Word Embedding Models for Metaphor Identification

Anonymous ACL submission

Abstract

Identification of metaphoric language in text is critical for generating accurate semantic representations. Previous approaches to automatic metaphor identification have largely relied on heuristic based models or feature-based machine learning, using hand-crafted lexical resources. While recent work has started to leverage deep learning through neural models, the benefits of word embeddings have yet to be fully realized for metaphor classification. This work shows word embeddings can be used along with random forest classification to achieve state of the art results that are robust across domain, effective for all parts of speech, and do not rely on any hand-crafted lexical resources.

1 Background

1.1 Metaphor Background

Metaphor is the conceptualization of an entity or concept through the structure of another. A typical metaphor is ARGUMENT IS WAR, in which ARGUMENT is structured through the domain of WAR:

1. He *defended* his position through his publications.

2. Her speech *attacked* his viewpoint.

The term *linguistic metaphor* is used to indicate these types of words and phrases. We will focus on linguistic metaphor, as identifying these utterances as metaphoric is critical for generating correct semantic interpretations for them. For instance, in the examples above, literal semantic interpretations of 'defend' and 'attack' will yield nonsensical utterances : a physical position cannot reasonably be defended by a publication, nor can a speech physically attack anything.[1]

There are many metaphor theories that have implications for automatic identification. The most prevalent is the cognitive framework of Lakoff and Johnson (1980), whose conceptual metaphor theory underlies most modern linguistic metaphor research. Conceptual metaphor theory holds that mappings are not purely linguistic, but present in all aspects of cognition, with linguistic metaphors being productions based on the conceptual metaphoric mappings.

The selectional preference theory of Wilks (1978) also carries significant computational implications. Wilks posits that metaphors are recognizable when there is a selectional preference mismatch between arguments. For example:

3. The car *drinks* gasoline.

In 1.1, we have the verb 'drink', which requires an animate agent. Given that in this utterance the agent is inanimate, we have an instance of linguistic metaphor. This theory provides a basis for many modern metaphor identification systems. Various methods are used to simulate selectional preferences for lexical items, which are used to fuel heuristic and machine-learning based approaches. However, there are many linguistic metaphors in which selectional preferences don't appear to be violated:

4. Djokovic *killed* Federer.

In this example, both arguments are people, thus satisfying the selectional preference of 'kill' by having an animate agent and an animate patient. But domain knowledge that the people are

[1] There are possible interpretations, such as a stack of publications standing in the way of an attacker, but they are certainly dispreferred to the metaphoric interpretation.

tennis rivals indicates that we should prefer the metaphoric reading, in which the domain of sports is understood through the lens of physical conflict. Despite these kinds of examples, this approach is promising computationally, as many linguistic metaphors do show evidence of selectional preference mismatch, and preferences can be modeled through lexical features.

1.2 Computational Approaches

There have been a wide variety of metaphor identification systems developed, starting as early as proposals by (Wilks, 1978), and the full interpretation system of (Martin, 1990). For full review of older computational metaphor systems see (Martin, 1996) and (Shutova, 2011). These early approaches often rely primarily on hand-crafted examples. More recently, there has been increasing interest in using corpus linguistics approaches to enable data driven analysis of what triggers metaphors, how they are produced and used, what conceptual metaphors may exist across languages, and how learning algorithms may be developed to detect and interpret them.

The VUAMC corpus (Steen et al., 2010) has been a vital resource for computational metaphor studies. This corpus contains approximately 200,000 words of text in four different registers (news, academics, fiction, conversations). The data is tagged at the token level: each word is tagged for metaphoricity, with additional subcategorization including possible metaphoric words ("When In Doubt, Leave It In"), markers for whether a word directly or indirectly indicates metaphor, and other related information. The corpus is also tagged for part of speech. The VUAMC has enabled the development of supervised machine learning algorithms, as well as providing an evaluation set for both supervised and unsupervised approaches.

