Black Holes and White Rabbits: Metaphor Identification with Visual Features

Ekaterina Shutova
Computer Laboratory
University of Cambridge
es407@cam.ac.uk

Douwe Kiela
Computer Laboratory
University of Cambridge
dk427@cam.ac.uk

Jean Maillard
Computer Laboratory
University of Cambridge
jean@maillard.it

Abstract

Metaphor is pervasive in our communication, which makes it an important problem for natural language processing (NLP). Numerous approaches to metaphor processing have thus been proposed, all of which relied on linguistic features and textual data to construct their models. Human metaphor comprehension is, however, known to rely on both our linguistic and perceptual experience, and vision can play a particularly important role when metaphorically projecting imagery across domains. In this paper, we present the first metaphor identification method that simultaneously draws knowledge from linguistic and visual data. Our results demonstrate that it outperforms linguistic and visual models in isolation, as well as being competitive with the best-performing metaphor identification methods, that rely on hand-crafted knowledge about domains and perception.

1 Introduction

Metaphor lends vividness, sophistication and clarity to our thought and communication. At the same time, it plays a fundamental structural role in our cognition, helping us to organise and project knowledge (Lakoff and Johnson, 1980; Feldman, 2006). Metaphors arise due to systematic associations between distinct, and seemingly unrelated, concepts. For instance, when we talk about "the *turning wheels* of a political regime", "*rebuilding* the campaign *machinery*" or "*mending* foreign policy", we view *politics* and *political systems* in terms of *mechanisms*, they can *function, break, be mended* etc. The existence of this association allows us to transfer knowledge and imagery from the domain of *mechanisms* (the source domain) to that of *political systems* (the target domain). According to Lakoff and Johnson (1980), such metaphorical mappings, or *conceptual metaphors*, form the basis of metaphorical language.

Metaphor is pervasive in our communication, which makes it important for NLP applications dealing with real-world text. A number of approaches to metaphor processing have thus been proposed, using supervised classification (Gedigian et al., 2006; Mohler et al., 2013; Tsvetkov et al., 2013; Hovy et al., 2013; Dunn, 2013a), clustering (Shutova et al., 2010; Shutova and Sun, 2013), vector space models (Shutova et al., 2012; Mohler et al., 2014), lexical resources (Krishnakumaran and Zhu, 2007; Wilks et al., 2013) and web search with lexico-syntactic patterns (Veale and Hao, 2008; Li et al., 2013; Bollegala and Shutova, 2013). So far, these and other metaphor processing works relied on textual data to construct their models. Yet, several experiments indicated that perceptual properties of concepts, such as concreteness and imageability, are important features for metaphor identification (Turney et al., 2011; Neuman et al., 2013; Gandy et al., 2013; Strzalkowski et al., 2013; Tsvetkov et al., 2014). However, all of these methods used manually-annotated linguistic resources to determine these properties (such as the MRC concreteness database (Wilson, 1988)). To the best of our knowledge, there has not yet been a metaphor processing method that employed information learned from both linguistic and visual data. Ample re-

Proceedings of NAACL-HLT 2016, pages 160–170,
San Diego, California, June 12-17, 2016.

search in cognitive science suggests that human meaning representations are not merely a product of our linguistic exposure, but are also grounded in our perceptual system and sensori-motor experience (Barsalou, 2008; Louwerse, 2011). Semantic models integrating information from multiple modalities have been shown successful in tasks such as modeling semantic similarity and relatedness (Silberer and Lapata, 2012; Bruni et al., 2014), lexical entailment (Kiela et al., 2015a), compositionality (Roller and Schulte im Walde, 2013) and bilingual lexicon induction (Kiela et al., 2015b). Using visual information is particularly relevant to modelling metaphor, where imagery is ported across domains.

In this paper, we present the first metaphor identification method integrating meaning representations learned from linguistic and visual data. We construct our representations using a skip-gram model of Mikolov et al. (2013a) trained on textual data to obtain linguistic embeddings and a deep convolutional neural network (Kiela and Bottou, 2014) trained on image data to obtain visual embeddings. Linguistic word embeddings have been previously successfully used to answer analogy questions (Mikolov et al., 2013b; Levy and Goldberg, 2014). These works have shown that such representations capture the nuances of word meaning needed to recognise relational similarity (e.g. between pairs "*king : queen*" and "*man : woman*"), quantified by the respective vector offsets ($king - queen \approx man - woman$). In our experiments, we investigate how well these representations can capture information about source and target domains and their interaction in a metaphor. We then enrich these representations with visual information. We first acquire linguistic and visual embeddings for individual words and then extend the methods to learn embeddings for longer phrases. The focus of our experiments is on metaphorical expressions in verb–subject, verb–direct object and adjectival modifier–noun constructions. We thus learn embeddings for verbs, adjectives, nouns, as well as verb–noun and adjective–noun phrases. We then use a set of arithmetic operations on word and phrase embedding vectors to classify phrases as literal or metaphorical. To the best of our knowledge, our approach is also the first one to apply word or phrase embeddings to the task of metaphor identification.

