The word ‘head’ serves as the effector or the location of effectors of all the sensory modalities and is highly embodied, while its modified phrase ‘Pembridge Investments’ demonstrates no obvious sense modality or action effector. The sense modalities and their action effectors provide information about the physical world; their lexical realizations are the sources of linguistic devices utilized in figurative language. These assumptions motivate the current work to explore their interactions and hence contribute to metaphor detection.^a

1.4. Research design and objectives

We propose to incorporate the perceptual-actional information associated with words, as provided by the sensorimotor norms (Lynott *et al.* 2019) (cf. detailed introduction in Section 3), for metaphor detection. Such information, as discussed above, links each lexical concept to the physical world, and in turn plays a central role in the interpretation and modeling of metaphor. As the sensorimotor norms represent knowledge of embodied cognition, the theoretical premise of the current study is that the incorporation of direct information about the embodied physical world facilitates metaphor detection.

To utilize the ubiquitous dual-mapping anchored from the perceptual and actional experiences, we propose a series of sensorimotor-enriched machine learning models for metaphor detection based on two publicly available benchmark datasets—the VUA corpus (Steen 2010) and the TOEFL corpus (Klebanov, Leong, and Flor 2018) (cf. Section 3). A series of machine learning models are adopted to attest to the generic power of the sensorimotor-enhanced models for metaphor detection, including statistical Machine Learning models, word embeddings, and Deep Learning models. We use the sensorimotor norms for constructing a conceptual representation of the target word and its surrounding words, combining advanced NLP technologies in representing the semantic and conceptual information of words.

The sensorimotor feature of the word “reduced” in the VUA corpus is presented in a JSON line in Example 2, where *x* stores the values of the sensorimotor predictors, with a feature dictionary of 64 attribute values. As the sensorimotor norms for each word constitute 64 dimensions of features, the customized sensorimotor representation for each word in the datasets hence contains a 64-dimension vector of perception-action ratings [cf. the original data in Lynott *et al.* (2019)].

^aEven though the sensorimotor norms data is originally collected based on special knowledge, it has now been widely replicated and scaled using automatic norms prediction models [e.g., Chersoni *et al.* (2020) made it a reality of framing such knowledge in a scalable way.]

Such information is used for concatenation with word embeddings for knowledge enhancement. In addition to providing a cognitively and linguistically motivated model of word representations (cf. Sections 3.1 and 4.2), we will also test the extent of how such information may enhance performance for automatic metaphor detection (cf. Section 5).

Example 2.

```
{“y”:1, “word”:“reduced”, “id”:“clp-fragment01_745_9”, “x”: < feature.dictionary >}
```

Examples of < *feature.dictionary* >:

```
{“Max_strength.action”:1.238, “N_known.perceptual”:21, . . .
```

```
“Foot_leg.Mean”:0.428, . . . “Max_strength.sensorimotor”:2.809 . . .}
```

2. Related work

Figurative, or non-literal, language has been one of the most challenging and theoretically interesting topics in NLP for the past two decades. Computational approaches to the study of metaphor can be traced back to as early as J. Martin’s 1988 thesis (Martin 1990), and Fass (1991). The 2003 ACL Workshop on the Lexicon and Figurative Language marked the start of the recent surge of NLP studies of metaphor detection with two papers dedicated to this topic: Ahrens *et al.* (2003), and Veale (2003). Later, Veale *et al.* (2016) presented a comprehensive survey of NLP studies up to the time of publication.

In general, NLP studies of metaphor detection can be categorized into three approaches, based on the primary sources of information utilized in the detection process. Note that since NLP studies often incorporate information from multiple sources, these approaches are not mutually exclusive. The papers reviewed below are classified by the core elements of their research designs, especially in terms of their innovation and contribution.

First, the linguistic information-based approach was introduced the earliest, including Wilks *et al.* (2013) and the above mentioned (Martin 1990; Fass 1991; Ahrens *et al.* 2003; Veale 2003). This approach typically relies on lexical or contextual linguistic information to identify a metaphorical usage. Ahrens *et al.* (2003) stand out in adopting an ontology-based mapping theory of metaphor, hence integrated the meta-linguistic ontological information. This approach is later elaborated with additional grammatical features: such as semantic classes (Klebanov *et al.* 2015), and constructions and frames (Hong 2016). More recent studies adopting this approach tend to also involve statistic models for the actual processing but continue to incorporate significant linguistically encoded information such as bigrams (Bizzoni and Ghanimifard 2018), and emotion (Dankers *et al.* 2019). Lastly, we also consider studies that leverage a linguistic task that presupposes shared linguistic information as an innovative extension of this approach. Shutova (2010)’s work leveraging paraphrasing is a good example.

The second is the machine learning based approach that, unlike the linguistic approach, does not rely on explicitly represented information. That is, various learning algorithms, especially the recently dominant Deep Learning methods, are applied to ‘learn’ to identify metaphors from a large training data sets. This approach typically requires pre-trained data sets, likely derived from earlier studies, but requires no external knowledge. Standard technologies adopted include statistical machine learning, deep neural networks, transformer-based pre-trained models, etc. Typical statistical models include Naive Bayes, Support Vector Machine and Decision Trees. Deep neural networks use many layers of nodes to derive high-level functions from input information, such as CNN, RNN and LSTM. A transformer model (such as BERT) is also a neural network that can

learn context and, thus, meaning by tracking relationships in sequential data (i.e., words in a sentence) by applying an evolving set of mathematical techniques, called attention or self-attention. Transformers are among the newest, and one of the most powerful classes, of models invented to date. More details are given below, with the caveat that machine learning is used in almost all current NLP studies.

Lastly, a cognition-language incorporation approach has emerged recently. Broadly motivated by the theory of embodied cognition, the approach typically integrates behavioral or neuro-cognitive data, such as visual information (Shutova, Kiela, and Maillard 2016). More specifically, this approach often leverages lexically linked perceptual or behavioural information such as sensorimotor knowledge (Barsalou 1999; Wilson 2002). The sensorimotor norms data was originally collected based on behavioral experiments thus typically smaller in scale; they can be replicated and scaled up using automatic norms prediction models [e.g., Chersoni *et al.* (2020)]. The scaling up, and cross-lingual bootstrapping makes this dataset suitable as part of the training data for deep learning approaches. The motivation to incorporate sensorimotor norms, similar to Shutova *et al.* (2016)'s extraction of visual information from text, is to model the contribution of perceptual input to interpretation. More specifically, note that current theories of metaphor generally agree that metaphors involve mapping from more embodied concepts to less embodied concepts. Hence the perceptual input of embodied information would contribute to the identification of the embodied possible sources. Recently, Kennington (2021) enriched language models with the Lancaster norms and image vectors. These current approaches tend to incorporate some aspects of the earlier two approaches. A detailed review on these works is summarized below.

2.1. Linguistic information based approach

First NLP studies on metaphor detection adopted linguistic theories guidelines for identifying and categorizing metaphors [e.g., Ahrens *et al.* (2003); Mao, Lin, and Guerin (2019); Shutova (2010); Veale (2003)]. Earlier works [e.g., Martin (1990); Fass (1991); Ahrens *et al.* (2003); Veale (2003)] applied linguistic information directly as rule-based heuristics. Per recent trends in NLP, later studies leverage linguistic information for the preparation of training data. Such studies can be considered as the pivots between the linguistic information-based approach and the machine learning based approach. Two sets of linguistically informed guidelines are mostly commonly adopted for labeling metaphorical expressions. The first one is the Metaphor Identification Procedure (MIP), proposed by the Pragglejaz Group (2007), based on the principle that a metaphor can be recognized based on semantic gaps, that is when the contextual meaning of a word is different from its literal meaning. Another guideline is the Selectional Preference Violation (SPV) principle proposed by Wilks (1975), which identifies a metaphoric expression as a target word with a meaning that violates the selectional restrictions of its neighboring words (Wilks *et al.* 2013). For example, in “Don’t twist my words”, the word ‘twist’ denotes an abstract meaning of misunderstanding someone given the context of “my words”, which is different from the basic lexical sense of bending and distorting a physical body. In addition, ‘word’ is a non-physical object, but the basic sense of ‘to twist’ requires a physical object, hence there is a violation of selectional restriction. These above are how MIP and SPV identify metaphorical expressions respectively. In principle, the two well-known procedures identify metaphors by judging the semantic gap/or violation of the target word with its context. A natural extension is the combination of both sets of principles: MIP and SPV (Zhang and Liu 2022).

Although both MIP and SPV acknowledge the mismatched meanings between the verb ‘to twist’ and the noun ‘words’, they are unable to pinpoint the metaphoric expression. For MIP, the meaning of ‘words’ in this context refers to something the subject expressed (by speaking or writing) that is different from its basic meaning of lexical/linguistic units in a dictionary. For SPV, it is clear that the collocation of the two expressions violates selectional restrictions. But there

is no objective way to determine the directionality of violation. As such, both methods require substantial human intervention, which is more vulnerable to inter-rater disagreement. Thus, NLP studies rarely rely on linguistic theory based methods only.^b

2.2. Machine learning based approach

Recent studies of metaphor detection, just like other NLP tasks, converged on the machine learning methods. These models are neural networks trained on the training data set to learn to identify metaphorical usages. Some popular deep learning models used for metaphor detection include Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and Transformer-based models like BERT (Devereux, Shutova, and Huang 2018) and RoBERTa (Liu *et al.* 2019).

A representative work by Wu *et al.* (2018) adopted both Bi-LSTM and CNN for metaphor detection. Feature-wise, Wu *et al.* (2018) employed Word2Vec (Mikolov *et al.* 2013) as basic features and test different linguistic features such as part-of-speech (POS) and word clustering information. More relevant works include Gao *et al.* (2018) who employed Bi-LSTM as an encoder using GloVe (Pennington, Socher, and Manning 2014) and ELMo (Peters *et al.* 2018) as text input representation; Brooks and Youssef (2020) built up an ensemble model of RNNs together with attention-based Bi-LSTMs for metaphor detection. Chen *et al.* (2020) adopted BERT to obtain sentence embeddings, and then applied a linear layer with softmax to each token for metaphoricity predictions. DeepMet (Su *et al.* 2020) utilized RoBERTa with various linguistic features. IlliniMet (Gong *et al.* 2020) combined RoBERTa to obtain word embeddings in concatenation with linguistic features (e.g., WordNet, VerbNet, POS, topicality, concreteness), and then fed them into a fully-connected feedforward network to make predictions. MelBert (Choi *et al.* 2021) proposed metaphor-aware late interaction over BERT, combining pre-trained contextualized models with metaphor identification theories. With strong representation power and generalization capability, such methods show leading performances in almost all fields of work in NLP.

In sum, supervised (or semi-supervised) machine learning has been proven useful for metaphor detection. Typically, metaphorical words or expressions are manually annotated by human experts, which are then used as knowledge-incorporated external resources for many metaphor detection tasks (Leong *et al.* 2018, 2020). In addition to the annotated training set, recent deep learning studies using BERT or other large neural language models require substantial computational resources for pre-training and fine-tuning. In several recent studies, people aim to adopt alternative techniques to compress BERT-based models for more reasonable production environments using methods such as weight pruning, matrix factorization, and knowledge distillation (Zafrir *et al.* 2019; Rogers, Kovaleva, and Rumshisky 2020; Zafrir *et al.* 2021).