Background The majority of automatic metaphor detection approaches have either been heuristic or feature-based machine learning algorithms. Many effective systems rely on databases of lexical features - imageability and concreteness from the MRC psycholinguistic database (Wilson, 1988) and the concreteness ratings of (Brysbaert et al., 2014) have been widely used, as well as values for valence, dominance, and arousal from the Affective Norms for English Words dataset (Bradley and Lang, 1999). For example, consider 5:

5. The woman *attacked* his beliefs.

In 5, 'attack' has an agent 'The woman' and a theme 'his beliefs'. 'woman' as a noun is both concrete and imageable, while 'beliefs' is neither.[2] The idea behind these approaches is that the differences in lexical features are indications that the words involved are used metaphorically. This example shows the importance of context : neither 'woman' nor 'viewpoint' is used metaphorically in this context. Rather, their mismatching lexical features indicate that the verb is metaphoric. This closely matches the theory of Wilks : the selectional preferences of the verb are violated, as 'attack' requires both a concrete and imageable agent as well as a concrete and imageable theme.

These types of features have recently led to significant advancements using supervised machine learning approaches on the VUAMC dataset. Dunn (2013) used a logistic regression model with features from the SUMO ontology, reporting an F1 of .455. Klebanov et al (2015) focus on achieving high recall through reweighting of lexical features, particularly concreteness. While their F1 is only .511, they achieved a much higher recall score (.670) than any other approach.

Haagsma et al (2016) use a selectional preference feature-based approach, focusing on verbs, reporting an F1 score of .58. Rai et al (2016) use a feature-heavy conditional random field (CRF) model, extending some of the previously noted lexical resources with WordNet, and while they focus on accuracy and precision, their F1 score reaches .609.

Since most of these approaches rely extensively on external hand-coded lexical information, gaps in coverage can affect performance when applied to actual data. For example, over 20 percent of tokens in the VUAMC database are not covered by the MRC concreteness or imageability rating. This makes the systems that are dependent on these resources difficult to apply to novel data and other languages, in addition to being difficult to develop and maintain.

[2] In the MRC, 'woman' has a concreteness value of 580 and imageability of 626 (scale from 100 to 700). 'beliefs' has a 328 for concreteness and 270 for imageability. Brysbaert has 'woman' at 4.46/5 for concreteness, with 'beliefs' at 1.19/5.

	F1	P	R	Model	Ev. Method
Dunn 2013	.455	.538	.394	Log Reg, SUMO Features	100 fold-CV
Klebanov 2015	.511	.438	**.670**	Concreteness/Imagability, feature reweighting	n-fold CV*
Dinh 2016	.561	.611	.536	MLP with pretrained embeddings	76 train, 12 dev, 12 test
Haagsman 2016	.578	-	-	Log Reg with selectional preference features	10-fold CV, Verbs Only
Rai 2016	**.609**	**.633**	.587	Feature-based CRF	10-fold CV

Table 1: Recent Results (*9-12 folds used for different registers)

They also have another drawback, illustrated by the "is a" construction from (Glucksberg, 2001):

6. That surgeon is a butcher.

In this example, the surgeon is said to have attributes of a butcher (such as disregard for life, clumsy slicing, etc). The utterance meaning here is not literal, but both arguments are highly concrete and highly imageable. [3] In fact, "is a" constructions seem to follow the opposite pattern : utterances in which one argument is significantly less concrete or imageable than the other seem to accept literal readings more easily:

7. A dog is an animal.

8. Gold is an element.

In the literal utterances 7 and 8, 'dog' is a more concrete than the abstract 'animal', and 'gold' is more concrete than the abstract 'element'. More research needs to be done to determine the strength of these correlations for "is a" constructions, but the certainly don't adhere to the heuristic that mismatched concreteness of entities leads to metaphoric utterances. A system built on this distinction using these kinds of lexical resources will not be able to reliably identify these kinds of metaphor.

We present a system that outperforms these approaches in F1 for metaphoric instances, and is both robust across domains and parts of speech, as well as independent of any hand-crafted lexical resources.

2 Word Embeddings

Word embeddings have proven fruitful in nearly all modern NLP tasks. In addition to using embeddings as input to neural network models, standard feature-based machine learning models can be improved by including word embeddings as features. Many metaphor detection systems rely on lexical databases for semantic features (imageability, concreteness, etc), and the abstract semantic information provided by word embeddings should be able to capture any kind of discrete semantic feature of a word by virtue of its distribution. Take the following example, repeated from above:

9. Djokovic killed Federer.

In 9, the sentence can be interpreted both literally and metaphorically. Given knowledge that the two nouns are actually tennis rivals, we should choose the metaphoric interpretation, mapping from the domain of conflict to the domain of sports. As both arguments are people, they are likely equal in concreteness, imageability, and ANEW scores.

