How well do Embedding Models capture Non-compositionality? A View from Multiword Expressions

Navnita Nandakumar **Bahar Salehi** **Timothy Baldwin**
School of Computing and Information Systems
The University of Melbourne
Victoria 3010, Australia
nnandakumar@student.unimelb.edu.au {tbaldwin|salehi.b}@unimelb.edu.au

Abstract

In this paper, we apply various embedding methods to multiword expressions to study how well they capture the nuances of non-compositional data. Our results from a range of word-, character-, and document-level embbedings suggest that word2vec performs the best, followed by fastText and infersent. Moreover, we find that recently-proposed contextualised embedding models such as BERT and ELMo are not adept at handling non-compositionality in multiword expressions.

1 Introduction

Modern embedding models, including contextual embeddings, have been shown to work impressively well across a range of tasks (Peters et al., 2018; Devlin et al., 2018). However, study of their performance on data with a mix of compositionality levels, whose meaning is often not easily predicted from that of its constituent words, has been limited (Salehi et al., 2015; Hakimi Parizi and Cook, 2018; Nandakumar et al., 2018).

At present, there exists no definitive metric to measure the modelling capabilities of an embedding technique across a spectrum of non-compositionality, especially in the case of newer, contextualised representations, such as ELMo and BERT.

In this study, we apply various embedding methods to the task of determining the compositionality of English multiword expressions ("MWEs"), specifically noun–noun and adjective–noun pairs, to test their performance on data representing a range of compositionality (Sag et al., 2002). Compositionality prediction can be modeled as a regression task (Baldwin and Kim, 2010) that involves mapping an MWE onto a continuous scale, representing its compositionality as a whole or with respect to each of its components. For example, *application form* can be considered to be quite compositional, while *sitting duck*[1] is considered to be idiomatic or non-compositional. *Close shave*[2] could be seen as partially compositional, heavily compositional with regards to the first word and less compositional with regards to the second. In this study, we focus on predicting the compositionality of the MWE as a whole. Although we conduct our experiments on English datasets, they can be applied to other languages with ease as we do not perform any kind of language-specific manipulation of the data.

The main contributions of this paper are:

(i) we compare embeddings over 3 different MWE datasets, focusing on noun–noun and adjective–noun pairs;
(ii) we experiment with 7 character-, word-, and document-level embedding models, including contextualised models;
(iii) we show that, despite their success on a range of other tasks, recent embedding learning methods lag behind simple word2vec in capturing MWE non-compositionality.

2 Related Work

Although vector space models have been popular since the 1990s, it was only after Collobert and Weston (2008) proposed a unified neural network architecture to learning distributed word representations and demonstrated its performance on downstream tasks, that embedding learning established a footing in NLP, with word2vec (Mikolov et al., 2013a) being the catalyst to the "embedding revolution".

Language embeddings are an example of an unsupervised representation learning application done well. They are preferred primarily because they can be learned from unannotated corpora and,

[1]A *sitting duck* means "a person or thing with no protection against an attack or other source of danger."

[2]A *close shave* is "a narrow escape from danger or disaster."

therefore, eliminate the need for manual annotation (which is expensive and time-consuming).

Salehi et al. (2015) were the first to apply word embeddings to the task of predicting the compositionality of MWEs. The assumption is that the compositionality of an MWE is proportional to the relative similarity between each of the components and the overall MWE, represented by their respective embeddings. This method was recently tuned variously by Cordeiro et al. (2019) and remains state-of-the-art for the task of MWE compositionality prediction, but has the downside that it requires automatic token-level pre-identification of each MWE in the training corpus in order to train a model (i.e. all occurrences of *sitting duck* need to be pre-tokenised to a single token, such as *sitting_duck*). This is not ideal, as it means the model will need to be retrained for a new set of MWEs (as the tokenisation will necessarily change). It also requires "complete" knowledge of the MWEs before the training step, which is impractical in most cases.

Character-level embedding models (Hakimi Parizi and Cook, 2018) are one possible solution to the fixed-vocabulary problem, in being able to handle an unbounded vocabulary, including MWEs. Document embeddings (Le and Mikolov, 2014; Conneau et al., 2017a) are also highly relevant to dynamically generating embeddings for MWEs, as they generate representations of arbitrary spans of text, which are potentially able to capture the context of use of the MWE.