2.3. Cognition-language incorporation approach

Metaphors, as described in the introductory section, are a cognitive mechanism to express new and novel experiences or knowledge using old and familiar experiences. As a linguistic device, its most prominent feature is the lack of any overt linguistic marking. This is why the linguistic knowledge based and machine learning based approaches, at least in the early data preparation stage, require human expertise to manually annotate the training data. Given the cost of human intervention in annotation for both linguistic and machine learning based approaches, there are some recent attempts to incorporate cognitive knowledge in metaphor identification. One of the first cognition-language incorporation resources applied was the shared conceptual knowledge of ontology, and SUMO ontology in particular (Ahrens *et al.* 2003; Huang *et al.* 2007; Dunn 2014).

^bPlease see Rai and Chakraverty (2020) for a summary of other related issues in these approaches.

The rationale is that since ontologies are shared conceptual knowledge, they are a potentially powerful tool to represent the differences between and map from source domain to the target domain. Following the same rationale, and based on the common belief that wordnets are linguistic ontologies, WordNet, FrameNet, and VerbNet were also adopted (Zhang and Barnden 2013; Jang *et al.* 2015; Klebanov, Leong, and Gutierrez 2016).

Another characteristic of metaphor that different theories generally agree upon is the embodied cognition hypothesis: that a metaphor is the use of a more embodied (i.e., concrete) concept to describe a less embodied (i.e., abstract) concept [e.g., Gibbs *et al.* (2004); Lakoff (2012)]. Since embodied concepts are attested by sensory inputs (Barsalou 1999), NLP studies explored the incorporation of sensory input-related resources, such as sensory lexicon, synaesthesia, and vision-based information (Shutova *et al.* 2016; Tekiroğlu, Özbal, and Strapparava 2015), property norms (Zayed, McCrae, and Buitelaar 2018), information about concreteness, imageability (Maudslay *et al.* 2020), or emotion (Rai *et al.* 2019), etc.

An emergent trend that is highly interdisciplinary is to leverage neuro-cognitive research outcomes in NLP. Although no metaphor processing study has adopted this new paradigm yet, this new trend may well be the next direction to go. The crucial breakthrough is the direct adaptation of behavioral or brain activity data, instead of the lexical information from the sensory domain as reported earlier. The main goal is to synergize neurocognitive and computational approaches for significant breakthroughs. Two recently founded/revamped workshop series spear-headed this new development: Linguistic and Neuro-Cognitive Resources (Devereux *et al.* 2018), and cognitive ordering and computational linguistics (Chen *et al.* 2021). The eye-tracking dataset is the earliest adopted data set to NLP, such as Long *et al.* (2019) and Barrett and Hollenstein (2020), Baroni and Lenci (2010). There was even a shared task on using NLP to predict eye-tracking results (Hollenstein *et al.* 2021). In addition, there are several attempts in CL to incorporate brain measurement data, such as fMRI and EEG, for modeling word embedding results (Chen *et al.* 2021). Continuing in this direction, our current study proposes to utilize sensory domain behavioral data for automatic metaphor detection.

2.4. Innovation of our work

To date, metaphor detection remains a challenging task because the semantic and ontological differences between metaphorical and non-metaphorical expressions are often subtle and contextually dependent. Existing methods show different strengths for detecting metaphors, yet each has its respective disadvantages. Knowledge- or theory-based methods (Hong 2016; Mao *et al.* 2019) tend to show generalization problems that are not widely applicable in real settings. State-of-the-art ML-based NLP models (Devereux *et al.* 2018; Liu *et al.* 2019; Brooks and Youssef 2020) are less interpretive to understand the intrinsic properties of metaphors. The innovation of our work is to model word meaning representations for metaphor detection via both word embeddings and perception-action knowledge. We take the foundational perception-action knowledge from embodied cognition and combine it with word embeddings and deep neural networks, with the expectation of a robust model metaphor processing that will both improve NLP performance and inform our understanding of embodied cognition.

Among previous studies, Tekiroğlu *et al.* (2015) is the most similar to our work in that they incorporated both the sensorial features and linguistic synesthesia of the five sense modalities. Different from us, this work did not incorporate either sensory modality exclusivity or sensorimotor norm and relied mainly on information based on the Sensicon extracted from WordNet. They also focus on the adjective-noun pairs extracted from a dependency-parsed corpus (DPC). Note that at the time of their study sensory modality exclusivity was already available, although the sensorimotor norms had not been published yet. Thus our current study is distinguished from Tekiroğlu *et al.* (2015), especially in terms of incorporating the lexicon-based behavioral norms from empirical studies. In addition, following results from recent studies, we added interoceptive

as the sixth sense modality. The incorporation of the two behavioral norms is crucial as it allows empirical data of embodied cognition to play a role in the detection of metaphors. We use the complete sensorimotor ratings for each word to form a word vector space, which is used to complement word embeddings. In addition, our study probes further into the correlation of the various modalities and actions for predicting metaphors, as well as conducts cross-sectional experiments to look into a wider range of factors, for example text genre, POS, and language proficiency.

These research objectives on metaphor, as preliminarily attested by the several prototype experiments in metaphor detection competitions (Leong *et al.* 2020; Wan *et al.* 2020a, 2020b; Wan and Xing 2020), will fill a much-needed gap in linguistic and computational research on metaphor identification. Note that the current work is motivated by yet distinct from these preliminary studies. For instance, Wan *et al.* (2020a) adopted simple statistical models (e.g., Logistic Regression) and some basic linguistic features apart from the conceptual and embodiment features. In addition, no deep neural network models were attested for further comparisons. As for the paper in Wan and Xing (2020); Wan *et al.* (2020b), though they employed neural network models, the experiments were run on only one dataset-VUA, and there are no subcategory experiments to look into the variation across genre, POS and language proficiency. In addition, the current work probes further into the issue of metaphors with more in-depth linguistic introspection and case-error analysis.

3. Data

3.1. The sensorimotor norms

As observed in the VUA corpus, the prevalence of concept mapping in terms of modality senses or bodily involvement between the target word and its immediate context implies a high probability of metaphorical usage. We propose to leverage perception-action information for modeling such concept mapping to facilitate metaphor detection, as well as to probe into the mapping mechanism of finer categories.

The Lancaster Sensorimotor norms collected by Lynott *et al.* (2019) are adopted for enriching word representations in this study. The data includes a most comprehensive measure of the sensorimotor strength (0–5 scale indicating different degrees of sense modalities and action effectors) for around 40K English words across six perceptual modalities: *touch*, *hearing*, *smell*, *taste*, *vision*, and *interoception*, as well as five action effectors: *mouth/throat*, *hand/arm*, *foot/leg*, *head* (excluding mouth/throat), and *torso*. These norms represent the largest ever set of semantic norms for English, incorporating almost 40,000 words; they provide far greater lexical coverage than has been possible with previous norms, encompassing the majority of words known to an average adult speaker of English [i.e., approximating a full-size adult conceptual system; Brysbaert *et al.* (2014)]. In the two datasets of the current work, 95% of the lexical words are covered in the sensorimotor lexicon. The data has been published in a top-tier journal in psycho-linguistics (*Research Behavior Methods*) which has gone through a rigorous review on the quality and reliability of the annotation (The mean alpha across all dimensions was $\geq .8$ and each individual dimension had $\alpha \geq .7$ (i.e., very good agreement overall).)

Among the six modality senses, the *visual* sense allows us to see the external world; the *hearing* sense permits us to hear the sounds; the *gustatory* sense considers tastes; the *olfactory* sense accounts for odors; the *tactile* sense perceives temperature, pain, and textures of objects; the *interoceptive* sense detects the internally stimulated feelings of hunger, exhaustion, disgust, etc. We use our body parts to undertake and experience these perceptions, as foregrounded in Figure 1. For example, we need to move our *head* or *torso* when we view the surroundings; we will open our *mouth* and articulate with our *throat* when we talk; we use our *hands* (and arms) to reach and grasp for objects, and we also resort to our *feet* (and legs) to walk or to kick. The scale of these

Table 1. The perceptual-actual ratings of five sample words in the sensorimotor norms

Dimension	Word	<i>Adopt</i>	<i>Big</i>	<i>Daze</i>	<i>Eat</i>	<i>Learn</i>
Perception	Auditory	1.222	0.944	0.455	1.263	3.941
	Gustatory	0.056	0.167	0.000	4.526	0.765
	Haptic	1.056	2.722	0.000	2.158	1.765
	Visual	1.889	3.889	1.953	2.632	3.882
	Olfactory	0.111	0.111	0.000	2.421	0.588
	Interoceptive	1.222	0.333	3.253	2.474	1.529
Action	Foot_leg	0.650	1.000	0.350	0.050	0.810
	Hand_arm	1.650	1.900	0.850	2.900	1.381
	Head	2.050	3.100	3.200	2.000	4.476
	Mouth	1.350	1.350	0.950	4.850	2.286
	Torso	0.900	0.950	0.850	1.600	0.667

The dominant modality and action effector are highlighted in bold.

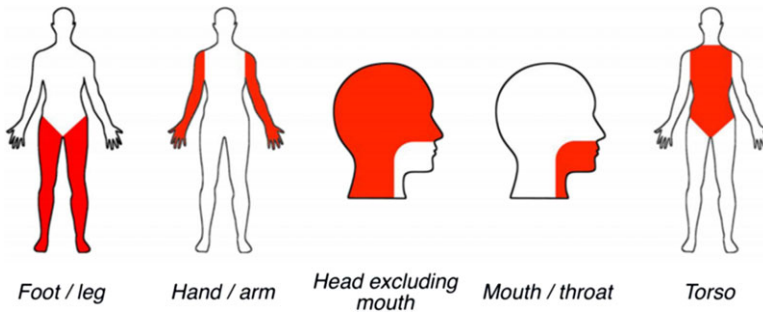


Figure 1. Avatar images describing the area of each effector during action strength norming. Figure downloaded from Lynott et al. (2019).

dimensions is judged in terms of how the participants experience these concepts through each of the perceptions and actions they perceive, on a scale from 0 = no feelings at all, to 5 = very strong feelings. In addition, the dominant effectors and exclusivity^c were likewise assigned to each word, expanding the paradigm of those in the modality exclusivity studies (Lynott and Connell 2009) with the addition of action effectors.

The augmentation and consolidation of perception and motion in this large-scale norming study compared to other relevant studies of other minority languages demonstrate a more fine-grained picture depicting how the specific sensorimotor information is encoded in the language. To demonstrate the structure of the sensorimotor norms, we provide five sample words and their six sensory scores and five action effectors, as provided by the sensorimotor norms, in Table 1 for illustration.