Our results demonstrate that the joint model incorporating linguistic and visual representations outperforms the linguistic model in isolation, as well as being competitive with the best-performing metaphor identification methods that rely on hand-crafted information about domains, concreteness and imageability.

2 Related work

A strand of metaphor processing research cast the problem as a classification of linguistic expressions as metaphorical or literal. They experimented with a number of features, including lexical and syntactic information and higher-level features such as semantic roles and domain types. Gedigian et al. (2006) classified verbs related to MOTION and CURE within the domain of financial discourse. They used the maximum entropy classifier and the verbs' nominal arguments and their semantic roles as features, reporting encouraging results. Dunn (2013a) used a logistic regression classifier and high-level properties of concepts extracted from SUMO ontology, including domain types (ABSTRACT, PHYSICAL, SOCIAL, MENTAL) and event status (PROCESS, STATE, OBJECT). Tsvetkov et al. (2013) also used logistic regression and coarse semantic features, such as concreteness, animateness, named entity types and WordNet supersenses. They have shown that the model learned with such coarse semantic features is portable across languages. The work of Hovy et al. (2013) is notable as they focused on compositional rather than categorical features. They trained an SVM with dependency-tree kernels to capture compositional information, using lexical, part-of-speech tag and WordNet supersense representations of sentence trees. Mohler et al. (2013) aimed at modelling conceptual information. They derived semantic signatures of texts as sets of highly-related and interlinked WordNet synsets. The semantic signatures served as features to train a set of classifiers (maximum entropy, decision trees, SVM, random forest) that map new metaphors to the semantic signatures of the known ones.

Turney et al. (2011) hypothesized that metaphor is commonly used to describe abstract concepts in terms of more concrete or physical experiences. Thus, Turney and colleagues expected that there would be some discrepancy in the level of concrete-

ness of source and target terms in the metaphor. They developed a method to automatically measure concreteness of words and applied it to identify verbal and adjectival metaphors. Neuman et al. (2013) and Gandy et al. (2013) followed in Turney's steps, extending the models by incorporating information about selectional preferences.

Heintz et al. (2013) and Strzalkowski et al. (2013) focused on modeling topical structure of text to identify metaphor. Their main hypothesis was that metaphorical language (coming from a different domain) would represent atypical vocabulary within the topical structure of the text. Strzalkowski et al. (2013) acquired a set of topic chains by linking semantically related words in a given text. They then looked for vocabulary outside the topic chain and yet connected to topic chain words via syntactic dependencies and exhibiting high imageability. Heintz et al. (2013) used LDA topic modelling to identify sets of source and target domain vocabulary. In their system, the acquired topics represented source and target domains, and sentences containing vocabulary from both were tagged as metaphorical.

Other approaches addressed automatic identification of conceptual metaphor. Mason (2004) automatically acquired domain-specific selectional preferences of verbs, and then, by mapping their common nominal arguments in different domains, arrived at the corresponding metaphorical mappings. For example, the verb *pour* has a strong preference for *liquids* in the LAB domain and for *money* in the FINANCE domain, suggesting the mapping MONEY is LIQUID. Shutova et al. (2010) pointed out that the metaphorical uses of words constitute a large portion of the dependency features extracted for abstract concepts from corpora. For example, the feature vector for *politics* would contain GAME or MECHANISM terms among the frequent features. As a result, distributional clustering of abstract nouns with such features identifies groups of diverse concepts metaphorically associated with the same source domain (or sets of source domains). Shutova et al. (2010) exploit this property of co-occurrence vectors to identify new metaphorical mappings starting from a set of examples. Shutova and Sun (2013) used hierarchical clustering to derive a network of concepts in which metaphorical associations are learned in an unsupervised way.