3 Methodology

Following Salehi et al. (2015) and Nandakumar et al. (2018), we compute the overall compositionality of an MWE with three broad metrics: direct composition, paraphrase similarity, and a combined metric. In all experiments, the similarity of a pair of vectors is measured using cosine similarity.

3.1 Direct Composition

Intuitively, an MWE appearing in similar contexts to its components is likely to be compositional. We directly compare the vector embedding of the MWE (described in Section 4.2) with that of its component words, in one of two ways: (1) performing an element-wise sum to obtain a 'combined' vector, which is then compared with the vector of the MWE ($\text{Direct}_{\text{pre}}$); and (2) a post-hoc combination of the scores obtained by individually comparing the component vectors with that of the MWE via a weighted sum ($\text{Direct}_{\text{post}}$). Formally:

$$\begin{aligned}\text{Direct}_{\text{pre}} &= \cos(\mathbf{mwe}, \mathbf{w_1} + \mathbf{w_2})\\ \text{Direct}_{\text{post}} &= \alpha\cos(\mathbf{mwe}, \mathbf{w_1}) +\\ &\quad (1-\alpha)\cos(\mathbf{mwe}, \mathbf{w_2})\,,\end{aligned}$$

where: $\mathbf{mwe}$, $\mathbf{w_1}$, and $\mathbf{w_2}$ are the embeddings for the combined MWE, first component and second component, respectively;[3] $\mathbf{w_1}+\mathbf{w_2}$ is the element-wise sum of the vectors of each of the component words of the MWE; and $\alpha \in [0,1]$ is a scalar which allows us to vary the weight of the respective components in predicting the compositionality of the compound. This helps us effectively capture the compositionality of the MWE with regards to each of its individual constituents.

We do not perform any tuning of α over held-out data and are, as such, overfitting as we select the best-performing α post hoc. We do, however, present analysis of hyper-parameter sensitivity in Section 5.

3.2 Paraphrase Similarity

Assuming access to paraphrases of an MWE, another intuition is that if the MWE appears in similar contexts to the component words of its paraphrases, it is likely to be compositional (Shwartz and Waterson, 2018). Each paraphrase provides an interpretation of the semantics of the MWE, e.g. *ancient history* is "in the past", "old news" or "forever ago" (note how each paraphrase brings out a slightly different interpretation). The RAMISCH MWE dataset (described in Section 4.1) provides one or more paraphrases for each MWE contained in it. We calculate the similarity of the embeddings of the MWE and its paraphrases using the following three formulae:

$$\begin{aligned}\text{Para\_first} &= \cos(\mathbf{mwe}, \mathbf{para_1})\\ \text{Para\_all}_{\text{pre}} &= \cos(\mathbf{mwe}, \sum_i \mathbf{para_i})\\ \text{Para\_all}_{\text{post}} &= \frac{1}{N}\sum_{i=1}^{N}\cos(\mathbf{mwe}, \mathbf{para_i})\,,\end{aligned}$$

where $\mathbf{para_1}$ and $\mathbf{para_i}$ denote the embedding for the first (most popular) and i-th paraphrases, respectively.

[3] All methods are presented and evaluated in terms of two-element MWEs in this work, but are trivially generalisable to multi-element MWEs.

In the case of Para_all$_{\text{post}}$, we considered computing the maximum instead of the average (as we report here) of the similarity scores between each paraphrase and its MWE, following the intuition that an MWE would be similar to at least one reported paraphrase, rather than all of them. However, the results for the average similarity were empirically higher across models.

3.3 Combined Metric

Finally, we present the combined results from the two metrics stated above:

$$\text{Combined} = \beta \max\left(\text{Direct}_{\text{pre}}, \text{Direct}_{\text{post}}\right) + (1-\beta) \max\left(\text{Para\_first}, \text{Para\_all}_{\text{pre}}, \text{Para\_all}_{\text{post}}\right) ,$$

where $\beta \in [0, 1]$ is a scalar weighting factor used to balance the effects of the two methods, in order to measure the extent to which the compositionality is determined by each of the methods. The choice of the $\max$ operator here to combine the sub-methods for each of the direct composition and paraphrase methods is that all methods tend to underestimate the compositionality (and empirically, it was found to be superior to taking the mean).