Table 1 and Example 2 demonstrate how words are represented and rated in terms of the perception-action dimension scale with examples of five words. The vector for each word (in each column) is composed of 64 dimensions of perception and action values, as well as some derived

^cEffector exclusivity is calculated similarly to modality exclusivity, that is a measure of the extent to which a particular concept is experienced through a single dimension (0–1, typically expressed) as in Lynott and Connell (2009).

statistics, such as standard deviation, modality/action exclusivity, dominant modality/action effector, etc. In Table 1, we only display the mean of the six perception modalities and the five action effector scores here to provide a brief overview. The perception and action effector with the highest scores (highlighted in bold) respectively mark the dominant sense modality and action effector for each word, such as ‘Visual’ and ‘Head’ for the word ‘Big’; ‘Gustatory’ and ‘Mouth’ for the word ‘Eat’. The fact that the perception and action as the two embodied bases of human cognition can be captured by these two attributes in the model suggests that it could serve as very useful resource for linguistic processing (Chersoni *et al.* 2020).

3.2. The metaphor detection datasets

For the metaphor detection experiments, we adopt two benchmark datasets—the VUA and TOEFL corpora that are commonly used by NLP system competitions on Figurative Language Processing. The series of shared tasks on metaphor detection have also adopted the two datasets for open competition (Leong *et al.* 2018, 2020). The two datasets are regarded as the most representative and large-scale data of metaphors in general with human validation of the metaphor labeling based on the MIPVU protocol. Although there are two other widely adopted datasets in the metaphor detection literature—MOH-X (Mohammad, Shutova, and Turney Peter 2016) and TroFi (Birke and Sarkar 2006), they are designed for verbal metaphors instead of all lexical categories and are much smaller in size.^d There are a total of four tracks for evaluation: VUA AllPOS, VUA Verbs, TOEFL AllPOS, and TOEFL Verbs. The AllPOS track is concerned with the detection of all content words, that is nouns, verbs, adverbs and adjectives while the Verbs track is concerned only with verbs. Function words are not considered for evaluation.

The metaphorical labels in the two datasets are prelabelled by Tekiroğlu *et al.* (2015) and Klebanov *et al.* (2018) for all lexical words (ALLPOS) that have been widely used by many previous works on metaphor detection (Leong *et al.* 2018, 2020). The current work, and many other reported systems, have evaluated on model performance by either (1) focusing on verbs only; and/or (2) on all POS labels. Many existing studies look at verbal metaphors in particular because verbs denote more metaphoric meaning, as also testified by our result in Figure 5. Our methodology is not just driven by verb-based metaphors, as our experimental results (as shown in Section 5.4.1) have also attested the possible POS effects on metaphor detection, which demonstrate a consistent improvement of adding sensorimotor information to the model for either verbs or other lexical categories, despite slight variances within each category.

All the results of the testing systems are publicly available online.^e Details of the two datasets are provided below.

3.2.1. The VUA dataset

The first dataset is the VU Amsterdam Metaphor Corpus (VUA) (Steen 2010). This corpus is a benchmark dataset released for the shared tasks of metaphor detection (Leong *et al.* 2018, 2020), which is publicly available for standard reference. It is a subcorpus of the British National Corpus with manually annotated labels indicating the metaphoricity of each token in the corpus. It consists of 115 text fragments sampled across four genres: Academic, News, Conversation, and Fiction. The text genre composition is provided in Table 2.

The data has been annotated using the MIPVU procedure (Steen 2010) with a strong inter-annotator agreement ($k > 0.8$). Examples of the sentences in the corpus are demonstrated in Figure 2. Each token in the corpus is encoded with a unique id composed of the text fragment

^dMOH-X is a verb metaphor detection dataset with the sentences from WordNet and TroFi is also a verb metaphor detection dataset, including sentences from the 1987–89 Wall Street Journal Corpus Release 1.

^e<https://competitions.codalab.org/competitions/22188>

Table 2. Text genre composition of the VUA corpus

Text genres	Sentences	Tokens	Fragments	%M_Verbs
Academic	8526	49,561	16	31
Conversation	9653	48,001	24	15
Fiction	7588	44,892	12	25
News	6698	45,116	63	42
TOTAL	32,465	187,570	115	28

“%M_Verbs” denotes the proportion of metaphorical verbs in each text genre.

```

txt_id sentence_id sentence_txt
ale-fragment01 1 Latest corporate unbundler M_reveals laid-back M_approach
ale-fragment01 2 By FRANK KANE
ale-fragment01 3 IT SEEMS that Roland Franklin , the latest unbundler to a
ale-fragment01 4 He has not properly investigated the M_target 's dining fa
ale-fragment01 5 The 63-year-old M_head of Pembridge Investments , M_throug
ale-fragment01 6 If he had M_taken his own rule seriously , he would have i
ale-fragment01 7 There are other M_things he has , M_on his own M_admissior
ale-fragment01 8 When the bid was M_launched last week , Mr Franklin M_face
ale-fragment01 9 He M_regards the M_charges as unfounded .
ale-fragment01 10 M_On property , he is M_blunt .

```

Figure 2. Sample of the annotated data in the VUA corpus.

id, sentence id and the token sequence number in each sentence. Thus, the word ‘corporate’ in the first sentence of Figure 2 has the id of ‘ale-fragment01-1-2’. Each metaphorical expression is marked by the label of ‘M_’ for distinction. Based on these gold labels, we can conduct supervised machine learning experiments.

3.2.2. The TOEFL dataset

The second dataset labeled for metaphor was sampled from the publicly available ETS Corpus of Non-Native Written English, which was first introduced by Klebanov *et al.* (2018). The annotated data comprises essay responses to eight persuasive/argumentative prompts, for three native languages of the writers (Japanese, Italian, Arabic), and for two proficiency levels—medium and high. The argumentative metaphors are annotated with average inter-annotator agreement $k = 0.56 - 0.62$ by Klebanov *et al.* (2018). We use the data partition of 180 essays as training data and 60 essays as testing data. Table 3 shows some descriptive characteristics of the two datasets: the number of texts (#text), sentences (#sentence), tokens (#token), and metaphorical proportion (%M) in the data.^f

The statistics show that verbs contain a much higher portion of metaphors than the other lexical categories as suggested by the %M. This may be due to the fact that verbs are shown to be more mutable [i.e., more likely to change meaning, see Gentner and France (1988); Ahrens (1999)]. We will revisit this issue in the discussion of results from the perspective of POS in Section 5.

^fThe test datasets were not released with annotations for public access, hence we are unable to calculate the metaphorical proportions in test sets.

Table 3. Data partition for both VUA and TOEFL datasets

Datasets	VUA		TOEFL	
	Train	Test	Train	Test
#text	90	27	180	60
#sentence	12,123	4081	2741	968
#token	72,611	22,196	26,737	9014
%M (ALLPOS)	18%	–	7%	–
%M (Verbs)	29%	–	13%	–

4. Methodology

This work proposes an innovative method for metaphor detection based on the idea that the basic modality senses (*touch, hearing, smell, taste, vision* and *interoception*) and action effectors (*mouth/throat, hand/arm, foot/leg, head, torso*) of words as indicated by the sensorimotor norms (Lynott *et al.* 2019) provide crucial information for metaphoricity inference. We utilize both feature engineering and deep neural networks for implementation,[§] as detailed in the following subsections.

4.1. Baseline methods

We adopt the following three strong baselines for peer comparisons to our proposed models.

- **B1 (Baseline 1):** The first baseline is a feature based statistical Machine Learning model proposed by Klebanov *et al.* (2014) which has been widely adopted as a common strong baseline for many metaphor detection shared tasks. Despite a simple method, it demonstrates surprisingly better performance than many advanced deep learning models. The features include lemmatized unigrams, generalized WordNet semantic classes, and differences in concreteness ratings between verbs/adjectives and nouns (UL + WordNet + CCDB). This baseline is similar to our first model as we also utilize both knowledge features and statistical ML models (*e.g.*, logistic regression). Therefore, it can serve as an effective baseline for our model comparisons.
- **B2 (Baseline 2):** The second baseline is the approach proposed by Brooks and Youssef (2020) which uses bidirectional attention mechanisms for metaphor detection. In this model, each word is represented by an 11-gram which contains the target word in the center together with five neighboring words as the context; each word in the 11-gram is represented by a 1324 dimensional word embedding (concatenation of ELMo and GloVe embeddings). Brooks and Youssef (2020) experimented with ensembles of models that implement different architectures (in terms of attention) trained on POS information. This baseline is similar to our second model which also utilizes attention based Bi-LSTM, but the difference is that we adopt Sensorimotor vector instead of ELMo. Therefore, it can serve as another effective comparison to our deep learning method.
- **B3 (Baseline 3):** The third baseline is a minimal adaptation of our second model (cf. SGNN in Section 4.2.2) that replaces the sensorimotor vector of each word with an equal

[§]Code and data available at: https://github.com/ClaraWan629/Metaphor-Detection_Journal

dimension of random vectors. This baseline is used to rule out the possibility that the effect of incorporating sensorimotor knowledge into the deep neural networks is caused by increased vector space. Hence we model a baseline architecture similar to SGNN except that randomly generated vectors are concatenated as the additional vector space to word embeddings.

4.2. Sensorimotor-enriched modeling

4.2.1. SFeature (Statistical models with sensorimotor feature)

Our first model is based on feature engineering with statistical machine learning. For comparison purposes, we modeled three other categories of features in addition to perceptual norms (i.e., sensorimotor features). These include word-ngram, lemma-ngram and POS-ngram, word embeddings, and cosine similarity between the target and its neighboring words, as well as B1 as mentioned in the above section. We use three statistical models and ensemble learning strategies during training so as to test the cross-model consistency of the various features, as detailed below:

- **Sensorimotor Feature:** We model the perceptual-actional knowledge to a word by mapping the target word to the sensorimotor norms data and acquiring the sensorimotor vector space for each word in the corpus. The acquired vector space for all the words in the corpus forms a sensorimotor feature matrix, which contains dictionaries of four key-value pairs, including the target word, its id, the feature attribute in terms of sensorimotor ratings (x), as well as the metaphoricity label of each word in the corpus (y). Such feature structure is formatted in JSON lines to work with Unix-style text processing tools and shell pipelines, as demonstrated in Example 2. Unmatched words (those not covered in the sensorimotor norms data) are assigned the average sensorimotor values for each feature dimension.
- **Collocations:** Three sets of collocational features are constructed to represent the lexical, syntactic, and grammatical information of the target nodes and their neighbors: Trigram, FL (Fivegram Lemma), FPOS (Fivegram POS tags). In the preliminary experiments, we tested on different window sizes ranging from 2 to 10 for the POS ngrams. The results show that trigrams and fivegrams produced superior performances and we focus on the two features for the collocation baseline. The two corpora are lemmatized using the NLTK WordNetLemmatizer and POS tagged using the NLTK averaged perception tagger (Loper and Bird 2002) before constructing such features.
- **Word Embeddings:** We also utilize word embeddings to capture the semantic information of words based on the distributional hypothesis on word meaning [e.g., Lenci (2008)]. Three models are used: GoogleNews.300d, Internal-W2V.300d (pre-trained using the VUA and TOEFL corpora), and the GloVe vectors. GoogleNews in this work is pre-trained using the continuous bag-of-words architecture for computing vector representations of words (Church 2017). GloVe is an unsupervised learning algorithm for obtaining vector representations for words. We use the 300d vectors pre-trained on Wikipedia 2014+Gigaword 5 (Pennington *et al.* 2014).
- **Cosine Similarity:** We also investigate the cosine similarity (CS) measures for computing word sense distances between each word and their neighboring lexical words in a given sentence. This approach can be traced back to earlier vector space neighborhood models of lexical meaning [e.g., Ploux and Victorri (1998)], as well as cosine similarity measurements of semantic relations and semantic distance [e.g., Turney and Littman (2005); Baroni and Lenci (2010)]. It has been applied to detect other non-literal meanings (Xu *et al.* 2015), as well as the differentiation of different semantic relations in the TOEFL dataset (Santus

Table 4. Parameter setting for the three statistical classifiers

Classifier	Parameter
Logistic Regression (LR)	'class weight': 'balanced', 'max iter': 5000, 'tol': 1
Linear SVC (LSVC)	'class weight': 'balanced', 'max iter': 50000, 'C': 10
Random Forest Classifier (RFC)	'min samples split': 8, 'max features': 'log2', 'oob score': 'True', 'random state': 10, 'class weight': 'balanced'

et al. 2016). Recently, it has also been applied to metaphor detection (Rai *et al.* 2018). The neighboring lexical words are syntactically close content words that occur in the same clause of the target word. The CS was computed based on the averaged cosine distance between the word embedding vectors. Three different sets of CS features are constructed in this work by using the above three different word embedding models: CS-Google, CS-GloVe, CS-Internal (word vectors trained on the VUA and TOEFL corpora).