3 Method

3.1 Learning linguistic representations

We obtained our linguistic representations using the log-linear skip-gram model of Mikolov et al. (2013a). Given a corpus of words w and their contexts c, the model learns a set of parameters θ that maximize the overall corpus probability

$$\arg\max_{\theta} \prod_{w} [\prod_{c \in C(w)} p(c|w;\theta)], \tag{1}$$

where $C(w)$ is a set of contexts of word w and $p(c|w;\theta)$ is a softmax function:

$$p(c|w;\theta) = \frac{e^{v_c \cdot v_w}}{\sum_{c' \in C} e^{v_{c'} \cdot v_w}}, \tag{2}$$

where v_c and v_w are vector representations of c and w. The parameters we need to set are thus v_{c_i} and v_{w_i} for all words in our word vocabulary V and context vocabulary C, and the set of dimensions $i \in 1, \ldots, d$. Given a set D of word-context pairs, embeddings are learned by optimizing the following objective:

$$\begin{aligned} &\arg\max_{\theta} \sum_{(w,c) \in D} \log p(c|w) = \\ &\sum_{(w,c) \in D} (\log e^{v_c \cdot v_w} - \log \sum_{c' \in C} e^{v_{c'} \cdot v_w}) \end{aligned} \tag{3}$$

We used a recent dump of Wikipedia[1] as our corpus. The text was lemmatized, tagged, and parsed with Stanford CoreNLP (Manning et al., 2014). Words that appeared less than 100 times in their lemmatized form were ignored. The 100-dimensional word and phrase embeddings were learned in two stages: in a first pass, we obtained word-level embeddings (e.g. for *white* and *rabbit*) using the standard skip-gram with negative sampling of Eq. (3); we then obtained phrase embeddings (e.g. for *white rabbit*) through a second pass over the same corpus. In the second pass, the vectors v_c and $v_{c'}$ of Eq. (3) were set to their values from the first pass, and kept fixed. Verb-noun phrases were extracted by finding `nsubj` and `dobj` arcs with VB head and NN dependent; analogously, adjective-noun phrases were extracted by finding `amod` arcs with NN head and JJ dependent. No frequency cutoff was applied for

[1] https://dumps.wikimedia.org/enwiki/20150805/

phrases. All embeddings were trained on the corpus for 3 epochs, using a symmetric window of 5, and 10 negative samples per word-context pair.

3.2 Learning visual representations

Visual embeddings were obtained in a manner similar to Kiela and Bottou (2014). Using the deep learning framework Caffe (Jia et al., 2014), we extracted image embeddings from a deep convolutional neural network that was trained on the ImageNet classification task (Russakovsky et al., 2015). The network (Krizhevsky et al., 2012) consists of 5 convolutional layers, followed by two fully connected rectified linear unit (ReLU) layers that feed into a softmax for classification. The network learns through a multinomial logistic regression objective:

$$J(\theta) = -\sum_{i=1}^{D}\sum_{k=1}^{K}\mathbf{1}\{y^{(i)}=k\} \log\frac{\exp(\theta^{(k)\top}x^{(i)})}{\sum_{j=1}^{K}\exp(\theta^{(j)\top}x^{(i)})} \quad (4)$$

where $\mathbf{1}\{\cdot\}$ is the indicator function and we train on D examples with K classes. We obtain image embeddings by doing a forward pass with a given image and taking the 4096-dimensional fully connected layer that precedes the softmax (typically called FC7) as the representation of that image.

To construct our embeddings, we used up to 10 images for a given word or phrase, which were obtained through Google Images. It has been shown that images from Google yield higher quality representations than comparable resources such as Flickr and are competitive with hand-crafted datasets (Fergus et al., 2005; Bergsma and Goebel, 2011). We created our final visual representations for words and phrases by taking the average of the extracted image embeddings for a given word or phrase.

3.3 Multimodal fusion strategies

While it is desirable to jointly learn representations from different modalities at the same time, this is often not feasible (or may lead to poor performance) due to data sparsity. Instead, we learn uni-modal representations independently, as described above, and then combine them into multi-modal ones. Previous work in multi-modal semantics (Bruni et al., 2014) investigated different ways of combining, or *fusing*, linguistic and perceptual cues. When calculating similarity, for instance, one can either combine the representations first and subsequently compute similarity scores; or compute similarity scores independently per modality and afterwards combine the scores. In contrast with joint learning (which has also been called *early fusion*), these two possibilities represent *middle* and *late* fusion, respectively (Kiela and Clark, 2015).

We experiment with middle and late fusion strategies. In middle fusion, we L-2 normalise and concatenate the vectors for linguistic and visual representations and then compute a metaphoricity score for a phrase based on this joint representation. In late fusion, we first compute the metaphoricity scores based on linguistic and visual representations independently, and then combine the metaphoricity scores by taking their average.

3.4 Measuring metaphoricity

We investigate a set of arithmetic operations on the linguistic, visual and multimodal embedding vectors to determine whether the two words in the phrase belong to the same domain or rather a word from one domain is metaphorically used to describe another.

