Phrase-Level Metaphor Identification using Distributed Representations of Word Meaning

Omnia Zayed, John P. McCrae, Paul Buitelaar
Insight Centre for Data Analytics
Data Science Institute
National University of Ireland Galway
IDA Business Park, Lower Dangan, Galway, Ireland
firstname.lastname@insight-centre.org

Abstract

Metaphor is an essential element of human cognition which is often used to express ideas and emotions that might be difficult to express using literal language. Processing metaphoric language is a challenging task for a wide range of applications ranging from text simplification to psychotherapy. Despite the variety of approaches that are trying to process metaphor, there is still a need for better models that mimic the human cognition while exploiting fewer resources. In this paper, we present an approach based on distributional semantics to identify metaphors on the phrase-level. We investigated the use of different word embeddings models to identify verb-noun pairs where the verb is used metaphorically. Several experiments are conducted to show the performance of the proposed approach on benchmark datasets.

1 Introduction

Metaphor is a stylistic device used to enrich the language and represent abstract concepts using the properties of other concepts. It is considered as an analogy between a tenor (target concept) and a vehicle (source concept) by exploiting common similarities. The sense of a concept such as *"harmful plant"* can be transferred to another concept's sense such as *"poverty"* by exploiting the properties of the first concept. This then can be expressed in our everyday language in terms of linguistic metaphoric expressions such as *"...eradicate poverty"*, *"...root out the causes of poverty"*, or *"...the roots of poverty are..."*[1] (Lakoff and Johnson, 1980; Veale et al., 2016). In this work, a word or an expression is a metaphor if it has at least one basic/literal sense (more concrete, physical) and a secondary metaphoric sense (abstract, non-physical) which resonates semantically with the basic sense (Steen et al., 2010; Hanks, 2016).

Metaphor processing is one of the most challenging problems for many natural language processing tasks such as machine translation, text summarization and text simplification. Moreover, metaphor processing could be helpful for wider applications such as political discourse analysis (Charteris-Black, 2011) and psychotherapy (Witztum et al., 1988; Gutiérrez et al., 2017).

Understanding metaphors requires deeper levels of language processing that go beyond the sentence surface level. Among the main challenges of the computational modelling of metaphors is their pervasiveness in language which makes them occur frequently in everyday language. Moreover, metaphors are often conventionalised to such an extent that they exhibit no defined lexical patterns or signals. Previous approaches relies on extensive lexical resources to identify metaphors and to capture their semantic features. Feature extraction from an annotated corpus is a challenge as well, not only due to the complexity of the task itself but also due to the lack of high quality annotated corpora. The process of creating such a corpus depends on the task definition as well as the targeted application and often requires significant effort and time.

In this paper, we introduce a semi-supervised approach that makes use of distributed representations of word meaning to capture metaphoricity. We focus on identifying verb-noun pairs where the verb is used metaphorically. We extract verb-noun grammar relations using the Stanford parser (Chen and Manning, 2014). We then employ pre-trained word embeddings models to measure the semantic similarity between the candidate and a predefined seed set of metaphors. A similarity threshold, which was optimised on a sample dataset, is used to classify the given candidate. Evaluation

[1] These examples could be found in the United Nations Parallel Corpus (Ziemski et al., 2016).

Proceedings of the Workshop on Figurative Language Processing, pages 81–90
New Orleans, Louisiana, June 6, 2018. ©2018 Association for Computational Linguistics

of the presented approach was carried out on various test sets using different word embeddings algorithms. Additionally, a performance comparison is carried out against the results of the state-of-the-art approach on benchmark datasets.

2 Related Work

One of the most common tasks of the computational processing of metaphors is "metaphor identification" which is concerned with recognising (detecting) the metaphoric expressions in the input text. Metaphor detection could be done on the word-level (token-level) or on the phrase-level by extracting grammatical relations.

In this paper, we are interested in phrase-level linguistic metaphor detection, focusing on verb-noun phrases (grammatical relations) by employing semantic representation of word meaning. Therefore, due to space limitation, we will discuss the most relevant research in this regard in this section. An extensive literature review is presented in (Zhou et al., 2007; Shutova, 2015). Some recent work on metaphor detection has been looking into the utilization of semantic representations through word embeddings representations to design supervised systems for metaphor detection (Rei et al., 2017; Bulat et al., 2017; Shutova et al., 2016). Our approach also utilises such representations but in a semi-supervised manner to avoid the need for large training corpora.