4 Experiments

4.1 Datasets

We used three datasets for our experiments, evaluating each model's performance using Pearson's correlation coefficient (r) to compare the similarity scores obtained with the annotated compositionality scores provided in the dataset.

REDDY The dataset of Reddy et al. (2011) contains 90 binary English noun compounds ("NCs"), along with human-annotated scores of their overall compositionality and component-specific compositionality, both ranging from 0 to 5. For our experiments, we consider the overall compositionality scores only.

RAMISCH Similar to REDDY, the English dataset of Ramisch et al. (2016) contains 90 binary noun compounds with annotated scores of compositionality ranging from 0 to 5, both overall and component-specific (of which we use only the former). It also contains a list of paraphrases for each NC, presented in decreasing order of popularity among the annotators.

Dataset	μ	σ
REDDY	53.2	30.0
RAMISCH	52.6	35.0
DISCO	68.1	21.7
Overall	59.7	29.0

Table 1: Mean (μ) and standard deviation (σ) of the compositionality scores for the three datasets used in this research, over a normalised range $[0, 100]$.

DISCO$_{\text{ADJ}}$ The English dataset from the DiSCo shared task (Biemann and Giesbrecht, 2011) containing a total of 348 binary phrases, comprising adjective–noun, verb–noun$_{\text{subj}}$, and verb–noun$_{\text{obj}}$ pairs, along with their overall compositionality rating ranging from 0 to 100. The phrases were extracted semi-automatically and their relations were assigned by patterns and checked manually. The compositionality scores were collected from Amazon Mechanical Turk, where workers were presented 4–5 randomly sampled sentences from the UK English WACKy corpora. We focus on the 144 adjective–noun pairs in this study.

The breakdown of compositionality scores across the three datasets in Table 1 indicates there is a reasonable distribution of data in terms of compositionality, with REDDY and RAMISCH being roughly comparable and covering a broad (and somewhat balanced) spectrum of compositionalities, while DISCO is more skewed towards compositional usages, with lower standard deviation.

4.2 Embeddings

We made use of various embeddings, ranging from character- to document-level, in our study. Below is a description of each model along with how they are trained. Where available, we made use of pre-trained models as is standard practice in NLP. As the different models were trained on different corpora, we are not attempting to perform a controlled comparative evaluation of the different models, so much as a comparison of the standard pre-trained versions of each. If we were to retrain our own models over a standard dataset such as English Wikipedia, we would expect the results for the document-level embedding methods in particular to drop.

4.2.1 Word-level

A word embedding captures the context of a word in a document (in relation to other words) in the form of a vector representation. It tokenises text at the word level.

word2vec We trained word2vec (Mikolov et al., 2013b) on a recent English Wikipedia dump,[4] after pre-processing (removing the formatting and punctuation) and concatenating each occurrence of the multiword expressions in our datasets (e.g. every occurrence of *close shave* in the corpus becomes *closeshave*). We make the greedy assumption that every occurence of the component words in sequence is an occurrence of the expression. We perform this token-level identification and manipulation of the corpus in order to obtain a single embedding for the expression, instead of a separate embeddings for the individual component words. In cases where the model still fails to generate an embedding (2 for REDDY, 8 for RAMISCH and 25 for DISCO) for the expression (due to low token frequency), we assign a default compositionality score of 0.5 (neutral; based on a range of $[0, 1]$). For paraphrases, we compute an element-wise sum of the embeddings for each of the component words to serve as the embedding of the phrase. We do this because token-level identification of each paraphrase in the training corpus is not practical.

4.2.2 Character-level

Character-level embeddings can generate vectors for words based on n-gram character aggregations. This means they can generate embeddings for out-of-vocabulary (OOV) words, as well new words or misspelled words. It tokenises text at the character level.

fastText We used the 300-dimensional fastText model pre-trained on Common Crawl and Wikipedia using CBOW (fastText$_{\text{pre}}$), as well as one trained over the same Wikipedia corpus[4] using skip-gram (fastText). Again, since fastText (Bojanowski et al., 2017) assumes all words to be whitespace delimited, we preprocess our MWE and paraphrases the same way as above (removing the space between them so that *armchair critic* becomes *armchaircritic*, say).