This example shows the difficulty of using hand-crafted lexical features like concreteness and imageability to automatically detect linguistic metaphor. However, given enough training data, embeddings can capture finer-grained distinctions in the semantics of these words.

Despite their prevalence in other NLP tasks, there has been relatively little work done in incorporating word embeddings into metaphor identification models. (Shutova et al., 2016) uses a word2vec model to create word and image embeddings, and cite that they are the first to employ word embeddings for metaphor identification. Another embedding-based approach is that of (Do Dinh and Gurevych, 2016), who use a multi-layered perceptron neural network, with pretrained Google News word embeddings, and report an overall F1 score of .59 on the VUAMC dataset.

3 Data Analysis

As an initial analysis of the data, we calculated the average embedding of the word types that are used metaphorically compared to those used literally.

[3] The MRC database doesn't contain 'surgeon', but it does rate doctor as 600 for concreteness and 575 for imageability (on a scale of 100-700). 'butcher' has almost the same exact scores : 595 for concreteness and 556 for imageability. Brysbaert ranks 'surgeon' as 4.66/5 and 'butcher' as 4.65/5 for concreteness

	All	Only Lit	Only Met	Lit and Met
All	21570	3988	1386	16196
Nouns	11740	1694	594	9452
Verbs	4634	1048	634	2952
Adj/Adv	5378	1060	509	3809
Preps	359	118	9	232

Table 2: Word Type Counts in VUAMC

For this task we use the pre-trained Google News vectors (Mikolov et al., 2013). We split words into those used only metaphorically, and those used only literally. For an overview of counts of various parts of speech and their metaphoric and literal usages, see Table 2. We examined nouns, verbs, prepositions, and the combination of adjectives and adverbs. This includes total word type counts for each part of speech, as well as counts for those types used both metaphorically and literally, those used only metaphorically, and those used only literally. Nouns in particular are typically ambiguous, while prepositions are almost never used exclusively metaphorically.

We then calculated the following mean embeddings for each part of speech (nouns, verbs, adjectives and adverbs, prepositions):

- The mean embedding for word types that are *only used literally*
- The mean embedding for word types that are *only used metaphorically*

This information allows us to assess the embeddings of literal and metaphoric words, but gives no information about their context. This is insufficient for a task of disambiguation, as context is vital for determining a word's metaphor status : we cannot determine whether a word token that is typically used literally is being used metaphorically (and vice-versa) without knowing its context. As syntactic constructions have been shown to determine metaphor patterns (Sullivan, 2013) and dependency relations have been used successfully in other machine learning approaches (i.e. (Gargett and Barnden, 2015)), we also examined each word's dependency relations.

For this work, we will refer to the word being assessed as the 'target', and use its head and dependent relations from the Stanford Dependency Tagger (Chen and Manning, 2014). For each part of speech, we extracted the following embeddings:

- The mean embedding for all head words for target words that are *only used literally*
- The mean embedding for all dependent words for target words that are *only used literally*
- The mean embedding for all head words for target words that are *only used metaphorically*
- The mean embedding for all dependent words for target words that are *only used metaphorically*

This yields six data points for each part of speech :

1. **L** : Mean of only literal words
2. **M** : Mean of only metaphoric words
3. **L-H** : Mean of heads of literal words
4. **L-D** : Mean of dependents of literal words
5. **M-H** : Mean of heads of metaphoric words
6. **M-D** : Mean of dependents of metaphoric words

We then used principal component analysis (PCA) to represent the embedding averages in two dimensions, and plot them by part of speech in figures 1-5.

3.1 All Words (Fig 1)

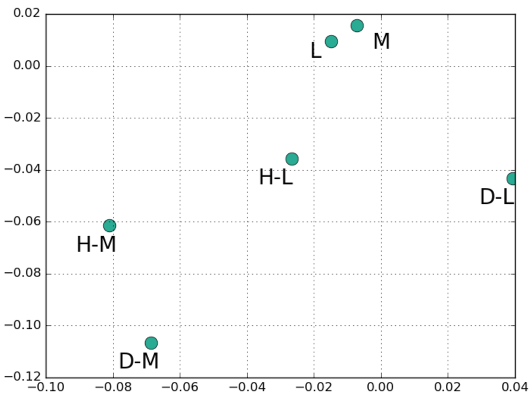

Figure 1: All Embeddings

The mean embeddings for target words are fairly consistent whether the word is metaphorically used or literally used. However, the heads and dependents of metaphoric words differ greatly from those of literal words. This indicates that word embeddings combined with dependency relations should be effective in distinguishing metaphoric targets from literal ones.