Rei et al. (2017) introduced a neural network architecture to detect adjective-noun and verb-noun metaphoric constructions. Their system comprises three main components which are: word gating, vector representation mapping and a weighted similarity function. The word gating is used to model the association between the properties of the source and target domains which is done via a non-linear transformation of the word embeddings vectors of the given candidate pair. The word embeddings used in this step are obtained from a pre-trained model. Then, a vector representation mapping is carried out to prepare a "new metaphor-specific" vector space using the original word embeddings. Finally, a weighted cosine similarity function is used to automatically select the important vector dimensions for the metaphor detection task. The authors experimented with different pre-trained word representations, namely skip-gram model and an attribute-based model. Two different datasets, which were referred to as the TSV dataset (Tsvetkov et al., 2013) and the MOH dataset (Mohammad et al., 2016), were used to train the system and optimise its parameters as well as to assess its performance.

Bulat et al. (2017) is a recent approach that investigated whether property-based semantic word representation can provide better concept generalisation for detecting metaphors than dense linguistic representation. The authors proposed property-based vectors through cross-modal mapping between dense linguistic representations and a property-norm semantic space. The authors built a count-based distributional vector and employed a skip-gram model trained on Wikipedia articles as their dense linguistic representations. The property-norm semantic space is obtained from the property-norm dataset (McRae et al., 2005). The TSV dataset is used to train and test a support vector machine (SVM) classifier to classify adjective-noun pairs using the introduced cognitively salient properties as features.

An interesting approach, which employed multi-model embeddings of visual and linguistic features to detect metaphoricity in text, is introduced by Shutova et al. (2016). The proposed approach obtained linguistic word embeddings using a log-linear skip-gram model trained on Wikipedia text and obtained visual embeddings using a deep convolutional neural network trained on image data. This was done for both the words and phrases of adjective-noun and verb-noun pairs individually. Then, the cosine similarity function has been employed to measure the distance between the phrase vector and the corresponding vectors of its constituent words. Metaphor classification is done based on an optimised threshold output of the cosine similarity function. The authors used the TSV and the MOH datasets to train and test their system in addition to optimising the classification thresholds.

Modelling metaphor in a distributional semantic space through linear transformation to improve vector representation has been investigated by Gutiérrez et al. (2016). The authors introduced a compositional distributional semantic framework to identify adjective-noun metaphoric expressions.

A variety of lexical and semantic features including lexical abstractness and concreteness, imageability, named entities, part-of-speech tags, and the word's supersenses[2] using WordNet (Fell-

[2] the WordNet lexicographer name of the words first sense

baum, 1998) have been employed to develop supervised systems to detect metaphors (Köper and Schulte im Walde, 2017; Tsvetkov et al., 2013; Hovy et al., 2013; Turney et al., 2011).

Shutova et al. (2010) was among the earliest approaches to computational modelling of metaphor, avoiding task-specific hand-crafted knowledge and huge annotated resources. They introduced a semi-supervised approach to identify verb-noun metaphors using corpus-driven distributional clustering. Their strategy is based on clustering abstract nouns based on their contextual features in order to capture the metaphorical senses associated with the source concept. The system exploits a small set of metaphoric expressions as a seed to detect metaphors in a semi-supervised manner. In a follow-up work, Shutova and Sun (2013) investigated the use of hierarchical graph factorization clustering to derive a network of concepts in order to learn metaphorical associations in an unsupervised way which then was used as features to identify metaphors. We consider the work introduced by Shutova et al. (2010) as a baseline for our proposed approach, thus we are going to explain its reimplementation details in subsection 3.3.

Birke and Sarkar (2006) introduced TroFi, which is considered the first statistical system to identify the metaphorical senses of verbs in a semi-supervised way. The authors adapted a statistical similarity-based word sense disambiguation approach to cluster literal and non-literal senses. A predefined set of seed sentences is utilised to compute the similarity between a given sentence and the seed sentences.