3.4.1 Word-level embeddings

In our first set of experiments, we compare embeddings learned for individual words in order to determine whether they come from the same domain. This is done by determining similarity between the representations of the two words in a phrase:

$$\text{sim}(word_1, word_2), \quad (5)$$

where $word_1$ is either a verb or an adjective, $word_2$ is a noun, and similarity is defined as cosine similarity:

$$\cos(x, y) = \frac{x \cdot y}{||x||||y||} \quad (6)$$

We expect the similarity of word representations to be lower for metaphorical expressions (where one word comes from the source domain and one from the target), than for the literal ones (where both words come from the target domain). We will further refer to this method as WORDCOS.

3.4.2 Phrase-level embeddings

In our second set of experiments, we investigate compositional properties of metaphorical phrases by comparing the embeddings learned for the whole phrase with those of the individual words in the phrase. This allows us to determine which properties the phrase shares with each of the words, providing another criterion for metaphor identification. We expect that the embeddings of literal phrases will be more similar to the embeddings of individual words in the phrase (or a combination thereof) than those of metaphorical phrases. We use the following measures to test this hypothesis:

$$\text{PHRASCOS1: } \cos(phrase - word_1, word_2) \quad (7)$$

$$\text{PHRASCOS2: } \cos(phrase - word_2, word_1) \quad (8)$$

$$\text{PHRASCOS3: } \cos(phrase, word_1 + word_2), \quad (9)$$

where $phrase$ is the phrase embedding vector, and $word_1$ and $word_2$ are defined as above.

3.4.3 Classification

We use a small development set (a collection of phrases annotated as metaphorical or literal) to determine an optimal classification threshold for each of the above scoring methods. We have optimized the threshold by maximizing classification accuracy on the development set.[2] All instances with values above the threshold were considered literal and those with values below the threshold metaphorical. The thresholds were then applied to classify the test instances as literal or metaphorical.

4 Experiments

4.1 Annotated datasets

We evaluate our method using two datasets manually annotated for metaphoricity:

Mohammad et al. dataset (MOH) Mohammad et al. (2016) annotated different senses of WordNet (Fellbaum, 1998) verbs for metaphoricity. They extracted verbs that had between three and ten senses in WordNet and the sentences exemplifying them in the corresponding glosses. The verb uses in the sentences (1639 in total) were then annotated for metaphoricity by 10 annotators each via the crowdsourcing platform CrowdFlower[3]. Mohammad et al. selected the verbs that were tagged by at least 70% of the annotators as metaphorical or literal to create their dataset. We extracted verb–direct object and verb–subject relations of the annotated verbs from this dataset, discarding the instances with pronominal or clausal subject or object. This resulted in a dataset of 647 verb–noun pairs, 316 of which were metaphorical and 331 literal. Figure 1 shows some examples of annotated verbs from Mohammad et al.'s dataset.

Verb noun	Class	Relation
blister foot	literal	SV
blister administration	metaphorical	VO
blur haze	literal	SV
blur vision	literal	VO
blur distinction	metaphorical	SV
boost economy	metaphorical	VO
boost voltage	literal	VO
bounce ball	literal	SV
bounce people	metaphorical	VO
bow person	literal	SV
bow government	metaphorical	SV
breathe person	literal	SV
breathe life	metaphorical	VO
breathe fabric	metaphorical	SV
breathe wine	metaphorical	SV

Figure 1: Annotated verb–direct object and verb–subject pairs from MOH

Tsvetkov et al. dataset (TSV) Tsvetkov et al. (2014) created a large dataset of adjective–noun pairs that they annotated for metaphoricity. Starting with a 1000 frequent adjectives, they extracted nouns they co-occur with in TenTen Web Corpus[4] using SketchEngine and in collections of metaphor on the Web. Tsvetkov et al. divided the data into a training set (containing 884 literal and 884 metaphorical pairs) and test set (111 literal and 111 metaphorical pairs). We will refer to their training set as TSV-TRAIN and to the test set as TSV-TEST. The test set was annotated for metaphoricity by 5 annotators with an inter-annotator agreement of $\kappa = 0.76$. Figure 2 shows a portion of the anno-

[2] We have also experimented with optimizing F-score on the development set and the results exhibited similar trends across methods.

[3] www.crowdflower.com

[4] https://www.sketchengine.co.uk/xdocumentation/wiki/Corpora/enTenTen

Metaphorical:	**Literal:**
bald assertion	cold beer
blind alley	cold weather
breezy disregard	huge number
dry wit	dead animal
dumb luck	deep sea
foggy brain	gold coin
healthy balance	dry skin
hollow mockery	honest opinion
honest meal	empty can
juicy scandal	good idea
spicy language	foggy night
stale cliché	frosty morning
steep discount	firm mattress

Figure 2: Annotated adjective–noun pairs from TSV-TEST

tated test set. Metaphorical phrases that depend on wider context for their interpretation (e.g. *drowning students*) were removed. The training set was annotated by one annotator only, and it is thus likely that the annotations are less reliable than those in the test set. We thus evaluate our methods on Tsvetkov et al.'s test set (TSV-TEST). However, we will also report results on TSV-TRAIN to confirm whether the observed trends hold in a larger, though likely noisier, dataset.