These features provide meaning representations of the target words and their neighbors in terms of their senses modalities, action effectors, exclusivity etc., as illustrated by the various predictors of each word in Example 2 [cf. more information on the data structure in Lynott *et al.* (2019)]. Wan *et al.* (2020a), reporting our pilot study, showed that these features are highly indicative of metaphorical uses and are hence hypothesized as more distinctive features than the strong baselines.

With the above features, three traditional classifiers are used for predicting the metaphoricity of the tokens, including Logistic Regression (LR), Linear Support Vector Classification (LSVC) and a Random Forest Classifier (RFC). The Machine Learning experiments are run through utilities provided in the SciKit-Learn Laboratory (SKLL) (Pedregosa *et al.* 2011). For parameter tuning, we use grid search to find optimal parameters for the learners, as in Table 4.

4.2.2. SGNN (*Sensorimotor with Glove Neural Network*)

SGNN is the second model we propose based on the sensorimotor information and neural networks. In the SGNN model, words are processed with the integration of sensorimotor vectors and word embeddings, as depicted in Figure 3. Since we aim to compare a sensorimotor features driven system with a pre-trained method (e.g., word embedding with BERT), we do not combine these two models. BERT may show overwhelming performance given the large pretrained models and training data used, as well as the vast amounts of hyper-parameters that rely on high-capacity GPU. For most of the fine-tuning experiment in the BERT-based models, more than 16 GB of GPU memory for BERT-Large is needed. However, one major purpose of our study is to introspect metaphors in-depth (in a cost-effective way) from the perspective of the perceptual and actional features to interpret the language mechanism in modeling metaphors. We make use of these features and look into their sub-dimension effects on metaphor detection through sub-experiments and look at the variances through the model performances instead of just pursuing model enhancement.

In the SGNN model, we map the words to the sensorimotor norms and obtain the modality representations (64 dimensions for each word). At the same time, we obtain the word vectors (300 dimension for each word) using GloVe and then concatenate them as inputs to neural networks. The red boxes represent the vector of sensorimotor information for each input word; the blue boxes are the word embeddings. For those words not mapped in the sensorimotor norms, we assigned three kinds of values, including all zeros, random values following normal distribution,

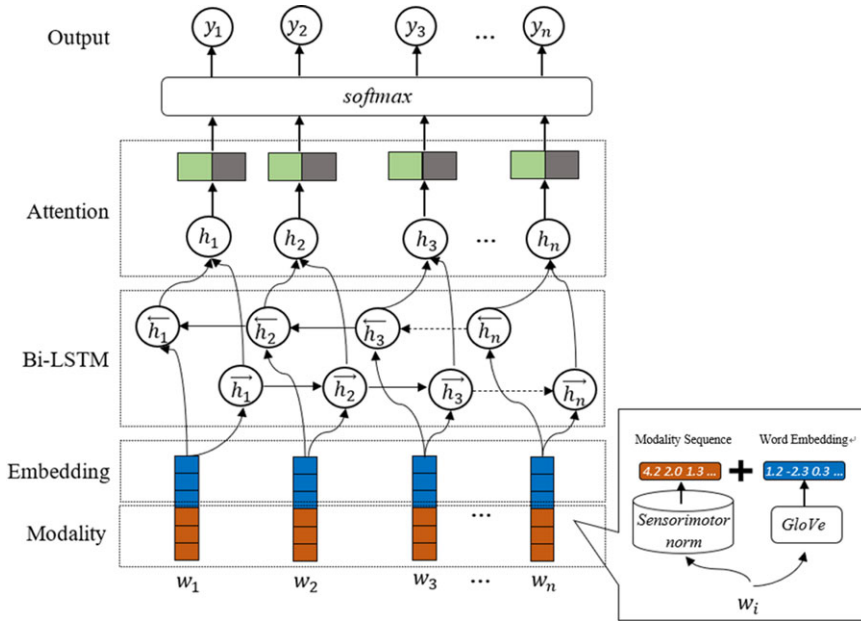


Figure 3. The architecture of the SGNN model.

as well as average scores of the sensory words in the corpus. In the end, we chose to use the average score for each dimension of the out-of-dictionary words to optimize our results.

The Bi-LSTM layer produces a hidden status for each word (w_i) in a given sentence. We use this status to calculate an attention weight which is multiplied by the output of the Bi-LSTM layer. The green box corresponds to the attention weight for each word, and the grey box represents the hidden vector. Let $H \in R^{d \times N}$ be a matrix consisting of hidden vectors $[h_1, h_2 \dots h_N]$ that is produced by LSTM, where d is the size of hidden layers and N is the length of the given sentence. The attention mechanism will generate an attention weight α . The final sentence representation is given by:

$$h = H \times \alpha^T$$

We also add a Linear layer. The final probability distribution is:

$$y = \text{softmax}(W_s h + b_s)$$

Let y be the target distribution for the sentence, b_s be the bias offset, and \hat{y} be the predicted metaphoricity distribution. We train the model to minimize the cross-entropy error between y and \hat{y} for all sentences.

$$\text{loss} = - \sum_i \sum_j y_i^j \log \hat{y}_i^j + \lambda || \theta ||^2$$

Then, we get a probability distribution of 0–1 label to train the model and get the predictions. Our evaluation adopts the commonly used metrics precision (P), recall (R), and F1-measure (F1). In addition, we use the default hyperparameters of attention-based Bi-LSTM and estimate them by using a grid search within a reasonable range. Each value of the hyperparameters is shown in Table 5.

Table 5. The hyperparameter setting of the SGNN model

Hyperparameters	Value
Input layer size	364
Number of hidden layers	1
Hidden layer size	300*2
Output layer size	1
Dropout	0.5
Loss function	NLLLoss
Optimization algorithm	Adam
Epochs	15
Batch size	512
Activation function	Softmax
Learning rate	0.01

Table 6. Feature evaluation on the VUA Verbs track

Major	Feature Secondary	Classifier		
		LR	LSVC	RFC
B1	UL + WordNet + CCDB	0.632	0.621	0.618
Collocation	Trigram	0.626	0.625	0.612
	FL	0.624	0.623	0.621
	FPOS	0.378	0.369	0.335
Word Embeddings	GoogleNews	0.605	0.607	0.603
	GloVe	0.630	0.627	0.633
	Internal	0.569	0.555	0.568
CS	GoogleNews	0.448	0.451	0.445
SFeature		0.637	0.636	0.634

The best performance for each classifier is highlighted in bold.

5. Results and discussions

5.1. Evaluation of SFeature

This evaluation section focuses on the salience of the various features for metaphor detection as well as their fitness to the three statistical classifiers by focusing on the VUA Verbs track. The evaluation results on the individual features in terms of the F1-score are summarized in Table 6.

Results in Table 6 show that the best individual feature is the sensorimotor vectors with the LR classifier, followed by B1, W2V.GloVe, Trigram and FL. These results verify the potential contribution of sensorimotor features to metaphor detection. The sensorimotor consistently led to

Table 7. Comparison of SFeature to B1 on all the four tracks

Track	B1	B1 + SFeature	SFeature
VUA-Verbs	0.600	0.642	0.652
VUA-AllPOS	0.589	0.597	0.603
TOEFL-Verbs	0.555	0.581	0.596
TOEFL-AllPOS	0.543	0.552	0.560

The best performance is highlighted in bold.

better results for metaphor detection as compared to the other features. The performances of the three classifiers are quite close for each feature set, with LR performing slightly better.

In addition to the evaluation of individual features, we use the best feature set and classifier (LR) in the above evaluation for testing on all four tracks. The results of our method on the test sets of the four tracks in terms of F1-score are summarized in Table 7.^h

In Table 7, ‘B1+SFeature’ stands for ‘Sensorimotor feature fused with baseline 1’ and the best results for the four tracks are highlighted in bold. The sensorimotor feature shows consistent improvement (1–5%F1) over baseline 1 and is also more effective when used alone. The evaluation results demonstrate the effectiveness of using the sensorimotor feature for metaphor detection. In addition, the results show that the models perform much better for predicting verbs than other lexical categories for both datasets. As highlighted in Table 7, the SFeature model shows 5.2% F1 and 4.1 F1% improvement for verbs over B1 in the two datasets respectively, while the improvements for the averaged performance of the four POS categories in the two datasets are much smaller: 1.4% F1 and 1.7% F1 respectively. Overall, using SFeature shows enhancement for metaphor detection on all lexical categories.

5.2. Evaluation of SGNN

Recall that, in addition to traditional classifiers, we also concatenated sensorimotor information with word embeddings and applied the model to deep neural networks to further explore the performance of sensorimotor-enriched modeling. The evaluation results in this section are summarized in Table 8 in terms of P(recision), R(ecall), and F1(-score).

For a meaningful comparison to other work, we focus on the Verbs track first and randomly select a development set (4,380 tokens) from the training set (17,240 tokens) in proportion to the Train/Test ratio of the task in Leong *et al.* (2020). The current focus on the Verbs track is because most of the reported work for the same task and on the same datasets conduct their experiments on the Verbs track so that we could compare the results directly. The reported results on the ALLPOS track are incomplete and hence incomparable. We report the peer results in Table 9 to observe the model performance across external groups (between models), and then look further into the POS variances within groups (between lexical categories) in Section 5.4.1.