3 Methodology

The idea behind our approach is based on finding synonyms and near-synonyms of metaphors. Our approach employs vector representation and semantic similarity to classify verb-noun pairs extracted from a sentence using a parser as potential candidates for metaphoric classification. A candidate is classified as a metaphor or not by measuring its semantic similarity to a predefined small seed set of metaphors which acts as our existing known metaphors sample. Metaphoric classification is performed based on a previously calculated similarity threshold value on a development dataset. The following subsections explain the hypothesis behind this work and our proposed approach in addition to the reimplementation of the state-of-the-art semi-supervised system used as our baseline system.

3.1 Hypothesis

Our hypothesis in this work is that a given candidate should have common characteristics and semantic features with some positive examples of metaphors. However, simply calculating the similarity between a given verb-noun candidate and a metaphoric seed is not enough due to the effect of each of the verb and the noun on the overall similarity score. For example, consider a metaphoric seed such as *"break agreement"* and two given candidates such as *"break promise"* and *"break glass"*. The semantic similarities between the word embeddings vectors of the seed and the two candidates measured by the cosine similarity function are 0.5304 and 0.6376, respectively, using a pre-trained Word2Vec (Mikolov et al., 2013) word embedding model on the Google News dataset. This indicates that both candidates are similar to the seed and there is not enough information to tell which one should be classified as a metaphor. Table 1 shows the similarity values of the two candidates and the most similar metaphoric seeds from the predefined seed set. We decided to look into the individual words of the candidate considering the fact that semantically similar or related words will be placed near each other in the embeddings space while unrelated words will be far apart. Therefore, we expect that the noun *"promise"* will be in the neighbourhood of *"agreement"* in the semantic space, while *"glass"* will not. So if both candidates share similar verbs, classification could be done based on the similarity of the nouns; in that case, *"break promise"* can be classified as metaphor due to the vicinity of its noun to the noun of the metaphoric seed while *"break glass"* will not. Since using one positive (metaphoric) example is not enough for precise classification, we used a small set of verb-noun pairs, hereafter referred to as the seed set, where the verb is used metaphorically. The specification of the seed set will be explained in detail in section 4.

3.2 Approach

We start with the seed set of metaphoric verb-noun pairs as $S = \{(V, N)\}$. Given a target verb-noun candidate (v_t, n_t) that needs to be classified, we calculate the distance between every verb v_s in S and the verb of the candidate v_t using the cosine distance measure as follows:

Candidate	Metaphoric Seed	Cosine Similarity	Candidate	Metaphoric Seed	Cosine Similarity
break promise	break agreement	0.6376	break glass	break agreement	0.5304
	hold back truth	0.4560		hold back truth	0.3435
	fix term	0.3653		frame question	0.3109
	spell out reason	0.3385		face hour	0.2949
	seize moment	0.3384		block out thought	0.2701
	glimpse duty	0.3224		seize moment	0.2677
	grasp term	0.3019		throw remark	0.2583
	frame question	0.2959		skim over question	0.2509
	accelerate change	0.2927		mend marriage	0.2375
	throw remark	0.2776		spell out reason	0.2354

Table 1: The cosine similarity between the candidates "break promise" and "break glass" and the top 10 metaphoric seeds in the seed set using a pre-trained Word2Vec word embedding model on Google News dataset.

$$D_{ts} = d(v_t, v_s) \ \ \forall v_s \in S$$

This gives a list of verbs ranked according to the distance to the verb of the candidate; we then select the top n nearest verbs and we get the nouns associated with them in the seed set as follows:

$$Y_{v_t} = top_n\{n_s : (v_s, n_s) \in S\} \text{ by } D_{ts}$$

Finally, the average of the distances between these nouns and the target noun in the candidate phrase is calculated. If this average is less than a threshold δ then the candidate phrase will be classified as a metaphoric expression as follows:

$$\frac{1}{|Y_{v_t}|} \sum_{n_s \in Y_{v_t}} [d(n_t, n_s)] \leq \delta$$

Table 2 shows the cosine distance between the verbs and the nouns of the candidates *"break promise"* and *"break glass"* verses the verbs and the nouns of the top 10 metaphoric seeds from the seed set using a pre-trained Word2Vec word embedding model on the Google News dataset; those 10 seeds have the most similar (nearest in terms of distance) verbs to the candidate verb.