Contextualised Embeddings Unlike classical embedding techniques, contextualised embeddings capture the semantics of a word or phrase in a manner which is sensitised to the context of usage.

We used the pretrained implementations of ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018) found in the Flair framework.[5] The framework also has a contextualised string embedding model of its own, also named Flair (Akbik et al., 2018).

We supplied sentences extracted from the Brown corpus where available in order to derive a contextualised interpretation. We extracted 25 sentences at random per MWE, except where there were fewer sentences in the corpus.

However, we also included a naive context-independent implementation in our study, consistent with the other models, following the intuition that the relative compositionality of even a novel compound can often be predicted from its component words alone (e.g. *giraffe potato* having the plausible compositional interpretation of a potato shaped like a giraffe vs. *couch intelligence* having no natural interpretation).

4.2.3 Document-level

Document embeddings aggregate from words to documents, generating vector representations for entire documents. Since document and sentence embeddings are capable of generating a single embedding for a span of text, we are able to generate representations of the MWEs and paraphrases without preprocessing them (to remove space). We treat each constituent word as a single word document to generate embeddings.

infersent We used two versions of infersent (Conneau et al., 2017b): infersent$_{\text{GloVe}}$ and infersent$_{\text{fastText}}$. Each generates a representation of 300 dimensions, trained over the 1,000,000 most popular English words using GloVe (Pennington et al., 2014) and fastText, respectively.

doc2vec We used the gensim implementation of doc2vec (Le and Mikolov, 2014; Lau and Baldwin, 2016) pretrained on Wikipedia data using the word2vec skip-gram models pretrained on Wikipedia and AP News.[6]

[4] Dated 07-Jan-2019

[5] https://github.com/zalandoresearch/flair

[6] https://github.com/jhlau/doc2vec

Emb. method	$\text{Direct}_{\text{pre}}$	$\text{Direct}_{\text{post}}$	Para_first	$\text{Para\_all}_{\text{pre}}$	$\text{Para\_all}_{\text{post}}$	Combined
Flair	0.165	0.295 ($\alpha = 0.1$)	0.334	0.399	0.492	0.492 ($\beta = 0.0$)
$\text{Flair}_{\text{context}}$	0.181	0.314 ($\alpha = 0.1$)	0.357	0.411	0.522	0.522 ($\beta = 0.0$)
$\text{fastText}_{\text{pre}}$	0.395	0.446 ($\alpha = 0.7$)	0.242	0.531	0.703	0.703 ($\beta = 0.0$)
fastText	0.464	0.532 ($\alpha = 0.7$)	0.548	0.613	0.673	0.673 ($\beta = 0.0$)
BERT	0.071	0.086 ($\alpha = 1.0$)	0.242	0.531	0.583	0.583 ($\beta = 0.0$)
$\text{BERT}_{\text{context}}$	0.089	0.111 ($\alpha = 1.0$)	0.267	0.546	0.601	0.601 ($\beta = 0.0$)
ELMo	0.420	0.459 ($\alpha = 0.6$)	0.361	0.488	0.546	0.546 ($\beta = 0.2$)
$\text{ELMo}_{\text{context}}$	0.461	0.489 ($\alpha = 0.6$)	0.373	0.492	0.552	0.627 ($\beta = 0.2$)
word2vec	**0.581**	**0.571** ($\alpha = 0.6$)	0.443	0.510	0.504	0.677 ($\beta = 0.9$)
$\text{infersent}_{\text{GloVe}}$	0.321	0.427 ($\alpha = 0.7$)	**0.636**	0.700	**0.741**	**0.783** ($\beta = 0.5$)
$\text{infersent}_{\text{fastText}}$	0.169	0.221 ($\alpha = 0.6$)	0.488	**0.712**	0.636	0.774 ($\beta = 0.0$)
doc2vec	–0.157	0.039 ($\alpha = 1.0$)	0.388	0.334	0.373	0.419 ($\beta = 0.3$)

Table 2: Pearson correlation coefficient for compositionality prediction results on the RAMISCH dataset.