We selected the above two datasets since they include examples for different senses (both metaphorical and literal) of the same verbs or adjectives. This allows us to test the extent to which our model is able to discriminate between different word senses, as opposed to merely selecting the most frequent class for a given word.

4.2 Experimental setup

We divided the verb- and adjective-noun datasets into development and test sets. The verb–noun development set contained 80 instances from MOH (40 literal and 40 metaphorical), leaving us with the test set of 567 verb-noun pairs from MOH. We created the adjective–noun development set using 80 adjective-noun pairs (40 literal and 40 metaphorical) from TSV-TRAIN, leaving all of the 222 adjective–noun pairs in TSV-TEST for evaluation. In a separate experiment, we also applied our methods to the remainder of TSV-TRAIN (1688 adjective–noun pairs) to evaluate our system on a larger adjective dataset.

We used the development sets to determine an optimal threshold value for each of our scoring methods. The thresholds for verb-noun and adjective-noun phrases were optimized independently using the corresponding development sets. We experimented with the three phrase-level scoring methods on the development sets, and found that PHRASCOS1 consistently outperformed PHRASCOS2 and PHRASCOS3 for both verb–noun and adjective–noun phrases. We thus report results for PHRASCOS1 on our test sets.

Features	Method	P	R	$F1$
Linguistic	WORDCOS	0.67	0.76	**0.71**
	PHRASCOS1	0.38	0.94	0.54
Visual	WORDCOS	0.49	0.97	0.65
	PHRASCOS1	0.56	0.79	0.66
Multimodal	WORDMID	0.56	0.86	0.68
	PHRASMID	0.44	0.93	0.59
	WORDLATE	0.49	0.96	0.65
	PHRASLATE	0.41	0.92	0.57
	MIXLATE	0.65	0.87	**0.75**

Table 1: System performance on Mohammad et al. dataset (MOH) in terms of precision (P), recall (R) and F-score ($F1$)

We first evaluated the performance of WORDCOS and PHRASCOS1 using linguistic and visual representations in isolation, and then evaluated the multimodal models using middle and late fusion strategies. In middle fusion, we concatenated the linguistic and visual vectors, and then applied WORDCOS and PHRASCOS1 methods to the resulting multimodal vectors. We will refer to these methods as WORDMID and PHRASMID respectively. In late fusion, we used an average of linguistic and visual scores to determine metaphoricity. We experimented with three different scoring methods: (1) WORDLATE, where linguistic and visual WORDCOS scores were combined; (2) PHRASLATE, where linguistic and visual PHRASCOS1 scores were combined; and (3) MIXLATE, where linguistic and WORDCOS and visual PHRASCOS1 scores were combined.

4.3 Results and discussion

We evaluated the performance of our methods on the MOH and TSV-TEST test sets in terms of precision, recall and F-score and the results are presented in Tables 1 and 2 respectively. When using linguistic embeddings alone, WORDCOS outperforms

Features	Method	P	R	$F1$
Linguistic	WORDCOS	0.73	0.80	**0.76**
	PHRASCOS1	0.43	0.96	0.57
Visual	WORDCOS	0.50	0.95	0.66
	PHRASCOS1	0.60	0.91	**0.73**
Multimodal	WORDMID	0.59	0.85	0.70
	PHRASMID	0.54	0.93	0.68
	WORDLATE	0.69	0.72	0.70
	PHRASLATE	0.50	1.00	0.67
	MIXLATE	0.67	0.96	**0.79**

Table 2: System performance on Tsvetkov et al. test set (TSV-TEST) in terms of precision (P), recall (R) and F-score ($F1$)

PHRASECOS1 for both verbs and adjectives by 17-19%. This suggests that linguistic word embeddings already successfully capture domain and compositional information necessary for metaphor identification. In contrast, the visual PHRASECOS1 model, when applied in isolation, tends to outperform the visual WORDCOS model. PHRASCOS1 measures to what extent the meaning of the phrase can be composed by simple combination of the representations of individual words. In metaphorical language, however, a meaning transfer takes place and this is no longer the case. Particularly in visual data, where no linguistic conventionality and stylistic effects take place, PHRASCOS1 captures this property. For adjectives this trend was more evident than for verbs. The visual PHRASECOS1 model, even when applied on its own, attains a high F-score of 0.73 on TSV-TEST, suggesting that concreteness and other visual features are highly informative in identification of adjectival metaphors. This effect was present, though not as pronounced, for verbal metaphors, where the vision-only PHRASECOS1 attains an F-score of 0.66.