3.2 Nouns (Fig 2)

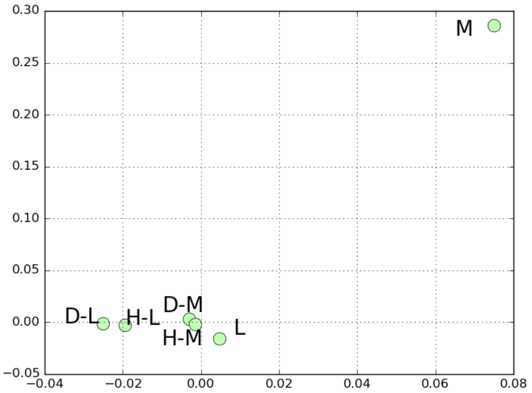

Figure 2: Noun Embeddings

Noun embeddings largely cluster together. Both head and dependent relations of literal and metaphoric words all have similar embeddings. However, the mean embedding for metaphoric target words is far apart from the rest. This is evidence that embeddings alone may indicate metaphoric noun usage, but adding dependency relations likely won't show significant improvement.

3.3 Verbs (Fig 3)

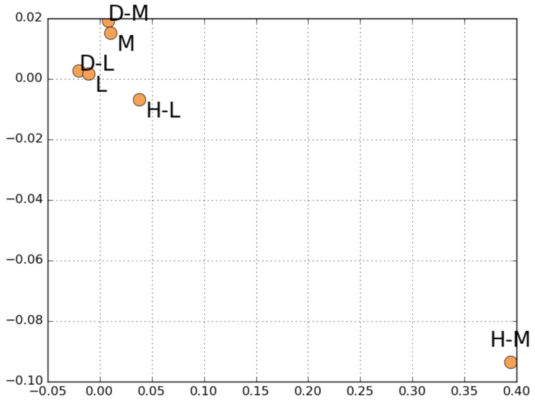

Figure 3: Verb Embeddings

Verb embeddings show a similar pattern to noun embeddings, with most means clustering together. The heads of metaphoric verbs vary greatly from the other mean embeddings, leading us to believe that adding head word embeddings for verbs will be critical in detecting their metaphoric use.

3.4 Adjectives/Adverbs (Fig 4)

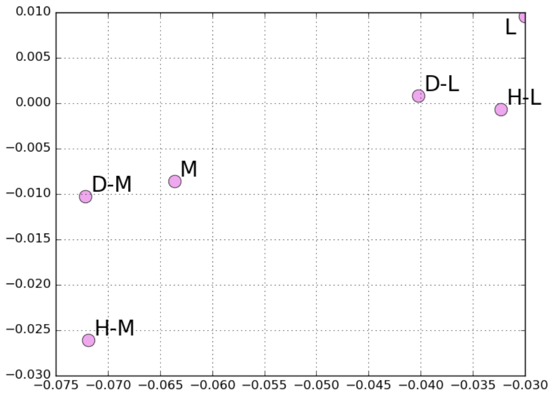

Figure 4: Adv/Adj Embeddings

We've lumped adjectives and adverbs into the same category, because of their similar semantic behavior : they both are used as modifiers for other words. This means they typically have rich head word information. From the PCA, we see that targets, heads, and dependents of metaphoric words cluster separately from their literal counterparts. This category is ideal for embeddings, and should see improvement both from target word embeddings and their dependency relations.

3.5 Prepositions (Fig 5)

Prepositions pose some difficulty, for multiple reasons. First, prepositions are used for a wide variety of semantic relations in English. They also have a wide variety of head and dependent words, making their average embeddings someone noisier than other categories. Our embeddings show one potentially useful pattern : all the metaphoric means are greater along the X axis, and both the target and head words are greater on the Y axis.