As introduced in Section 4.1, B2 and B3 are implemented for direct comparisons to SGNN. We also implement several other approaches with a minimal difference to SGNN for a more comprehensive comparison that could potentially differentiate the contribution of the linguistic features and neural networks. Table 8 clearly shows that the sensorimotor enhanced models perform better: a 2.4% F1 improvement of SGNN over B1, a 3.8% F1 improvement over B2, a 7% F1 improvement over a pure linguistic model, a 1.5% F1 improvement over the pure neural

^hThe performance difference of B1 is because results in Table 6 are generated from the evaluation experiment which is based on the training data only. We split the training data into two parts for training and development respectively to find the best features. In Table 7, prediction results are based on test sets.

Table 8. Results of sensorimotor-enriched models with neural networks on the Verbs track

Corpus	Category	Approach	P	R	F1
VUA	Baseline 2	ELMo + Glove + LSTM	0.722	0.745	0.737
	Baseline 3	Random Vector + Glove + LSTM	0.720	0.725	0.723
	Linguistic	Sensorimotor + LSTM	0.699	0.675	0.687
	Neural	Glove + LSTM	0.744	0.748	0.746
	SGNN	Sensorimotor + Glove + LSTM	0.767	0.755	0.761
TOEFL	Baseline 2	ELMo + Glove + LSTM	0.686	0.700	0.697
	Baseline 3	Random Vector + Glove + LSTM	0.676	0.689	0.682
	Linguistic	Sensorimotor + LSTM	0.654	0.678	0.666
	Neural	Glove + LSTM	0.714	0.689	0.702
	SGNN	Sensorimotor + Glove + LSTM	0.703	0.732	0.717

The best performance in terms of P, R, F1 is highlighted in bold.

Table 9. Comparison of our result to state-of-the-art works on the Verbs track of the VUA corpus

Work	Approach	F1
Wan <i>et al.</i> (2020a)	Modality + embodiment + LR	0.652
Kuo and Carpuat (2020)	Bi-LSTM + Embeddings + Unigram Lemmas + Spell Correction	0.686
Kumar and Sharma (2020)	Character embeddings + Similarity Networks + Bi-LSTM + Transformer	0.717
Liu <i>et al.</i> (2020)	BERT, XNET + POS tags + Bi-LSTM	0.730
Li <i>et al.</i> (2020)	ALBERT + Bi-LSTM	0.755
Devereux <i>et al.</i> (2018)	BERT: Pre-training of deep bidirectional transformers	0.756
<i>SGNN</i>	<i>Sensorimotor + Glove + LSTM</i>	0.761

The best performance is highlighted in bold.

network model. The improvements are salient and consistent in almost all cases, exception for the precision of the neural model for the TOEFL Corpus, although both the recall and F-score improved. Note that replacing the sensorimotor vectors with randomly generated vectors (B3) does not help improve the performance compared to the one without enhancement. It is shown that adding random vectors lowers the performances across the board. Both facts support our assumption that adding sensorimotor information into the model enhances performance, and that the enhancement is not due to a random effect or the increased vector space.

To further demonstrate the effectiveness of our second proposed model, we compare our results to recent related works on the same dataset focusing on the Verbs track, as displayed in Table 9. All the results are publicly available, as reported in Leong *et al.* (2020).

In Table 9, our method is highlighted in italics. Despite using simple neural networks, our method obtains very promising results: it outperforms all the other related works to various degrees (0.5–11% F1 gain), reaching state-of-the-art performance. Overall, our results are consistently superior to the three strong baselines and other linguistically-based or pure deep learning approaches. The above evaluations demonstrate the effectiveness of our model for metaphor

Table 10. Examples of erroneous predictions by B2 but not by SGNN

Sentence	Source domain	Target domain
<i>In the opposite, old people mostly have big <u>responsibilities</u>.</i>	Visual-Head	Interoception-Head
<i>So they do not want to waste their <u>time</u> by helping their communities.</i>	Visual-Head	Interoception-Head
<i>To seek <u>knowledge</u> is everyone's personal ambition.</i>	Visual-Head	Auditory-Head
<i>Also <u>oil prices</u> are increasing and not steady.</i>	Visual-Foot/Leg	Visual-Head
<i>Knowledge of many academic subjects gives many <u>choices</u> to people.</i>	Visual-Hand/Arm	Visual-Head
<i>Old people have more money and <u>time</u> than the people in young generation.</i>	Visual-Head	Visual-Hand/Arm & Interoception-Head
<i>We belong to the <u>internet era</u>; everything is flowing, everything is moving very quickly.</i>	Visual-Head	Interoception-Head & Visual-Foot/Leg
<i>Students will need to be able to think and reason, <u>computer</u> will help to connect the dots.</i>	Visual-Head	Visual-Hand/Arm
<i>I think that <u>future</u> must go in this direction to have a safe life for us and for future generations.</i>	Visual-Foot/Leg	Interoception-Head
<i>, then I broke through this <u>change</u> and my business and work completely changed.</i>	Visual-Hand/Arm	Visual-Head

The word (of the source domain) is highlighted in bold, and the word (of the target domain) is underlined.

detection, supporting our hypothesis that metaphor often anchors the perception-action schema via their source domains.

5.3. Case analysis

To shed light on the contribution of sensorimotor embedding to the detection of metaphors, we conduct case analysis of the improvement. That is, the examples that are correctly predicted by SGNN, but not by the B2 counterpart. Typical samples are provided in Table 10.ⁱ

The source and target domains of the target word and its immediate context are based on the pre-labeled data of the metaphorical expressions. We also map the source domain and target domain words to the sensorimotor norms data to get their dominant perception and action information. The dominant perception and action of the words as provided in the norms indicate the most salient perceptual and actional effectors of the word in a person's conceptual system. For instance, the word 'big' possesses a Visual-Head dominant sensorimotor, while its real meaning in the first example is to modify 'responsibilities' which possesses an Interceptive-Head dominant sensorimotor. The modality shift from Visual to Interceptive indexes a metaphor with a high probability and hence the sensorimotor-enriched model can correctly predict such cases. Other examples in Table 10 show similar patterns of having a sensorimotor dominance shift either from one modality to another or from one action to another, or both. Note that the sensorimotor information serves as a complementary representation to word embeddings and the effectiveness will not be maximized if used alone. There are also metaphors without sensorimotor mismatches between the target word and its context, as in the following examples.

ⁱThe examples and labels of the source and target domains are manually labeled by two linguists with a high agreement (Pearson correlation = 0.85, p -value = 0.0001)

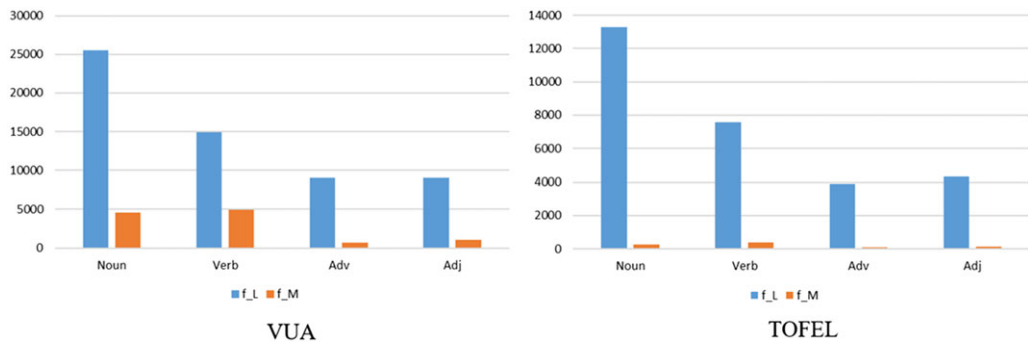


Figure 4. Distribution of metaphorical words across the four POS categories in the two datasets. (f_M: frequency of metaphorical words, f_L: frequency of literal words.)

Example 2-6.

- Ex.2: *Another reason is the **advancement** in Science.* Both Visual-Head
- Ex.3: *This is the general **scheme** of a normal Italian family's **dynamic**.* Both Visual-Hand/Arm
- Ex.4: ***fought** for equal rights, he believed that all men are created equal.* Both Visual-Head
- Ex.5: *Lastly, the family **structure** has been changing.* Both Visual- Hand/Arm
- Ex.6: *A lot of things become **forbidden** just for physical reasons.* Both Visual-Head

Metaphors involve various kinds of domain transfer which can anchor a wide range of basic cognition states such as modality and embodiment experiences. However, its detection is still challenging due to the subtle differences between the metaphorical expressions and their context, and the perception and interpretation of metaphors can vary from person to person. Some may be regarded as dead metaphors that are only used as part of an idiom chunk or other formulaic expressions. Such expressions should be recorded as constructions. Nevertheless, with the incorporation of sensorimotor vector space, word embedding, and linguistic theories, this work has shown promising results of the concerted efforts for metaphor detection, contributing to the interpretation and enhancement of the peer models.

5.4. Cross-sectional comparison

In this section, we explore the issue of whether the performance of the sensorimotor enhanced model is dependent on certain textual properties or not. We examine the performance of the model vis-à-vis parts-of-speech (POS) categories, text genres, and language proficiency levels, as detailed below.

5.4.1. Part of speech

Metaphorical expressions have been suggested to show variance among words of different lexical categories (Shinohara 1999; Ahrens and Huang 2002; Zhao 2018; Dong, Fang, and Qiu 2020). In particular, verbs are argued to be more likely to employ metaphors (Leong et al. 2020), perhaps because of the mutability of verbs (Gentner and France 1988; Ahrens 1999) and that the relational meanings of verbs often rely on metaphoricity (Gentner and Asmuth 2019; Song et al. 2021). We provide the distribution of literal words (f_L) and metaphorical words (f_M) across the four POS categories in the two datasets in Figure 4.

Table 11. Results of model performances across POS categories in the two datasets

Dataset	Approach	All-POS	Verbs	Adjectives	Nouns	Adverbs
VUA	B1	0.589	0.616	0.557	0.564	0.542
	SFeature	0.603	0.625	0.595	0.581	0.552
	gain	0.014	0.009	0.018	<i>0.017</i>	0.010
	B2	0.703	0.737	0.678	0.678	0.648
	SGNN	0.732	0.762	0.708	0.712	0.654
	gain	0.029	0.025	<i>0.030</i>	0.034	0.006
TOEFL	B1	0.528	0.543	0.618	0.415	0.531
	SFeature	0.560	0.587	0.630	0.462	0.517
	gain	0.032	<i>0.044</i>	0.012	0.047	-0.014
	B2	0.692	0.697	0.749	0.641	0.691
	SGNN	0.712	0.717	0.778	0.727	0.626
	gain	0.020	0.020	<i>0.029</i>	0.086	-0.065

The top and second performance gains are highlighted in bold and italics respectively.

According to Figure 4, the four POS categories show different metaphorical distributions in both datasets. Though the occurrences of metaphors in the four POS categories vary to a great extent, the distribution patterns of the metaphorical words and literal words for both datasets are uniform. That is, in terms of literal meaning, the frequency of the four POS categories is the same for both datasets: Noun, Verb, Adverb, and Adjective; in terms of metaphoric meaning, the frequency of the four POS categories are different but the pattern is similar for both datasets: Verb, Noun, Adjective, and Adverb. Among the four, verbs show the highest metaphorical uses, followed by nouns. We aim to investigate how the sensorimotor incorporated model performs in the four different lexical categories (results presented in Table 11), and to what degrees the performances vary.