The multimodal model, integrating linguistic and visual embeddings, outperforms the linguistic models for both verbs and adjectives, clearly demonstrating the utility of visual features across word classes. The late fusion method MIXLATE, which combines the linguistic WORDCOS score and the visual PHRASECOS1, attains an F-score of 0.75 for verbs and 0.79 for adjectives, which makes it best-performing among our fusion strategies. When the same type of scoring (i.e. either WORDCOS or PHRASCOS1) is used with both linguistic and visual embeddings, middle and late fusion techniques attain comparable levels of performance, with WORDCOS being the leading measure. The reason behind the higher performance of MIXLATE is likely to be the combination of different scoring methods, one of which is more suitable for the linguistic model and the other for the visual one.

The differences between verbs and adjectives with respect to the utility of visual information can be explained by the following two factors. Firstly, previous psycholinguistic research on abstractness and concreteness (Hill et al., 2014) suggests that humans find it easier to judge the level of concreteness of adjectives and nouns than that of verbs. It is thus possible that visual representations capture the concreteness of adjectives and nouns more accurately than that of verbs. Besides concreteness, it is also likely that perceptual properties in general are more important for the semantics of nouns (e.g. objects) and adjectives (their attributes), than for the semantics of verbs (actions), since the latter are grounded in our motor activity and not merely perception. Secondly, following the majority of multimodal semantic models, we used images as our visual data rather than videos. However, some verbs, e.g. stative verbs and verbs for continuous actions, may be better captured in video than images. We thus expect that using video data along with the images as input to the acquisition of visual embeddings is likely to improve metaphor identification performance for verbal metaphors. However, we leave the investigation of this issue for future work.

In an additional experiment, we evaluated our methods on the larger TSV-TRAIN dataset (specifically using its portion that was not employed for development purposes) and the trends observed were the same. MIXLATE attained an F-score of 0.71, outperforming language-only and vision-only models. The performance of all scoring methods on TSV-TRAIN was lower than that on the TSV-TEST. This may be the result of the fact that the labelling of TSV-TRAIN was less consistent than that of TSV-TEST. As TSV-TEST is a set of metaphors annotated by 5 annotators with a high agreement, the evaluation on TSV-TEST is likely to be more reliable (Tsvetkov et al., 2014).

It is important to note that, unlike other supervised approaches to metaphor, our methods do not

require large training sets to learn the respective thresholds. The results reported here were obtained using only 80 annotated examples for training. This is sufficient since the necessary lexical knowledge and the knowledge about domain, concreteness and visual properties of concepts is already captured in the linguistic and visual embeddings. However, we additionally investigated how stable the thresholds learned by the model are using the TSV-TRAIN dataset. For this purpose, we divided the dataset into 10 portions of approximately 170 examples (balanced for metaphoricity). We then trained the thresholds first on a small set of 170 examples and then increasing the dataset by 170 examples at each round. The thesholds appear to be relatively stable, with a standard deviation of 0.03 for MIXLATE; 0.02 for WORDCOS (linguistic); and 0.05 for PHRASECOS1 (visual). This suggests that our methods do not require a large annotated dataset and training on a small number of examples is sufficient.

Despite the limited need in training data and no reliance on hand-coded lexical resources, the performance of our method favourably compares to that of existing metaphor identification systems (Turney et al., 2011; Neuman et al., 2013; Gandy et al., 2013; Dunn, 2013b; Tsvetkov et al., 2013; Hovy et al., 2013; Hovy et al., 2013; Shutova and Sun, 2013; Strzalkowski et al., 2013; Beigman Klebanov et al., 2015), that typically use such resources. For instance, Turney et al. (2011) used hand-annotated abstractness scores for words to develop their system, and reported an F-score of 0.68 for verb–noun metaphors and an accuracy of 0.79 for adjective–noun metaphors (though the latter was only evaluated on a small dataset of 10 adjectives and Turney and colleagues did not report results in terms of F-score, which is likely to be lower). Our use of visual features is in line with Turney's hypothesis concerning the relevance of concreteness features to metaphor processing. However, our results indicate that extracting this information from image data directly is a more suitable way to capture the concreteness itself, as well as capturing other relevant perceptual properties of concepts. The method of Tsvetkov et al. (2014) used both concreteness features (which they extracted from the MRC concreteness database) and hand-coded domain information for words (which they extracted from WordNet). They report a high F-score of 0.85 for adjective–noun classification on TSV-TEST. The performance of our method on the same dataset is a little lower than that of Tsvetkov et al. However, we do not use any hand-annotated resources and acquire linguistic, domain and perceptual information in the data-driven way. It is thus encouraging that, even though resource-lean, our methods approach the performance level of the methods using hand-annotated features (as in case of Tsvetkov et al. (2014)) or outperform them (as in case of Turney et al. (2011), Neuman et al. (2013), Dunn (2013b), Mohler et al. (2013), Gandy et al. (2013), Strzalkowski et al. (2013), Beigman Klebanov et al. (2015) and many others). For further comparison with these approaches and their results see a recent review by Shutova (2015).