4 Hypotheses

From this data we have made four hypotheses about the effectiveness of word embeddings for classification:

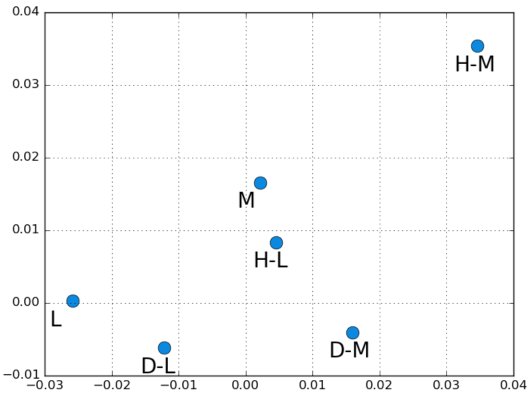

Figure 5: Preposition Embeddings

4.1 Embeddings are useful for classification.

From the graph of all embeddings, a well as each particular part of speech, we can see that embeddings for literal and metaphoric words differ, as do embeddings for their respective heads and dependents. In general, this indicates they will be effective for classification.

For examples like 6 and 9, we can see why this might be the case. Embeddings capture deeper semantic representations than lexical features, and thus should improve classification performance.

4.2 Adjective and adverb classification should be largely improved through both head and dependent embeddings.

Classification of adjectives and adverbs in particular should be improved through heads and dependents, as targets, heads, and dependents cluster together when the target is metaphoric.

4.3 Noun classification should be primarily based on the target word embedding.

Nouns should be more easily classified by the target word's embedding, and less by the head and dependency embeddings, as the dependency relations don't show any particular patterning.

4.4 Verb classification should be primarily improved by the head word embedding.

Verbs should be particularly improved by the embedding of their head word, as metaphoric targets show drastically different head word embeddings.

To test these hypotheses, we turn to machine learning applied to the VUAMC dataset.

5 Experiments

Our analysis shows that word embeddings are distinctly different between metaphoric and literal words, and the dependency relations of these words also show variation between literal and metaphoric targets.

To test our above hypotheses, we employed a Random Forest classifier, as they have been shown in previous work to be effective for metaphor detection using lexical features (Gargett and Barnden, 2015) (Tsvetkov et al., 2014)[4]. We included all words tagged as metaphor, and did not distinguish between the four registers in training or evaluation. This allows for the construction of a more robust classifier that should be capable of making better predictions on data from different domains.

We first established a baseline using only the embedding for the target word. We then added the embeddings for each word's head and the average of its dependents. Finally, we included the mean embedding of the entire sentence without the target word to test whether words lying outside of the target's syntactic context but still within the same sentence are accurate predictors of its metaphoric nature.

5.1 Results

We report the results for all words in the corpus, as well as results from training on individual parts of speech, as each part of speech has a different relation between the target's vector and its dependency relations. F1, precision, and recall for each part of speech are shown in Table 3.

Embeddings give strong performance across all parts of speech, with the best results being on verbs and prepositions and worst on adjectives and adverbs. Including the full sentence as context provides the best overall performance. This is likely because different parts of speech are dependent on different components, as we can see from other columns, and thus including all context words provides strong general classification.

Nouns generally don't respond to context, as the target embedding alone provides the best classification. Verbs are best classified using only their dependents, while adjectives and adverbs perform best with both head and dependents. Each part of speech has its own particular benefits from each of the syntactic dependency embeddings it uses,

[4] Model parameters/text processing available as supplemental material

	All			Noun			Verb			Adv/Adj			Prep		
	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R
Target Only	.624	.663	.589	**.644**	.720	**.582**	.701	.749	.659	.492	.618	.409	.690	.611	**.793**
Target, Head	.651	.715	.597	.626	.755	.537	.722	.736	.708	.530	.678	**.436**	.728	.734	.723
Target, Dependents	.637	.686	.593	.626	.765	.530	**.756**	.777	**.737**	.504	.632	.419	.689	.611	.791
Target, Head, Dependents	.653	.739	.597	.594	**.787**	.477	.749	.779	.721	**.535**	**.697**	.435	**.729**	.733	.726
Target, Full Context	**.664**	**.747**	**.598**	.639	.776	.543	.741	**.798**	.692	.517	.695	.412	**.729**	**.742**	.718

Table 3: Embedding Results by Part of Speech, with best result for each subcategory bolded

and we should use this as evidence that metaphor classifiers should use a combination of classifiers tuned to each particular part of speech.

5.2 Hypothesis Results

With regard to our other four hypotheses:

Embeddings are useful for classification. (4.1)

True : Embeddings show great utility in metaphor classification across all categories.