Table 11 shows that, based on the VUA dataset, sensorimotor methods outperform the baselines consistently across all the POS categories. In particular, the performance gains by sensorimotor methods are the greatest for the Nouns and Adjectives despite the fact verbs are the most frequent among all the metaphorical expressions. This result is consistent with that of the TOEFL dataset, except that Adverbs show no performance gain by the sensorimotor enhanced models. We suspect the reasons why the sensorimotor model works particularly well for nouns and adjectives are: (1) that a general machine learning method tends to perform better on the more frequently attested cases, that is verbs; hence leaves little room for improvement, and (2) that most synesthetic metaphors, mapping between two sensory domains such as ‘sweet voices’, occur with nouns and adjectives, such as in the example of “sweet voice”. Although adjective-noun expressions are not the highest structure among all the metaphorical expressions, they benefit the most from the sensorimotor-enhanced model according to the result in Table 11. In other words, the sensorimotor norms supplement information for the less frequently attested POS in training data. The sensorimotor knowledge for adjective-nouns provides informative information for identifying metaphors in the model, as demonstrated by the many synesthetic metaphors. Since other models tend to perform well on verbal metaphors, the enhancement of our model on verbs is relatively minor. However, the overall performance of the four POS categories is consistently improved by the sensorimotor model for both datasets, confirming the effectiveness of leveraging sensorimotor information for metaphor detection.

Table 12. Results of our methods in comparison to the two baselines across the four text genres

Approach	AllVUA	Academic	Conversation	Fiction	News
B1	0.589	0.721	0.472	0.458	0.606
SFeature	0.603	0.719	0.482	0.476	0.634
Gain	0.014	-0.002	0.010	<i>0.018</i>	0.028
B2	0.703	0.761	0.599	0.651	0.714
SGNN	0.731	0.765	0.656	0.690	0.744
Gain	0.028	0.004	0.057	<i>0.049</i>	0.030

The top and second performance gains are highlighted in bold and italics respectively.

5.4.2. Text genre

Lexical choices are well attested to vary across different texts. The VUA corpus consists of 115 fragments sampled across four genres from the British National Corpus: Academic, News, Conversation, and Fiction. Two previous shared tasks on metaphor detection have adopted this corpus for competition (Leong *et al.* 2018, 2020). The published results demonstrated a pattern that are highly consistent across the participant systems: metaphor detection of texts of Academic and News genres is substantially easier than Fiction and Conversation. This can be accounted for by the fact that Literary and Conversation genres are more creative and more likely to use novel metaphors. While metaphoric uses in Academic and News genres typically are dominated by conventional metaphors, conventional metaphors are also well-attested in training data and hence easier to detect. Given the nature of different usages of metaphors among different genres, they offer another good test to better understand the contribution of the sensorimotor norms to metaphor detection. Thus we conducted further experiments by dividing the dataset into the four subsets according to their text genres and trained the model with the entire training data, but tested on the same sample size from the four text genres respectively. Results are shown in Table 12.

As expected, all the metaphor detection models in Table 12 perform the best in the Academic texts. In addition, the sensorimotor enhanced model gained greater improvement in the other three text genres. In particular, SFeature outperforms the B1 model the most for the News genre with a 2.8% F1 gain, followed by the Fiction genre (a 1.8% F1 gain). In addition, SGNN outperforms the B2 model the most for the Conversation genre with a 5.7% F1 gain, followed by the Fiction genre with a 4.9% F1 gain. This result indicates that genre differences did not contribute to the gains achieved by our model and that the gains by the sensorimotor enhanced models are likely due to its ability to detect novel usages of metaphors.

5.4.3. Language proficiency

Metaphorical expressions have been regarded as an important linguistic index of the language proficiency of writers (Klebanov *et al.* 2018). That is one of the main reasons that the TOEFL corpus is structured according to two language proficiency levels (Medium and High). We experiment on the two subcategories of data in the TOEFL corpus to further explore the possible interaction of the sensorimotor enriched model with language proficiency level for metaphor detection. Results are provided in Table 13.

Interestingly, the results in Table 13 suggest no apparent relation between the language proficiency level, with the sensorimotor-enhanced models as the feature-based method showing a higher performance gain for high proficiency writing, and the neural network method showing

Table 13. Results of our methods in comparison to the two baselines for the two language proficiency levels

Approach	All	High	Medium
B1	0.528	0.533	0.524
SFeature	0.560	0.567	0.552
Gain	0.032	0.034	0.028
B2	0.692	0.713	0.671
SGNN	0.712	0.725	0.682
Gain	0.020	0.012	0.031

Best performance gain is highlighted in bold.

a higher performance gain for medium proficiency writing. Note that there is also a low proficiency subset in the original dataset of TOEFL. This portion of data was excluded from the metaphor labelling by the dataset developer due to too many grammar errors. This fact suggests that grammatical proficiency itself may compound the task of metaphor detection, and the different frequency of metaphoric uses may not be the salient factor. Given potential compounding factors as well as the relatively small size of annotated data of learners' corpus, the issue of correlation between proficiency and metaphor detection cannot currently be resolved.

5.5. Interplay of the 11 sensorimotor dimensions with metaphor prediction

The above models have incorporated all the sensorimotor features; thus, it is not possible to know which dimension plays a more salient role in predicting metaphors. To find the best predictors, we use binary logistic regression for modeling the relations between the 11 sensorimotor dimensions with the metaphoricity of words. We aim to see which dimension is more salient for predicting the metaphoricity in words. We use the TOEFL corpus for running the model. There are 26,736 lexical words in the TOEFL corpus; each word is mapped to the sensorimotor lexicon and a successful mapping returns an 11-dimension vector to the word in terms of the 11 sensorimotor ratings; y is the metaphoricity of the target word. Figure 5 demonstrates the data frame of the x (sensorimotor features) and y (metaphoricity gold label). Due to space limitations, only 20 instances are presented.

Logistic regression is a method for fitting a regression curve, $y = f(x)$, when y is a categorical variable. The typical use of this model is predicting y given a set of predictors x . The categorical variable y , in general, can assume different values. In the simplest case scenario y is binary meaning that it can assume either the value 1 or 0. We run the binomial model using R and the results are provided in Table 14.

The logistic regression result in Table 14 suggests that the interoceptive modality and the hand/arm action effector are the most reliable predictors for metaphors and both show a positive coefficient for predicting a metaphorical expression. In contrast, the olfactory, mouth/throat, and head effectors are negatively related to metaphors, also with significance. This aligns with the observation that metaphors are often expressed via various hand movement activities, as in the expression 'to break through'. Note also that the most significant positive predictor belongs to the action effector category (i.e., *hand/arm*) and the sensory modality (i.e., *interoceptive*). Action effectors are also body parts associated with embodied activities and often serve as the source domain of a conceptual metaphor. In contrast, the *interoceptive* modality, which is associated with mental states, typically occurs as target domains as they are highly abstract. Thus, we postulate that it is the strength of association to either a source domain or a target domain that provides the best predictions.

Table 14. The binary logistic regression results for predicting metaphoricality of words

Independent variable	Coefficient	Std error	p-value
Visual	0.15692	0.10478	0.134237
Auditory	-0.14349	0.10126	0.156481
Gustatory	0.45780	0.22619	0.042971*
Olfactory	-0.70168	0.23225	0.002518**
Tactile	0.06323	0.10590	0.550476
Interoceptive	0.43742	0.10213	1.84e-05***
Leg/foot	0.20644	0.11697	0.077592.
Hand/arm	0.37701	0.11225	0.000783***
Mouth/throat	-0.41800	0.13878	0.002595**
Head	-0.36979	0.12198	0.002434**
Torso	-0.14034	0.16007	0.380630

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

1	Auditory	Gustatory	Haptic	Interoceptive	Olfactory	Visual	Foot_leg	Hand_arm	Head	Mouth	Torso	y
2	1.94	0.82	1.00	0.82	0.82	2.06	0.25	0.50	2.25	1.10	0.10	0
3	2.53	1.88	1.94	2.76	2.06	3.00	1.30	1.50	3.25	1.90	1.15	0
4	2.84	0.37	0.53	2.47	0.42	2.84	0.30	0.85	2.95	2.00	0.35	0
5	3.47	2.76	3.24	2.59	2.65	4.24	2.88	3.13	3.56	3.06	2.50	0
6	1.06	0.28	0.28	1.78	0.28	2.50	1.14	1.43	2.71	1.57	1.33	0
7	1.25	0.00	1.81	0.44	0.00	3.75	1.50	2.10	2.00	1.10	1.35	1
8	2.47	0.21	0.32	1.11	0.21	2.63	0.40	1.10	2.70	2.95	0.45	0
9	3.82	0.91	3.00	1.45	2.09	4.45	2.48	3.38	4.48	3.81	2.43	0
10	1.78	0.22	0.17	0.56	0.17	0.72	0.35	0.75	1.45	1.50	0.25	0
11	2.63	0.05	0.21	1.00	0.58	4.00	0.85	0.90	2.40	0.85	0.85	0
12	1.47	0.37	0.68	0.95	0.21	2.16	0.70	0.65	2.25	1.30	0.85	0
13	2.72	0.22	1.22	0.67	0.44	3.56	2.86	3.29	3.29	2.81	2.24	0
14	1.10	0.20	0.55	2.25	0.20	1.55	0.80	0.90	3.35	1.30	1.15	1
15	2.60	2.15	2.80	1.60	2.30	3.60	1.05	1.75	2.00	0.70	0.90	0
16	2.50	0.56	0.94	0.81	0.44	2.38	0.42	1.47	2.47	1.79	0.37	0
17	2.30	0.30	1.35	1.05	0.30	3.35	0.95	2.00	3.05	1.47	0.84	0
18	2.31	0.56	2.31	2.06	0.88	3.56	2.89	3.67	3.83	2.78	2.72	0
19	1.74	0.47	0.74	1.89	0.58	2.42	3.60	1.95	2.60	1.75	1.75	0
20	1.84	0.47	0.58	0.74	0.47	1.68	0.30	0.60	1.30	1.35	0.30	0

Figure 5. Data frame of the sensorimotor and metaphor data.

6. Conclusion

Following the emergent but challenging trend to synergize neuro-cognitive information in NLP, this paper proposes a novel method for metaphor detection combining sensorimotor norms (Lynott *et al.* 2019) with word embeddings (Pennington *et al.* 2014). The basic perceptual senses (*touch, hearing, smell, taste, vision* and *interoception*) and action effectors (*mouth/throat, hand/arm, foot/leg, head* and *torso*), identified as such by the sensorimotor norms, provide crucial

information for metaphoricity inference and interpretation that is supplementary to word embeddings. That is, the occurrence of a mismatch in perception and (or) action of a target word with its context tends to be metaphorical. We experience and learn from the physical world through our five senses and actions. The sensory, as well as the embodiment lexicon as a collection of sensorimotor words, can hence be considered as basic units of conceptualization of our knowledge of the physical world. These conceptualization units can be utilized as a useful resource for linguistic modeling (Chersoni *et al.* 2020; Zhong, Ahrens, and Huang 2023). On such a basis, we have looked at how such information facilitates metaphor detection. We use statistical machine learning, Bi-LSTM, and other DL architectures for a comprehensive attestation of our proposed method. Results show that the proposed model with access to sensorimotor scores outperforms the counterpart models by at least 0.5% F1, proving that perceptual-conceptual indices are crucial for identifying metaphors. The proposed model achieves results that (1) are significantly higher than one without such information, and (2) show leading performances as compared to most related works on record. The results of the study are in line with results from our pilot study (Wan *et al.* 2020a, 2020b), and strengthen the conclusion that sensory modalities and motor effectors are crucial to metaphors (relevant to both our understanding of metaphors and to NLP).