5 Conclusion

We presented the first method that uses visual features for metaphor identification. Our results demonstrate that the multi-modal model combining both linguistic and visual knowledge outperforms language-only models, suggesting the importance of visual information for metaphor processing. Unlike previous metaphor processing approaches, that employed hand-crafted resources to model perceptual properties of concepts, our method learns visual knowledge from images directly, thus reducing the risk of human annotation noise and having a wider coverage and applicability. Since the method relies on automatically acquired lexical knowledge, in the form of linguistic and visual embeddings, and is otherwise resource-independent, it can be applied to unrestricted text in any domain and easily tailored to other metaphor processing tasks.

In the future, it would be interesting to apply multimodal word and phrase embeddings to automatically interpret metaphorical language, e.g. by deriving literal or conventional paraphrases for metaphorical expressions (similarly to the task of Shutova (2010)). Multimodal embeddings are also likely to provide useful information for the models of metaphor translation, as they have already proved successful in bilingual lexicon induction more generally (Kiela et al., 2015b). Finally, it would be interest-

ing to further investigate compositional properties of metaphorical language using multimodal phrase embeddings and to apply the embeddings to automatically generalise metaphorical associations between distinct concepts or domains.

Acknowledgment

We are grateful to the NAACL reviewers for their helpful feedback. Ekaterina Shutova's research is supported by the Leverhulme Trust Early Career Fellowship. Douwe Kiela is supported by EPSRC grant EP/I037512/1.

References

Lawrence W. Barsalou. 2008. Grounded cognition. *Annual Review of Psychology*, 59(1):617–645.

Beata Beigman Klebanov, Chee Wee Leong, and Michael Flor. 2015. Supervised word-level metaphor detection: Experiments with concreteness and reweighting of examples. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 11–20, Denver, Colorado, June. Association for Computational Linguistics.

Shane Bergsma and Randy Goebel. 2011. Using visual information to predict lexical preference. In *RANLP*, pages 399–405.

Danushka Bollegala and Ekaterina Shutova. 2013. Metaphor interpretation using paraphrases extracted from the web. *PLoS ONE*, 8(9):e74304.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *J. Artif. Intell. Res.(JAIR)*, 49:1–47.

Jonathan Dunn. 2013a. Evaluating the premises and results of four metaphor identification systems. In *Proceedings of CICLing'13*, pages 471–486, Samos, Greece.

Jonathan Dunn. 2013b. What metaphor identification systems can tell us about metaphor-in-language. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 1–10, Atlanta, Georgia.

Jerome Feldman. 2006. *From Molecule to Metaphor: A Neural Theory of Language*. The MIT Press.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database (ISBN: 0-262-06197-X)*. MIT Press, first edition.

Robert Fergus, Li Fei-Fei, Pietro Perona, and Andrew Zisserman. 2005. Learning object categories from google's image search. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1816–1823. IEEE.

Lisa Gandy, Nadji Allan, Mark Atallah, Ophir Frieder, Newton Howard, Sergey Kanareykin, Moshe Koppel, Mark Last, Yair Neuman, and Shlomo Argamon. 2013. Automatic identification of conceptual metaphors with limited knowledge. In *Proceedings of AAAI 2013*.

Matt Gedigian, John Bryant, Srini Narayanan, and Branimir Ciric. 2006. Catching metaphors. In *In Proceedings of the 3rd Workshop on Scalable Natural Language Understanding*, pages 41–48, New York.

Ilana Heintz, Ryan Gabbard, Mahesh Srivastava, Dave Barner, Donald Black, Majorie Friedman, and Ralph Weischedel. 2013. Automatic extraction of linguistic metaphors with lda topic modeling. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 58–66, Atlanta, Georgia.

Felix Hill, Anna Korhonen, and Christian Bentz. 2014. A quantitative empirical analysis of the abstract/concrete distinction. *Cognitive Science*, 38(1):162–177.

Dirk Hovy, Shashank Shrivastava, Sujay Kumar Jauhar, Mrinmaya Sachan, Kartik Goyal, Huying Li, Whitney Sanders, and Eduard Hovy. 2013. Identifying metaphorical word use with tree kernels. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 52–57, Atlanta, Georgia.

Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*.