Adjective and adverb classification should be largely improved by both heads and dependents. (4.2)

True : Adjectives and adverbs show poor performance when only using the target embedding, and while including only the head word embedding improves performance much more than including only the target, the best performance is achieved by including both. Further context beyond head and dependents doesn't improve adjective/adverb classification.

Noun classification should be primarily based on the target embedding. (4.3)

True : Nouns' single target embedding performance outperforms the more complicated models that add either head or dependent embeddings. This is predicted from our PCA analysis, as the embeddings of metaphoric targets are quite distinct, but also poses a new problem : without context, the ceiling for noun classification is much lower. Only using target embeddings means that all instances of any particular noun will be classified as metaphoric or literal, despite our observations that particular noun tokens can differ in their use based on context.

Verb classification should be primarily improved by the head word embedding. (4.4)

False : While verb performance is increased when head embeddings are added, it gets a much bigger boost from the dependent embeddings, despite the apparent similarity between literal and metaphoric dependents in the PCA.

One interesting note from the results is that simply using the word embedding of the target word is a very strong baseline. This indicates that many words are simply more or less likely to be used metaphorically, regardless of their context, which is reasonable given the word types used exclusively metaphorically or literally in 2. It also indicates a strength of using word embeddings: they overcome lexical sparsity in the training data. If a word in the test data hasn't been seen, the classifier can still function using latent semantics present in the embedding. This is true of concreteness and imageability as well - the lexical units are represented only by their lexical features, so unseen words in the test data are not an issue.

5.3 State of the art comparison

In order to compare our results to those of most recent approaches, we present our results from running 10-fold cross validation results over the VUAMC data, using our best random forest classifier. Results are show in Table 4.

The word-embedding based model with sentence context embeddings outperforms all recent metaphor classification systems with regard to F1 and precision, while Klebanov's still delivers the best recall. We believe there are two key reasons why our approach outperforms those relying on lexical features:

- Lexical features do not capture enough semantic information. It is possible that with the concatenation of many lexical features, systems will improve, but the latent semantics present in word embeddings currently provides more semantic information than the other lexical features.

- The resources that features are drawn from are sparse. Rai et al (2016) show that expanding lexical resources automatically is promising for identification, which is further evidence that the resources currently are lacking in coverage. The Google News vectors

Model	F1	P	R
Dunn 2013	.455	.538	.394
Klebanov 2015	.511	.438	**.670**
Dinh 2016	.561	.611	.536
Haagsman 2016 (Verbs Only)	.578	-	-
Rai 2016	.609	.633	.587
W2V w/ context, all tokens	**.664**	**.747**	.598

Table 4: Result Comparison

have better coverage : only 13 percent of the tokens in the VUAMC don't have an embedding, compared to 20 percent from the MRC data.

The effectiveness of embeddings as features for a machine learning model is encouraging. Embeddings are acquirable for any language with sufficient text, and we believe the latent semantic nature of the lexical representations makes this approach particularly robust to novel data, both in other languages and other domains.

6 Future Work

A key advantage to using word embeddings for metaphor detection is their flexibility. Our approach is easily applicable to any other domain, as well as any other language for which there is significant data to train embeddings. One possible avenue for future research is evaluating our methods on both English and Russian, based on the work of (Tsvetkov et al., 2013) and (Ovchinnikova et al., 2014).

We are also interested in applying other neural network-based models to the data. As Dinh (2016) has shown neural network models to be relatively effective at metaphor detection, we think different kinds of neural models could prove increasingly useful for this task. Long-short term memory networks (LSTMs) have achieved state of the art performance in Word Sense Disambiguation (WSD) (Kågebäck and Salomonsson, 2016), (Yuan et al., 2016), a task very similar to metaphor identification. Both involve determining an appropriate semantics for a word from options : we can model metaphor detection as a binary word-sense task. Because of the task similarity and effectiveness of neural models, we believe this could be an valuable area of inquiry.

We also plan on experimenting with different word embeddings to maximize accuracy in different registers. Dinh et al (2016) observe that they achieved their best results on the news part of the corpus, and attributed it to their embedding selection: Google News vectors likely provide better representations for news data. We believe training embeddings specific to particular domains may improve classification for that domain, and training more general embeddings could improve the flexibility of this approach when applying it to unseen or variable-domain data.

References

Margaret M. Bradley and Peter J. Lang. 1999. Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings. Technical report, The Center for Research in Psychophysiology, University of Florida.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods* 46(3):904–911. https://doi.org/10.3758/s13428-013-0403-5.