3.3 Baseline

We consider the system introduced by Shutova et al. (2010) as our baseline system. In this subsection, we are going to explain in detail the reimplementation of this approach and the related findings. The system consists of four main components which are: a seed set, a clustering component, a candidate extraction component, and a filtering component. The seed set is obtained from the British National Corpus (BNC) (Burnard, 2009) and consists of 62 metaphoric verb-noun pairs (more details are given in section 4). Spectral clustering (Meila and Shi, 2001) is used to cluster the abstract concepts (nouns) and the concrete concepts (verbs) then an association (mapping) is drawn between the two clusters using the seed set. The candidate extraction component employs the Robust Accurate Statistical Parsing (RASP) parser (Briscoe et al., 2006) to extract verb-subject and verb-direct object grammar relations. After that, the linked clusters (through the seed set) is used to identify potential metaphoric candidates. The filtering component is finally used to filter out these candidates based on a selectional preferences strength (SPS) measure (Resnik, 1993). The verbs exhibiting weak selectional preferences are considered to have lower metaphorical potential. An SPS threshold was set experimentally to be 1.32, thus, the candidates which verbs have an SPS value below this threshold are discarded.

In our reimplementation, we employed the Stanford Parser instead of the RASP Parser to extract the grammar relations and to implement the filtering component to calculate the SPS. SPS is calculated using a simplified Resnik model which models the association of the verb (predicate) with the noun (instead of a class) from the BNC corpus. The verb clusters were originally developed using VerbNet (Schuler, 2006) and the noun clustering were developed using the 2,000 most frequent nouns in the BNC corpus. Since the clusters were obtained from a relatively small dataset we suspected that it might lead to a limited coverage, which will be later shown in the system evaluation.

Cand. V	Seed's V	CosDist	Cand. N	Seed's N	CosDist	Cand. N	Seed's N	CosDist
break	break	0	promise	agreement	0.7479	glass	agreement	1.0093
	hold back	0.6591		truth	0.7736		truth	0.8872
	mend	0.6935		marriage	0.9381		marriage	0.9419
	fix	0.6952		term	0.8085		term	1.0252
	catch	0.6966		contagion	1.0126		contagion	0.9089
	throw	0.7035		remark	0.8513		remark	0.9559
	seize	0.7201		moment	0.8556		moment	0.9510
	impose	0.7350		decision	0.8207		control	0.9506
	impose	0.7350		control	0.9107		decision	0.9987
	frame	0.7371		question	0.8462		question	0.9424

Table 2: The cosine distance between the verbs and nouns of the candidates "break promise" and "break glass" verses the verbs and the nouns of the top 10 metaphoric seeds in the seed set using a pre-trained Word2Vec word embedding model on Google News dataset.

This is one of the limitations of this system; a candidate is either in the clusters or not. And if the candidate's noun appeared in a noun cluster but this cluster was not mapped to the cluster where the verb occurs the candidate will be discarded.

4 System Architecture

As described in Figure 1 below, our system consists of three main components: a parser, a seed set of metaphoric expressions and a pre-trained word embedding model.

Parser: Since our aim is to identify metaphors on the phrase-level, the Stanford parser is used to extract the grammar relations in a given sentence. We used the recurrent neural network (RNN) parser in the Stanford CoreNLP toolkit (Manning et al., 2014) to extract dependencies focusing on verb-subject and verb-direct object grammar relations.

Seed Set: We used the seed set of Shutova et al. (2010) to act as our set of existing known metaphoric expressions (positive examples). The seed set consists of 62 verb-subject and verb-direct object phrases where the verb is used metaphorically[3]. These seeds are extracted originally from a subset of the BNC corpus which contains 761 sentences. These sentences were annotated for grammatical relations to extract the specified grammar relations which are then filtered and manually annotated for metaphoricity. Examples of the metaphors in the seed set are *"mend marriage, break agreement, cast doubt, and stir excitement"*.

Word Embedding Model: This work utilises distributional vector representation of word meaning to calculate semantic similarity between a candidate and a seed set. Word2Vec and GloVe (Pennington et al., 2014) are two widely used word embeddings algorithms to construct embeddings vectors based on the distributional hypothesis (Firth, 1957) but using different machine learning techniques. In this work, we investigated the effect of using different pre-trained models and similarity measures as shown in detail in the next section.