Douwe Kiela and Léon Bottou. 2014. Learning Image Embeddings using Convolutional Neural Networks for Improved Multi-Modal Semantics. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-14)*.

Douwe Kiela and Stephen Clark. 2015. Multi- and cross-modal semantics beyond vision: Grounding in auditory perception. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2461–2470, Lisbon, Portugal, September. Association for Computational Linguistics.

Douwe Kiela, Laura Rimell, Ivan Vulić, and Stephen Clark. 2015a. Exploiting image generality for lexical entailment detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Beijing, China.

Douwe Kiela, Ivan Vulić, and Stephen Clark. 2015b. Visual bilingual lexicon induction with transferred convnet features. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal.

Saisuresh Krishnakumaran and Xiaojin Zhu. 2007. Hunting elusive metaphors using lexical resources. In *Proceedings of the Workshop on Computational Approaches to Figurative Language*, pages 13–20, Rochester, NY.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press, Chicago.

Omer Levy and Yoav Goldberg. 2014. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 171–180, Ann Arbor, Michigan, June.

Hongsong Li, Kenny Q. Zhu, and Haixun Wang. 2013. Data-driven metaphor recognition and explanation. *Transactions of the Association for Computational Linguistics*, 1:379–390.

Max M Louwerse. 2011. Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, 3(2):273–302.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Zachary Mason. 2004. Cormet: a computational, corpus-based conventional metaphor extraction system. *Computational Linguistics*, 30(1):23–44.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *Proceedings of ICLR*, Scottsdale, AZ.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of NAACL-HLT*, pages 746–751.

Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a medium for emotion: An empirical study. *Language Resources and Evaluation*, forthcoming.

Michael Mohler, David Bracewell, Marc Tomlinson, and David Hinote. 2013. Semantic signatures for example-based linguistic metaphor detection. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 27–35, Atlanta, Georgia.

Michael Mohler, Bryan Rink, David Bracewell, and Marc Tomlinson. 2014. A novel distributional approach to multilingual conceptual metaphor recognition. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, Dublin, Ireland.

Yair Neuman, Dan Assaf, Yohai Cohen, Mark Last, Shlomo Argamon, Newton Howard, and Ophir Frieder. 2013. Metaphor identification in large texts corpora. *PLoS ONE*, 8(4):e62343.

Stephen Roller and Sabine Schulte im Walde. 2013. A multimodal lda model integrating textual, cognitive and visual modalities. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1146–1157.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

Ekaterina Shutova and Lin Sun. 2013. Unsupervised metaphor identification using hierarchical graph factorization clustering. In *Proceedings of NAACL 2013*, Atlanta, GA, USA.

Ekaterina Shutova, Lin Sun, and Anna Korhonen. 2010. Metaphor identification using verb and noun clustering. In *Proceedings of Coling 2010*, pages 1002–1010, Beijing, China.

Ekaterina Shutova, Tim Van de Cruys, and Anna Korhonen. 2012. Unsupervised metaphor paraphrasing using a vector space model. In *Proceedings of COLING 2012*, Mumbai, India.

Ekaterina Shutova. 2010. Automatic metaphor interpretation as a paraphrasing task. In *Proceedings of NAACL 2010*, pages 1029–1037, Los Angeles, USA.

Ekaterina Shutova. 2015. Design and Evaluation of Metaphor Processing Systems. *Computational Linguistics*, 41(4).

Carina Silberer and Mirella Lapata. 2012. Grounded models of semantic representation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1423–1433. Association for Computational Linguistics.

Tomek Strzalkowski, George Aaron Broadwell, Sarah Taylor, Laurie Feldman, Samira Shaikh, Ting Liu, Boris Yamrom, Kit Cho, Umit Boz, Ignacio Cases, and Kyle Elliot. 2013. Robust extraction of metaphor from novel data. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 67–76, Atlanta, Georgia.

Yulia Tsvetkov, Elena Mukomel, and Anatole Gershman. 2013. Cross-lingual metaphor detection using common semantic features. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 45–51, Atlanta, Georgia.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection

with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 248–258, Baltimore, Maryland, June. Association for Computational Linguistics.

Peter D. Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 680–690, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tony Veale and Yanfen Hao. 2008. A fluid knowledge representation for understanding and generating creative metaphors. In *Proceedings of COLING 2008*, pages 945–952, Manchester, UK.

Yorick Wilks, Adam Dalton, James Allen, and Lucian Galescu. 2013. Automatic metaphor detection using large-scale lexical resources and conventional metaphor extraction. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 36–44, Atlanta, Georgia.

M.D. Wilson. 1988. The MRC Psycholinguistic Database: Machine Readable Dictionary, Version 2. *Behavioural Research Methods, Instruments and Computers*, 20:6–11.