Emb. method	$\text{Direct}_{\text{pre}}$	$\text{Direct}_{\text{post}}$
Flair	–0.127	0.024 ($\alpha = 0.0$)
$\text{Flair}_{\text{context}}$	0.012	0.172 ($\alpha = 0.0$)
$\text{fastText}_{\text{pre}}$	0.223	0.285 ($\alpha = 0.3$,0.4)
fastText	0.217	0.287 ($\alpha = 0.3$,0.4)
BERT	0.304	0.352 ($\alpha = 0.2$)
$\text{BERT}_{\text{context}}$	0.313	0.377 ($\alpha = 0.2$)
ELMo	0.339	0.406 ($\alpha = 0.5$)
$\text{ELMo}_{\text{context}}$	0.387	0.416 ($\alpha = 0.5$)
word2vec	**0.634**	**0.622** ($\alpha = 0.6$)
$\text{infersent}_{\text{GloVe}}$	0.413	0.500 ($\alpha = 0.5$)
$\text{infersent}_{\text{fastText}}$	0.401	0.527 ($\alpha = 0.6$)
doc2vec	–0.049	0.025 ($\alpha = 0.0$)

Table 3: Pearson correlation coefficient for compositionality prediction results on the REDDY dataset.

Emb. method	$\text{Direct}_{\text{pre}}$	$\text{Direct}_{\text{post}}$
Flair	0.261	0.291 ($\alpha = 0.4$)
$\text{Flair}_{\text{context}}$	0.280	0.315 ($\alpha = 0.4$)
$\text{fastText}_{\text{pre}}$	0.339	0.353 ($\alpha = 0.6$,0.7)
fastText	0.374	0.419 ($\alpha = 0.4$)
BERT	0.154	0.177 ($\alpha = 0.3$,0.4)
$\text{BERT}_{\text{context}}$	0.163	0.189 ($\alpha = 0.3$)
ELMo	0.253	0.287 ($\alpha = 0.5$)
$\text{ELMo}_{\text{context}}$	0.301	0.319 ($\alpha = 0.5$)
word2vec	**0.427**	**0.419** ($\alpha = 0.4$)
$\text{infersent}_{\text{GloVe}}$	0.321	0.315 ($\alpha = 0.4$)
$\text{infersent}_{\text{fastText}}$	0.001	0.202 ($\alpha = 1.0$)
doc2vec	–0.023	0.003 ($\alpha = 0.0$)

Table 4: Pearson correlation coefficient for compositionality prediction results on the $\text{DISCO}_{\text{ADJ}}$ dataset.

5 Results and Discussion

The results from our experiments on the RAMISCH, REDDY and DISCO datasets can be found in Tables 2, 3 and 4, respectively, with the best performing αs and βs for each embedding method.

We observe that the αs in Table 2 are high, implying the compound nouns in RAMISCH are more compositional in terms of their head (second) nouns. Similarly, the lower α scores in Table 3 suggest REDDY's compound nouns are more dependent on their modifiers, or first nouns. Table 4, on the other hand, shows the αs embracing the entire range of $[0, 1]$. This suggests the adjective–noun pairs in DISCO are spread in terms of their dependency on their constituents, which also depends on the embedding method used. Overall, the methods are sensitive to the choice of the α hyperparameter, with ELMo and infersent being particularly sensitive and showing substantial change in output with change in α (Figures 1,2 and 3).

We see that for RAMISCH (Table 2), word2vec achieves the highest scores among the direct combination metrics, while infersent outperforms the other methods among the paraphrase metrics, and word2vec falls behind character embedding models like fastText, ELMo and BERT (even when the latter two were performed without context). The lower β scores also show the other models favouring the paraphrase metrics, while the high β score for word2vec shows its preference for di-

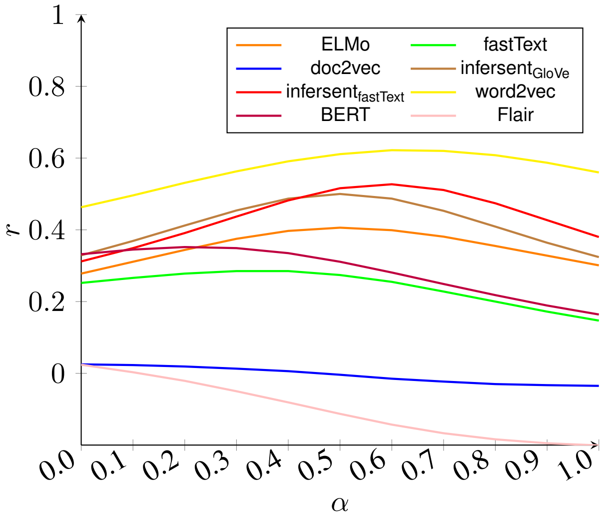

Figure 1: Sensitivity analysis of α (REDDY)