5 Experimental Settings

In this section, we give an overview of the experimental settings of our proposed approach and the test sets that are used to assess the performance of the methodology described above.

5.1 Models and Parameters

The utilised similarity measures, word embeddings models, and system's parameters are defined as follows:

Similarity Measures: We examined two similarity measures as follows:

- Cosine Distance Metric: The cosine similarity function measures the cosine of the angle between two vectors. Given the vectors u and v, the cosine distance can be defined as:

$$1 - \cos(u, v)$$

[3] The seed set provided to us by Shutova et al. (2010) consists of 52 pairs out of which 11 are verb-subjects and 41 are verb-direct object

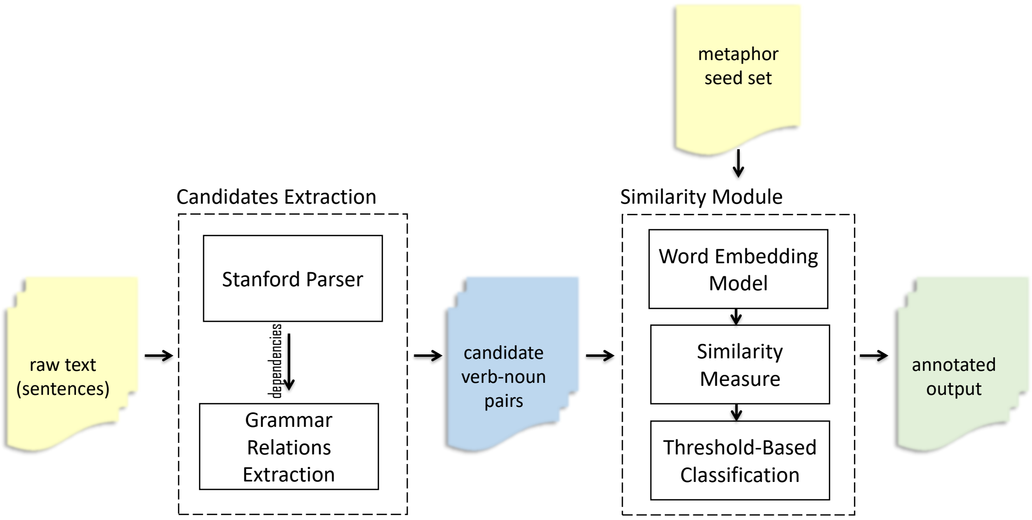

Figure 1: The Overall System Architecture.

- Word Mover's Distance (WMD) (Kusner et al., 2015): could be defined as the minimum travelling distance from one word embeddings vector to the other.

Embeddings Models: We experimented with two different pre-trained vector representations of word embeddings which are:

- Word2Vec Google News[4]: The model is trained on about 100 billion words from the Google News dataset and contains 300-dimensional vectors for 3 million words using the approach described in (Mikolov et al., 2013). The model is based on the skip-gram neural network architecture which employs the negative sampling training algorithm and sub-sampling frequent words using a window-size of 10.

- GloVe Common Crawl[5]: We used a pre-trained model on the Common Crawl dataset containing 840 billion tokens of web data (about 2 million words). The vectors are 300-dimensional using 100 training iteration.

For simplicity, we used a single vector representation for each word ignoring multi-word combinations such as phrasal verbs, examples of which include e.g. *"hold back, flip through"*; we are planing to address this issue in the future.

System's Parameters: We performed experiments on a development set to select the values of the parameters top_n and δ mentioned in subsection 3.2. The best value obtained for n is found to be top 10 nearest verbs. The suitable distance average threshold δ is found to be 0.80 for the GloVe Creative-Commons-840 model and 0.85 for the Word2Vec Google-News model. These values give a good trade-off between false positives and false negatives.