In addition, sub-experiments are conducted on different lexical categories (nouns, verbs, adjectives and adverbs) with different genres of data (conversation, news, fiction and academic writing) among people of different language proficiency (high and medium) to probe into the grammatical, stylistic and other influence on the model. In particular, the performance gains by sensorimotor methods are the greatest for nouns and adjectives although verbs take the largest proportion of all the metaphorical expressions. This is possibly because the majority of words in the sensorimotor lexicon are nouns or adjectives. It is also observed that most of the synesthetic metaphors occur among nouns and adjectives, such as in the example of “sweet voice”. In addition, although all the models perform the best in the Academic texts for the task of metaphor detection, our proposed methods show greater improvement in the other three genres of texts, including Conversation, News, and Fiction, indicating the high generalization ability of a sensorimotor enriched neural network, which captures the knowledge of semantic, cognitive, and conceptual information in one shot.

There are several interesting directions for future work: First, we will extend our research methods to other types of figurative language. Second, we will run our model on other datasets such as MOH-X (Mohammad *et al.* 2016) and TroFi (Birke and Sarkar 2006) to further attest to the efficiency of our model in terms of model robustness and generalization abilities. Third, through the case analysis, we found that multiple word metaphors affected the performance of the metaphor detection model. We will further consider multi-word metaphor detection using our current approach. We will also try alternative methods in introducing sensorimotor information using other paradigms rather than concatenating embedding with the hope of further enriching the content. In addition to perceptual and actional dimensions, we are also interested to explore other dimensions such as the emotional and imageability predictors for metaphors and probe into their possible relations.

In addition to modeling static conceptual features, some studies (Yee and Thompson-Schill 2016; Trott and Bergen 2022) have also addressed the effects of modeling contextualized features across time scales or on communication efficiency. However, these two studies were not specifically designed for metaphor detection. Still, the idea of contextualized sensorimotor representations can be a very interesting way to probe further into linguistically-enriched methods for metaphor detection in future work.

Note that the existing sensorimotor norms laid the ground work for other enriched neuro-cognitive information that is lexically anchored (Banks and Connell 2023; Reilly, Flurie, and Peelle 2020; Zhong and Ahrens 2023). In addition, sensorimotor norms are now available for Chinese (Zhong *et al.* 2022), Italian (Repetto *et al.* 2022), and Russian (Miklashevsky 2020), among

others. Lastly, Chersoni *et al.* (2020), Chersoni *et al.* (2021a), and Chersoni *et al.* (2021b) illustrated that it is possible to bootstrap sensorimotor norms of a different language based on existing norms. These ongoing developments suggest that our proposed approach has the potential of both bridging to richer neuro-cognitive resources and expanding to multi- and cross-lingual language processing.

References

- Ahrens K. (1999). The mutability of noun and verb meaning. *Chinese Language and Linguistics* 5(2), 335–371.
- Ahrens K., Chung S. F. and Huang C. R. (2003). Conceptual Metaphors: Ontology-based representation and corpora driven Mapping Principles. In *Proceedings of the ACL 2003 Workshop on the Lexicon and Figurative Language*, pp. 36–42.
- Ahrens K. and Huang C. R. (2002). Time passing is motion. *Language and Linguistics* 3(3), 491–519.
- Ahrens K. and Jiang M. (2020). Source domain verification using corpus-based tools. *Metaphor and Symbol* 35(1), 43–55.
- Banks B. and Connell L. (2023). Multi-dimensional sensorimotor grounding of concrete and abstract categories. *Philosophical Transactions of the Royal Society B* 378(1870), 20210366.
- Baroni M. and Lenci A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics* 36(4), 673–721.
- Barrett M. and Hollenstein N. (2020). Sequence labelling and sequence classification with gaze: Novel uses of eye-tracking data for Natural Language Processing. *Language and Linguistics Compass* 14(11), 1–16.
- Barsalou c. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences* 22(4), 577–609, disc. 609–660.
- Birke J. and Sarkar A. (2006). A clustering approach for nearly unsupervised recognition of nonliteral language. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 329–336.
- Bizzoni Y. and Ghanimifard M. (2018). Bigrams and Bi-LSTMs two neural networks for sequential metaphor detection. In *Proceedings of the Workshop on Figurative Language Processing*, pp. 91–101.
- Brooks J. and Youssef A. (2020). Metaphor detection using ensembles of bidirectional recurrent neural networks. In *Proceedings of the Second Workshop on Figurative Language Processing*, pp. 244–249.
- Brysaert M., Warriner A. B. and Kuperman V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods* 46(3), 904–911.
- Casasanto D. and Gijssels T. (2015). What makes a metaphor an embodied metaphor? *Linguistics Vanguard* 1(1), 327.
- Chen X., Hai Z., Wang S., Li D., Wang C. and Luan H. (2021). Metaphor identification: A contextual inconsistency based neural sequence labeling approach. *Neurocomputing* 428, 268–279.
- Chen X., Leong C., Flor M. and Klebanov B. B. (2020). Go Figure! Multi-task transformer-based architecture for metaphor detection using idioms: ETS team in 2020 metaphor shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pp. 235–243.
- Chersoni E., Hollenstein N., Jacobs C. L., Oseki Y., Prévot L. and Santus E. (2021b). *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*.
- Chersoni E., Santus E., Huang C. R. and Lenci A. (2021a). Decoding word embeddings with brain-based semantic features. *Computational Linguistics* 47(3), 663–698.
- Chersoni E., Xiang R. L., Q and Huang C. R. (2020). Automatic learning of modality exclusivity norms with crosslingual word embeddings. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pp. 32–38.
- Choi M., Lee S., Choi E., Park H., Lee J., Lee D. and Lee J. (2021). MeBERT: Metaphor detection via contextualized late interaction using metaphorical identification theories. arXiv preprint arXiv: 2104.13615.
- Church K. W. (2017). Word2Vec. *Natural Language Engineering* 23(1), 155–162.
- Dankers V., Rei M., Lewis M. and Shutova E. (2019). Modelling the interplay of metaphor and emotion through multi-task learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2218–2229.
- Devereux B., Shutova K. and Huang C. R. (2018). Proceedings of the first workshop on linguistic and neurocognitive resources. In *Proceedings of Workshops at the 2018 Language Resources and Evaluation Conference*.
- Dong M., Fang A. C. and Qiu X. (2020). Shell nouns as grammatical metaphor in knowledge construal: Variation across science and engineering discourse. *Lingua* 248, 102946.
- Dunn J. (2014, June). Measuring metaphoricity. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics Volume 2: Short Papers*, pp. 745–751.
- Fass D. (1991). met*: A method for discriminating metonymy and metaphor by computer. *Computational Linguistics* 17(1), 49–90.
- Gao G., Choi E., Choi Y. and Zettlemoyer Y. (2018). Neural metaphor detection in context. arXiv preprint arXiv: 1808.09653.
- Gentner D. (1988). Metaphor as structure mapping: The relational shift. In *Child Development*, pp. 47–59.

- Gentner D. and Asmuth J.** (2019). Metaphoric extension, relational categories, and abstraction. *Language, Cognition and Neuroscience* 34(10), 1298–1307.
- Gentner D. and France I. M.** (1988). The verb mutability effect: Studies of the combinatorial semantics of nouns and verbs. In *Lexical Ambiguity Resolution*. San Mateo, CA: Morgan Kaufmann, pp. 343–382.
- Gibbs R. W.** (1996). Why many concepts are metaphorical. *Cognition* 61(3), 309–319.
- Gibbs R. W.** (2006). Metaphor interpretation as embodied simulation. *Mind & Language* 21(3), 434–458.
- Gibbs Jr. R. W., Lima P. L. C. and Francozo E.** (2004). Metaphor is grounded in embodied experience. *Journal of Pragmatics* 36(7), 1189–1210.
- Gong H., Gupta K., Jain A. and Bhat S.** (2020). Illinimet: Illinois system for metaphor detection with contextual and linguistic information. In *Proceedings of the Second Workshop on Figurative Language Processing*, pp. 146–153.
- Hollenstein N., Chersoni E., Jacobs C. L., Oseki Y., Prévot L. and Santus E.** (2021). CMCL 2021 shared task on eye-tracking prediction. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pp. 72–78.
- Hong J.** (2016). Automatic metaphor detection using constructions and frames. *Constructions and Frames* 8(2), 295–322.
- Huang C. R., Prévot L., Su I. L. and Hong J. F.** (2007). Towards a conceptual core for multicultural processing: A multilingual ontology based on the Swadesh list. In *Intercultural Collaboration: First International Workshop, IWIC 2007 Kyoto, Japan, January 25-26, 2007, Invited and Selected Papers*. Berlin/Heidelberg: Springer, pp. 17–30.
- Jang H., Moon S., Jo Y. and Rosé C. P.** (2015). Metaphor detection in discourse. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 384–392.
- Kennington C.** (2021). Enriching language models with visually-grounded word vectors and the Lancaster sensorimotor norms. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pp. 148–157.
- Klebanov B. B., Leong C. and Flor M.** (2018). A corpus of non-native written English annotated for metaphor. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 86–91.
- Klebanov B. B., Leong C. and Gutierrez E. D.** (2016). Semantic classifications for detection of verb metaphors. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 101–106.
- Klebanov B. B., Leong C., Heilman M. and Flor M.** (2014). Different texts, same metaphors: Unigrams and beyond. In *Proceedings of the Second Workshop on Metaphor in NLP*, pp. 11–17.
- Klebanov B. B., Leong C., Heilman M. and Flor M.** (2015). Supervised word-level metaphor detection: Experiments with concreteness and reweighting of examples. In *Proceedings of the Third Workshop on Metaphor in NLP*, pp. 11–20.
- Kovcses Z.** (2000). The scope of metaphor. *Topics in English Linguistics* 30, 79–92.
- Kumar T. and Sharma Y.** (2020). Character aware models with similarity learning for metaphor detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pp. 116–125.
- Kuo K. and Carpuat M.** (2020). Evaluating a bi-lstm model for metaphor detection in toefl essays. In *Proceedings of the Second Workshop on Figurative Language Processing*, pp. 192–196.
- Lakoff G.** (2012). Explaining embodied cognition results. *Topics in Cognitive Science* 4(4), 773–785.
- Lakoff G. and Johnson M.** (1980). *Metaphors We Live By*. Chicago, IL: University of Chicago.
- Lenci A.** (2008). Distributional semantics in linguistic and cognitive research. *Italian Journal of Linguistics* 20(1), 1–31.
- Leong C., Klebanov B. B., Hamill C., Stemle E., Ubale R. and Chen X.** (2020). A report on the 2020 *vua* and *toefl* metaphor detection shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pp. 18–29.
- Leong C., Klebanov B. B. and Shutova E.** (2018). A report on the 2018 *vua* metaphor detection shared task. In *Proceedings of the Workshop on Figurative Language Processing*, pp. 56–66.
- Li S., Zeng J., Zhang J., Peng T., Yang L. and Lin H.** (2020). Albert-Bi-LSTM for sequential metaphor detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pp. 110–115.
- Liu J., O'Hara N., Rubin A., Draelos R. and Rudin C.** (2020). Metaphor detection using contextual word embeddings from transformers. In *Proceedings of the Second Workshop on Figurative Language Processing*, pp. 250–255.
- Liu Y., Ott M., Goyal N., Du J., Joshi M., Chen D., Levy M., Lewis M., Zettlemoyer L. and Stoyanov V.** (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv: 1907.11692.
- Long Y., Xiang R., Lu Q., Huang C. R. and Li M.** (2019). Improving attention model based on cognition grounded data for sentiment analysis. *IEEE Transactions on Affective Computing* 12(4), 900–912.
- Loper E. and Bird S.** (2002). Nltk: The natural language toolkit. arXiv preprint cs/0205028.
- Lynott D. and Connell L.** (2009). Modality exclusivity norms for 423 object properties. *Behavior Research Methods* 41(2), 558–564.
- Lynott D., Connell L., Brysbaert M., Brand J. and Carney J.** (2019). The Lancaster Sensorimotor Norms: Multidimensional measures of perceptual and action strength for 40,000 English words. *Behavior Research Methods* 52(3), 1271–1291.
- Mao R., Lin C. and Guerin F.** (2019, July). End-to-end sequential metaphor identification inspired by linguistic theories. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3888–3898.
- Martin J. H.** (1990). *A Computational Model of Metaphor Interpretation*. San Diego, CA: Academic Press Professional, Inc.
- Mason Z. J.** (2004). CorMet: a computational, corpus-based conventional metaphor extraction system. *Computational Linguistics* 30(1), 23–44.