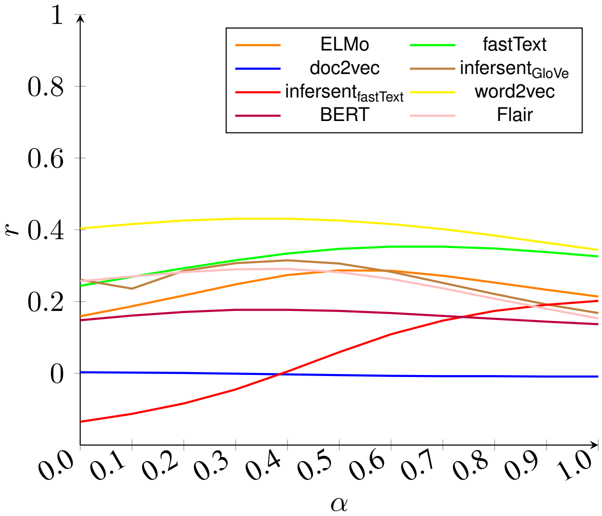

Figure 3: Sensitivity analysis of α (DISCO)

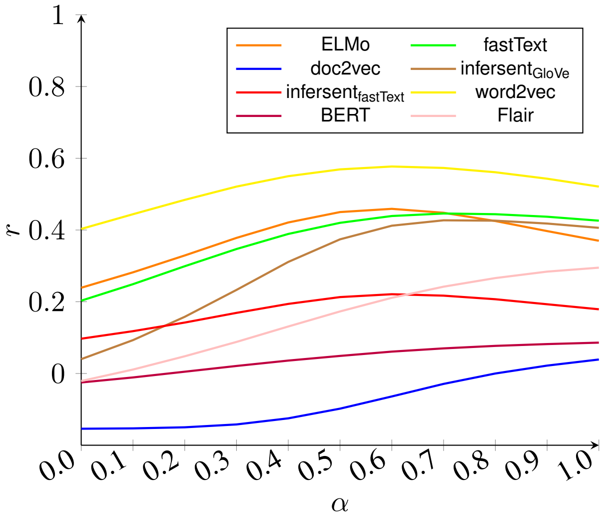

Figure 2: Sensitivity analysis of α (RAMISCH)

rect combination.

We observe that, consistent with its performance on RAMISCH, word2vec performs the best of all models for the direct combination methods.

Overall, we observe that word2vec is consistent in providing the best results based on the methods outlined in Section 3.1, while fastText and infersent come a close second and third, respectively. It is noteworthy, however, that word2vec required explicit modelling of the MWEs during the training procedure, while the other models did not.

It is not surprising that infersent, being a document-level embedding model, works better with paraphrase data than the other models. However, doc2vec has really poor scores overall across the three datasets. It does, however, redeem itself with the paraphrases, with substantially higher scores than the direct metric but still quite a way behind the top-scoring methods.

We also see that the paraphrase metric seems to achieve much greater results across all models, suggesting this could be a direction for future study (noting the requirement for paraphrase data for the MWE in order to apply this method, which has inherent scalability limitations). The combined metric seems to favour the paraphrase results as well, based on the relative β values.

One of the reasons word2vec did not work as well with the paraphrases could be the naive assumption that the Direct$_{\text{pre}}$ is a representation of the paraphrase itself. As we see from the results across the datasets and methods, Direct$_{\text{pre}}$ does not entirely capture the compositionality of the MWE, so it is reasonable to assume that a paraphrase would not be accurately represented by Direct$_{\text{pre}}$ either.

We see that fastText provides us with impressive scores throughout, and we notice a slight improvement when trained on the same corpus as word2vec. However, there is a huge gap in the performance between word2vec and fastText, especially in the case of REDDY (which could be an issue of a heavier representation of a particular level of compositionality, say).