5.2 Test Sets

Two different test sets are used to evaluate our approach as follows:

VUA Test Set: We use a subset of the training verbs dataset from the VU Amsterdam Metaphor Corpus (VUA) (Steen et al., 2010) provided by the NAACL 2018 Metaphor Shared Task [6]. The original VUA corpus is a subset of the BNC Baby corpus consists of 117 texts covering various genres which are academic, conversation, fiction, and news. Although the dataset is annotated on the token-level, its availability and the fact that it is

[4] https://code.google.com/archive/p/word2vec/

[5] https://nlp.stanford.edu/projects/glove/

[6] https://github.com/EducationalTestingService/metaphor/tree/master/NAACL-FLP-shared-task

already annotated encouraged us to use it for assessing our approach. The verbs dataset consists of around 17,240 annotated verbs; we retrieved the original sentences of these verbs from the VUA corpus, which yielded around 8,000 sentences. We then parsed these sentences using the Stanford Parser and extracted around 5,000 verb-direct object relations. Arbitrary 300 verb-noun pairs (160 positive and 145 negative examples) are selected to be our test set where the verb is used metaphorically or literally. Table 3 shows some examples from this test set.

MOH dataset: Shutova et al. (2016) introduced a manually annotated dataset of verb-subject and verb-object pairs. The dataset has been referred to as MOH as it was originally obtained from Mohammad et al. (2016) who annotated different senses of verbs in WordNet for metaphoricity. Verbs were selected if they have more than three senses and less than ten senses. Then the example sentences from WordNet for each verb were extracted and annotated by 10 annotators using crowd-sourcing. In a next step, the verb-subject and verb-direct object grammar relations were extracted out of the original dataset. The final dataset consists of 647 pairs out of which 316 instances are metaphorical and 331 instances are literal.

Metaphor	**Not Metaphor**
reveal approach	collect passport
break corporation	use power
make money	abolish power
see language	perform shuffle
make error	decorate wall
face criticism	put stage
give access	read book
lay foundation	research joke
make time	tell story
abuse status	give key

Table 3: Examples from the VUA test set.

6 Evaluation

In this section, we evaluate our approach using different test sets, pre-trained word embeddings models and similarity measures. Additionally, we compare the performance of our approach against the baseline system explained in subsection 3.3. We used four standard evaluation metrics, namely precision, recall, F-score and accuracy.

6.1 Results

We applied our system to the three test sets introduced above and compared it to the defined baseline system. Table 4 shows the results of the experiment carried out on the VUA test set. It also shows the results obtained from the baseline system. Table 5 shows the performance of our system on the whole MOH dataset.

6.2 Discussion and Analysis

It can be seen from the results above that our approach performs better using GloVe as the pre-trained word embedding model and using cosine distance as the similarity metric. It is also noted that the system suffers from a low recall when using the Word2Vec model with the cosine distance function. This might be due to the limited coverage of the seed set where the top 10 most similar metaphors are not enough to detect new candidates of metaphors. We manually examined our system's output on the MOH dataset. Our system was able to correctly detect metaphoric expressions such as *"absorb knowledge, attack cancer, blur distinction, buy story, capture essence, swallow word, visit illness, wear smile"* as well as literal ones such as *"attack village, build architect, leak container, steam ship, suck poison"*. Some of the false positives, where our system detection was metaphor while the gold label was not, include *"ascend path, blur vision, buy love, communicate anxiety, jam mechanism, lighten room, line book, push crowd"* which could be regarded as metaphors depending on the context.

Our system was able to spot some inconsistency in the annotations of the VUA test set. For example, the verb-noun pair *"win election"* is detected as metaphor by our system while we realised that it has 3 different annotations across the rest of the VUA dataset (the verb *"win"* annotated once as a metaphor and twice as not metaphor while having *"election"* as its direct object). Additionally, in the VUA corpus the verb *"win"* is annotated as metaphor with similar abstract concepts such as in *"win match"* and *"win bid"*. This is one of the differences between preparing a dataset for word-level detection as the VUA corpus or preparing a dataset for phrase-level detection. Moreover, it shows that a verb-noun pair may or may not be metaphoric based on the context. Also, it highlights the minor differences in the views of the

			Precision	**Recall**	**F–score**	**Accuracy**
Shutova et al. (2010) distributional clustering approach			**0.7500**	0.0197	0.0385	0.4915
Our approach	Word2Vec	WMD	0.556	0.8487	0.6719	0.5729
		cosine distance	0.7455	0.2697	0.3961	0.5763
	GloVe	WMD	0.5565	**0.9079**	0.6900	0.5797
		cosine distance	0.6377	0.8684	**0.7354**	**0.6780**

Table 4: Evaluation on the VUA test set of 300 verb-noun pairs and a performance comparison to the baseline system.