- Maudslay R. H., Pimentel T., Cotterell R. and Teufel S. (2020). Metaphor detection using context and concreteness. In *Proceedings of the Second Workshop on Figurative Language Processing*, pp. 221–226.
- Miklashevsky A. (2020). Sensorimotor norms for 506 Russian nouns. In *Proceedings of the Second Workshop on Linguistic and Neurocognitive Resources*, pp. 59–60.
- Mikolov T., Chen K., Corrado G. and Dean J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv: [1301.3781](https://arxiv.org/abs/1301.3781).
- Mohammad S. M., Shutova E. and Turney Peter D. (2016). Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pp. 23–33.
- Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Müller A., Nothman J., Louppe G., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M. and Duchesnay É. (2011). Scikit-Learn: Machine learning in Python. *Journal of machine Learning research* **12**, 2825–2830.
- Pennington J., Socher R. and Manning C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543.
- Peters M. E., Neumann M., Iyyer M., Gardner M., Clark C., Lee K. and Zettlemoyer L. (2018). Deep contextualized word representations. arXiv preprint arXiv: [1802.05365](https://arxiv.org/abs/1802.05365).
- Ploux S. and Victorri B. (1998). Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes. *Revue TAL* **39**, 161–182.
- Pragglejaz Group (2007). MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol* **22**(1), 1–39.
- Rai S. and Chakraverty S. (2020). A survey on computational metaphor processing. *ACM Computing Surveys (CSUR)* **53**(2), 1–37.
- Rai S., Chakraverty S., Tayal D. K. and Kukreti Y. (2018). A study on impact of context on metaphor detection. *The Computer Journal* **61**(11), 1667–1682.
- Rai S., Chakraverty S., Tayal D. K., Sharma D. and Garg A. (2019). Understanding metaphors using emotions. *New Generation Computing* **37**(1), 5–27.
- Raval S., Sedghamiz H., Santus E., Alhanani T., Ghassemi M. and Chersoni E. (2021). Exploring a unified sequence-to-sequence transformer for medical product safety monitoring in social media. arXiv preprint arXiv: [2109.05815](https://arxiv.org/abs/2109.05815).
- Reilly J., Flurie M. and Peelle J. E. (2020). The English lexicon mirrors functional brain activation for a sensory hierarchy dominated by vision and audition: Point-counterpoint. *Journal of Neurolinguistics* **55**, 100895.
- Repetto C., Rodella C., Conca F., Santi G. C. and Cattricalà E. (2022). The Italian Sensorimotor Norms: Perception and action strength measures for 959 words. *Behavior Research Methods*. <https://doi.org/10.3758/s13428-022-02004-1>
- Rogers A., Kovaleva O. and Rumshisky A. (2020). A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics* **8**, 842–866.
- Rohanian O., Rei M., Taslimipoor S. and Ha L. (2020, July). Verbal multiword expressions for identification of metaphor. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL.
- Santus E., Chiu T. S., Lu Q., Lenci A. and Huang C. R. (2016). What a nerd! Beating students and vector cosine in the ESL and TOEFL datasets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 4565–4572.
- Shinohara K. (1999). Constraints on motion verbs in the TIME IS MOTION metaphor. *Annual Meeting of the Berkeley Linguistics Society* **25**(1), 250–271.
- Shutova E. (2010, June). Automatic metaphor interpretation as a paraphrasing task. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 1029–1037.
- Shutova E., Kiela D. and Maillard J. (2016). Black holes and white rabbits: Metaphor identification with visual features. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 160–170.
- Song W., Zhou S., Fu R., Liu T. and Liu L. (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)*. Online. Association for Computational Linguistics, pp. 4240–4251.
- Steen G. E. (2010). *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*, vol. 14. Philadelphia, PA: John Benjamins Publishing.
- Su C., Fukumoto F., Huang X., Li J., Wang R. and Chen Z. (2020). DeepMet: A reading comprehension paradigm for token-level metaphor detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pp. 30–39.
- Tanasescu C., Kesarwani V. and Inkpen D. (2018, May). Metaphor detection by deep learning and the place of poetic metaphor in digital humanities. In *The Thirty-first International Flairs Conference*.
- Tekiroğlu S. S., Özbal G. and Strapparava C. (2015). Exploring sensorial features for metaphor identification. In *Proceedings of the Third Workshop on Metaphor in NLP*, pp. 31–39.
- Trott S. and Bergen B. (2022). Languages are efficient, but for whom? *Cognition* **225**, 105094.
- Turney P. D. and Littman M. L. (2005). Corpus-based learning of analogies and semantic relations. *Machine Learning* **60**(1-3), 251–278.

- Veale T. (2003). Systematicity and the lexicon in creative metaphor. In *Proceedings of the ACL. 2003 Workshop on the Lexicon and Figurative Language*, pp. 28–35.
- Veale T., Shutova E. and Klebanov B. B. (2016). *Metaphor: A Computational Perspective. Synthesis Lectures on Human Language Technologies*, vol. 31. San Rafael, CA: Morgan & Claypool Publishers.
- Wan M., Ahrens K., Chersoni E., Jiang M., Su Q., Xiang R. and Huang C. R. (2020a). Using conceptual norms for metaphor detection. In *Proceedings of the Second Workshop on Figurative Language Processing, ACL*, pp. 104–109.
- Wan M. and Xing B. (2020). Modality enriched neural network for metaphor detection. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING'2020)*, Barcelona, Spain, online.
- Wan M., Xing B., Su Q., Liu P. and Huang C. R. (2020b). Sensorimotor enhanced neural network for metaphor detection. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation (PACLIC34)*, online.
- Wilks Y. (1975). A preferential, pattern-seeking, semantics for natural language inference. *Artificial Intelligence* 6(1), 53–74.
- Wilks Y., Dalton A., Allen J. and Galescu L. (2013). Automatic metaphor detection using large-scale lexical resources and conventional metaphor extraction. In *Proceedings of the First Workshop on Metaphor in NLP*, pp. 36–44.
- Wilson M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review* 9(4), 625–636.
- Wu C., Wu F., Chen Y., Wu S., Yuan Z. and Huang Y. (2018). Neural metaphor detecting with CNN-LSTM model. In *Proceedings of the Workshop on Figurative Language Processing*, pp. 110–114.
- Xu H., Santus E., Laszlo A. and Huang C. R. (2015). LLT-PolyU: Identifying sentiment intensity in ironic tweets. In *Proceedings of the 9th International Workshop on Semantic Evaluation (Semeval 2015)*, pp. 673–678.
- Yee E. and Thompson-Schill S. L. (2016). Putting concepts into context. *Psychonomic Bulletin & Review* 23(4), 1015–1027.
- Yu N. (2003). Synesthetic metaphor: A cognitive perspective. *Journal of Literary Semantics* 32(1), 19–34.
- Zafrir O., Boudoukh G., Izsak P. and Wasserblat M. (2019, December). Q8bert: Quantized 8bit bert. In *2019 Fifth Workshop on Energy Efficient Machine Learning and Cognitive Computing-NeurIPS Edition (EMC2-NIPS)*. IEEE, pp. 36–39.
- Zafrir O., Larey A., Boudoukh G., Shen H. and Wasserblat M. (2021). Prune once for all: Sparse pre-trained language models. arXiv preprint arXiv: 2111.05754.
- Zayed O., McCrae J. P. and Buitelaar P. (2018). Phrase-level metaphor identification using distributed representations of word meaning. In *Proceedings of the Workshop on Figurative Language Processing*, pp. 81–90.
- Zhang L. and Barnden J. (2013). Towards a semantic-based approach for affect and metaphor detection. *International Journal of Distance Education Technologies (IJDET)* 11(2), 48–65.
- Zhang S. and Liu Y. (2022, October). Metaphor detection via linguistics enhanced Siamese network. In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 4149–4159.
- Zhao Q. (2018). *Synaesthesia, metaphor, and cognition: A corpus-based study on synaesthetic adjectives in Mandarin Chinese*. PhD Thesis, The Hong Kong Polytechnic University.
- Zhong Y. and Ahrens K. (2023). The emotion code in sensory modalities: An investigation of the relationship between sensorimotor dimensions and emotional valence-arousal. In *Chinese Lexical Semantics: 23rd Workshop, CLSW 2022, Virtual Event, May 14-15, 2022, Revised Selected Papers, Part II*. Cham: Springer, pp. 183–192.
- Zhong Y., Ahrens K. and Huang C. R. (2023). Entity, event, and sensory modalities: An onto-cognitive account of sensory nouns. *Humanities and Social Sciences Communications* 10(1), 255. <https://doi.org/10.1057/s41599-023-01677-z>
- Zhong Y., Wan M., Ahrens K. and Huang C.-R. (2022). Sensorimotor norms for Chinese nouns and their relationship with orthographic and semantic variables. *Language, Cognition and Neuroscience* 37(8), 1000–1022.