We also notice that, unlike the noun compounds in REDDY and RAMISCH, there is less variance in the relative scores of each method in the case of DISCO$_{\text{ADJ}}$, with overall results dropping appreciably, and the best-performing word2vec dropping back in raw r value compared to noun–noun pairs.

In terms of the contextualised embeddings, we notice that across the three models, there is only a slight increase in correlation when contextualised embeddings are used. This suggests that even with context, these modern embedding techniques are unable to capture non-compositionality as well as their simpler counterparts.

Further analysis reveals that most models struggle to accurately predict the compositionality of idiomatic noun compounds, as well as semi-compositional terms wherein one of the constituent words are used in a metaphoric sense. In REDDY, we observe this for *silver bullet* and *snail mail*. Interestingly, while BERT struggles to effectively model compositionality throughout, it is surprisingly the only model able to perfectly predict the compositionality of *snail mail* (which appears as an extreme outlier). This suggests that BERT might be more successful using a different metric. In the case of the adjective–noun phrases in DISCO, we see that the models are still unable to accurately predict the compositionality of non-compositional phrases (like *big fish*, *heavy metal* and *red tape*). This time, however, they are also unable to capture *mobile phone* and *floppy disk*, perhaps because of their relatively archaic use.

6 Conclusion

In this paper, we investigated the modelling capabilities of various embedding techniques applied to the specific task of predicting the MWE compositionality, to see how well they model a mixture of compositionalities in the dataset. Our results indicate that modern character- and document-level embedding methods are inferior to the simple word2vec approach. However, the promising results of fastText and infersent across the datasets indicate that, among the more modern methods, they are better equiped to handle non-compositionality as they did not require much manipulation of the corpus or knowledge of the MWEs beforehand. We also found that the paraphrase metric results in greater correlation scores across the models.

In future work, we intend to tune our hyperparameters over held-out data, and experiment with other languages and language-independent techniques, including other models.

Acknowledgments

We would like to thank the curators and annotators of the datasets used in this study, and the anonymous reviewers for their insightful comments. We would also like to thank Chris Biemann from the University of Hamburg for making the dataset from the DiSCo shared task available to us.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, USA.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In *Nitin Indurkhya and Fred J. Damerau (eds.) Handbook of Natural Language Processing, Second Edition, CRC Press*, pages 267–292, Boca Raton, USA.

Chris Biemann and Eugenie Giesbrecht. 2011. Distributional semantics and compositionality 2011: Shared task description and results. In *Proceedings of the Workshop on Distributional Semantics and Compositionality*, pages 21–28, Portland, USA.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, pages 160–167.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017a. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017b. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Silvio Cordeiro, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. 2019. Unsupervised compositionality prediction of nominal compounds. *Computational Linguistics*, 45(1).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ali Hakimi Parizi and Paul Cook. 2018. Do character-level neural network language models capture knowledge of multiword expression compositionality? In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 185–192, Santa Fe, USA.

Jey Han Lau and Timothy Baldwin. 2016. An empirical evaluation of doc2vec with practical insights into document embedding generation. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 78–86, Berlin, Germany.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML 2014)*, pages 1188–1196, Beijing, China.

Tomas Mikolov, G.s Corrado, Kai Chen, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *Proceddings of Workshop at the International Conference on Learning Representations*, pages 1–12.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

Navnita Nandakumar, Bahar Salehi, and Timothy Baldwin. 2018. A comparative study of embedding models in predicting the compositionality of multiword expressions. In *Proceedings of the Australasian Language Technology Association Workshop 2018*, pages 71–76, Dunedin, New Zealand.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, USA.

Carlos Ramisch, Silvio Cordeiro, Leonardo Zilio, Marco Idiart, and Aline Villavicencio. 2016. How naked is the naked truth? a multilingual lexicon of nominal compound compositionality. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 156–161, Berlin, Germany.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 210–218, Chiang Mai, Thailand.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, pages 1–15, Mexico City, Mexico.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2015. A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 977–983, Denver, USA.

Vered Shwartz and Chris Waterson. 2018. Olive oil is made of olives, baby oil is made for babies: Interpreting noun compounds using paraphrases in a neural model. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 218–224, New Orleans, USA.