			Precision	**Recall**	**F–score**	**Accuracy**
Shutova et al. (2010) distributional clustering approach			**1.0000**	0.0095	0.0189	0.5148
Our approach	Word2Vec	WMD	0.5321	0.8413	0.6519	0.5599
		cosine distance	0.8727	0.1524	0.2595	0.5739
	GloVe	WMD	0.5243	**0.8571**	0.6506	0.5490
		cosine distance	0.6317	0.7460	**0.6841**	**0.6625**

Table 5: Evaluation on the MOH dataset of 647 verb-noun pairs and a performance comparison to the baseline system.

definition of metaphor itself between Lakoff and Johnson (1980) and Steen et al. (2010), which in turn emphasises that the metaphorical sense does not depend solely on the properties of individual words (Gutiérrez et al., 2016).

The results also indicate that the baseline system has a very low recall on the introduced test sets. The reason behind that, as mentioned in subsection 3.3, is that it utilises clusters developed using the BNC corpus, which likely limit the coverage of the system adding into account the limitation of the small seed set (as in our approach). For example, out of the 300 pairs in the VUA test set only 7 candidates were included in the final classification as the rest of the words were not seen before in the clusters. Similarly, out of the 647 pairs in the MOH dataset only 4 were able to be recognised as candidates.

Our system's performance could be improved by increasing the size of the seed set and optimising the system's parameters accordingly (which we are planing to address in the future). In order to investigate this point, we did an additional experiment using 10-fold cross-validation of the MOH dataset in which we included 10 different splits from the dataset as our seed set of metaphors. The best results in terms of precision, recall, F-score, and accuracy are 0.5945, 0.756, 0.6657, and 0.6290, respectively. These results are obtained using the GloVe word embedding model pre-trained on the Common Crawl dataset and the cosine distance as similarity function with the same parameters values. In this experiment, we noticed that the values of n and the threshold δ should be adapted according to the increase in the number of seeds.

We did not to compare our results to Shutova et al. (2016) or Rei et al. (2017) as these systems are not directly comparable to ours. Shutova et al. (2016) is using a different test split from the MOH dataset to evaluate their system. Moreover, both works proposed fully supervised approaches in which they utilise negative (literal) examples as well as positive (metaphoric) examples to train their systems, whereas our approach is semi-supervised (similar to (Shutova et al., 2010)) which uses only the positive (metaphoric) examples. Therefore, carrying out a performance comparison will be imperfect.

7 Conclusion and Future Work

In this work, we presented a semi-supervised approach to detect metaphors using distributional representation of word meaning. Different word embeddings models have been investigated to identify phrase-level metaphors focusing on verb-noun expressions. The system utilises a predefined seed set of metaphoric expressions to detect unseen metaphoric expression(s) in a given sentence. As discussed, in contrast to other state-of-the-art approaches, our proposed approach employs fewer lexical resources and does not require annotated datasets or highly-engineered features. This gives it a flexibility to be easily adapted to new languages or text types. We have performed several experiments to assess the performance of our approach on benchmark datasets. As part of our future work, we are planning to investigate the effect of increasing the number of seeds on the system's coverage and to extend this approach to detect other metaphoric syntactic constructions taking into account multi-word expressions such as phrasal verbs.

Acknowledgments

This work was supported by Science Foundation Ireland under Grant Number SFI/12/RC/2289 (Insight).

References

Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '06, pages 329–336, Trento, Italy.

Ted Briscoe, John Carroll, and Rebecca Watson. 2006. The second release of the RASP system. In *Proceedings of the Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics*, COLING-ACL '06, pages 77–80, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Lou Burnard. 2009. About the British National Corpus. `http://www.natcorp.ox.ac.uk/corpus/index.xml`.

Jonathan Charteris-Black. 2011. Metaphor in Political Discourse. In *Politicians and Rhetoric: The Persuasive Power of Metaphor*, pages 28–51. Palgrave Macmillan UK, London.

Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP '14, pages 740–750, Doha, Qatar. Association for Computational Linguistics.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

John R. Firth. 1957. A synopsis of linguistic theory 1930-55. *Studies in Linguistic Analysis (special volume of the Philological Society)*, 1952-59:1–32.

E. Darío Gutiérrez, Guillermo A. Cecchi, Cheryl Corcoran, and Philip Corlett. 2017. Using automated metaphor identification to aid in detection and prediction of first-episode schizophrenia. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, EMNLP '17, pages 2923–2930, Copenhagen, Denmark.

E. Darío Gutiérrez, Ekaterina Shutova, Tyler Marghetis, and Benjamin Bergen. 2016. Literal and metaphorical senses in compositional distributional semantic models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 183–193, Berlin, Germany. Association for Computational Linguistics.

Patrick Hanks. 2016. Three kinds of semantic resonance. In *Proceedings of the 17th EURALEX International Congress*, pages 37–48, Tbilisi, Georgia.

Dirk Hovy, Shashank Srivastava, Sujay Kumar Jauhar, Mrinmaya Sachan, Kartik Goyal, Huiying 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 52–56, Atlanta, Georgia. Association for Computational Linguistics.

Maximilian Köper and Sabine Schulte im Walde. 2017. Improving verb metaphor detection by propagating abstractness to words, phrases and individual senses. In *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, SENSE '18, pages 24–30, Valencia, Spain.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *ICML'15*, pages 957–966, Lille, France.

George Lakoff and Mark Johnson. 1980. *Metaphors we Live by*. University of Chicago Press, Chicago, USA.

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, Baltimore, Maryland, USA. Association for Computational Linguistics.

Ken McRae, George S. Cree, Mark S. Seidenberg, and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37(4):547–559.

Marina Meila and Jianbo Shi. 2001. A random walks view of spectral segmentation. In *Proceedings of the 8th International Workshop on Artificial Intelligence and Statistics*, AISTATS 2001, Florida, USA.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, Lake Tahoe, Nevada, USA. Curran Associates Inc.

Saif M. Mohammad, Ekaterina Shutova, and Peter D. Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the 5th Joint Conference on Lexical and Computational Semantics*, *Sem '16, pages 23–33, Berlin, Germany.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP '14, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Marek Rei, Luana Bulat, Douwe Kiela, and Ekaterina Shutova. 2017. Grasping the finer point: A supervised similarity network for metaphor detection. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, EMNLP '17, pages 1537–1546, Copenhagen, Denmark. Association for Computational Linguistics.

Philip Stuart Resnik. 1993. *Selection and Information: A Class-based Approach to Lexical Relationships*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA, USA.

Karin Kipper Schuler. 2006. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA, USA.

Ekaterina Shutova. 2015. Design and evaluation of metaphor processing systems. *Computational Linguistics*, 41(4):579–623.

Ekaterina Shutova, Douwe Kiela, and Jean Maillard. 2016. Black holes and white rabbits: Metaphor identification with visual features. In *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '16, pages 160–170, San Diego, California, USA. The Association for Computational Linguistics.

Ekaterina Shutova and Lin Sun. 2013. Unsupervised metaphor identification using hierarchical graph factorization clustering. In *Proceedings of the 2013 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '13, pages 978–988, Atlanta, Georgia. The Association for Computational Linguistics.

Ekaterina Shutova, Lin Sun, and Anna Korhonen. 2010. Metaphor identification using verb and noun clustering. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 1002–1010, Beijing, China. Association for Computational Linguistics.

Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. Converging evidence in language and communication research. John Benjamins Publishing Company.

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. 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 2011 Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 680–690, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Tony Veale, Ekaterina Shutova, and Beata Beigman Klebanov. 2016. Metaphor: A Computational Perspective. *Synthesis Lectures on Human Language Technologies*, 9(1):1–160.

Eliezer Witztum, Onno van der Hart, and Barbara Friedman. 1988. The use of metaphors in psychotherapy. *Journal of Contemporary Psychotherapy*, 18(4):270–290.

Chang-Le Zhou, Yun Yang, and Xiao-Xi Huang. 2007. Computational mechanisms for metaphor in languages: A survey. *Journal of Computer Science and Technology*, 22(2):308–319.

Michal Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations Parallel Corpus v1.0. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, LREC '16, pages 3530–3534, Portoro, Slovenia.