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)*. Association for Computational Linguistics, Doha, Qatar, pages 740–750. http://www.aclweb.org/anthology/D14-1082.

Erik-Lân Do Dinh and Iryna Gurevych. 2016. Token-level metaphor detection using neural networks. In *Proceedings of the Fourth Workshop on Metaphor in NLP*. Association for Computational Linguistics, San Diego, California, pages 28–33. http://www.aclweb.org/anthology/W16-1104.

Jonathan Dunn. 2013. What metaphor identification systems can tell us about metaphor-in-language. In *Proceedings of the First Workshop on Metaphor in NLP*. Association for Computational Linguistics, Atlanta, Georgia, pages 1–10. http://www.aclweb.org/anthology/W13-0901.

Andrew Gargett and John Barnden. 2015. Modeling the interaction between sensory and affective meanings for detecting metaphor. In *Third Workshop on Metaphor in NLP*. Denver, CO, pages 21–30.

Sam Glucksberg. 2001. *Understanding Figurative Language*. Oxford University Press, London.

Hessel Haagsma and Johannes Bjerva. 2016. Detecting novel metaphor using selectional preference information. In *Proceedings of the Fourth Workshop on Metaphor in NLP*. Association for Computational Linguistics, San Diego, California, pages 10–17. http://www.aclweb.org/anthology/W16-1102.

Mikael Kågebäck and Hans Salomonsson. 2016. Word sense disambiguation using a bidirectional LSTM. *CoRR* abs/1606.03568. http://arxiv.org/abs/1606.03568.

Beata Beigman Klebanov, Chee Wee Leong, and Michael Flor. 2015. Supervised Word-Level Metaphor Detection: Experiments with Concreteness and Reweighting of Examples. In *Third Workshop on Metaphor in NLP*. Denver, CO, pages 11–20.

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

James H. Martin. 1990. *A Computational Model of Metaphor Interpretation*. Academic Press, Inc.

James H. Martin. 1996. Computational Approaches to Figurative Languge. *Metaphor and Symbolic Activity* 11(1):85–100.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *CoRR* abs/1310.4546. http://arxiv.org/abs/1310.4546.

Ekaterina Ovchinnikova, Ross Israel, Suzanne Wertheim, Vladimir Zaytsev, Niloofar Montazeri, and Jerry Hobbs. 2014. Abductive inference for interpretation of metaphors. In *Proceedings of the Second Workshop on Metaphor in NLP*. Association for Computational Linguistics, Baltimore, MD, pages 33–41. http://www.aclweb.org/anthology/W14-2305.

Sunny Rai, Shampa Chakraverty, and Devendra K. Tayal. 2016. Supervised metaphor detection using conditional random fields. In *Proceedings of the Fourth Workshop on Metaphor in NLP*. Association for Computational Linguistics, San Diego, California, pages 18–27. http://www.aclweb.org/anthology/W16-1103.

Ekaterina Shutova, Douwe Kiela, and Jean Maillard. 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*. Association for Computational Linguistics, San Diego, California, pages 160–170. http://www.aclweb.org/anthology/N16-1020.

Ekaterina V Shutova. 2011. Computational approaches to figurative language. Technical report, Cambridge University. http://www.cl.cam.ac.uk/.

Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna A. Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. *A Method for Linguistic Metaphor Identification*. John Benjamins Publishing Company.

Karen Sullivan. 2013. *Frames and Constructions in Metaphoric Language*. John Benjamins Publishing, Amsterdam/Philadelphia.

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)*. Association for Computational Linguistics, Baltimore, Maryland, pages 248–258. http://www.aclweb.org/anthology/P14-1024.

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*. Association for Computational Linguistics, Atlanta, Georgia, pages 45–51. http://www.aclweb.org/anthology/W13-0906.

Yorick Wilks. 1978. Making Preferences More Active. In *Words and Intelligence I*, Springer Netherlands, Dordrecht, pages 141–166. https://doi.org/10.1007/1-4020-5285-5_7.

Michael Wilson. 1988. Mrc psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers* 20(1):6–10. https://doi.org/10.3758/BF03202594.

Dayu Yuan, Ryan Doherty, Julian Richardson, Colin Evans, and Eric Altendorf. 2016. Word sense disambiguation with neural language models. *CoRR* abs/1603.07012. http://arxiv.org/abs/1603.07012.