Predicting Metaphor Paraphrase Judgements in Context

Yuri Bizzoni
University of Gothenburg
yuri.bizzoni@gu.se

Shalom Lappin
University of Gothenburg
shalom.lappin@gu.se

Abstract

We conduct two experiments to study the effect of context on metaphor paraphrase aptness judgments. The first is an AMT crowd source task in which speakers rank metaphor-paraphrase candidate sentence pairs in short document contexts for paraphrase aptness. In the second we train a composite DNN to predict these human judgments, first in binary classifier mode, and then as gradient ratings. We found that for both mean human judgments and our DNN modeling, adding document context compresses the aptness scores towards the centre of the scale, raising low out of context ratings and decreasing high out of context scores. We briefly consider two possible explanations for this compression effect.

1 Introduction

A metaphor is a way of forcing the normal boundaries of words' meaning in order to better express an experience, a concept or an idea. At least to a native speaker's ear, some metaphors sound more conventional (like the usage of the words *ear* and *sound* in this sentence), others more original. This is not the only way to judge a metaphor. One of the most important qualities of a metaphor is its appropriateness, its *aptness*. This poses the question of how good a metaphor is for conveying a given experience or concept. While a metaphor's degree of conventionality can be measured through probabilistic methods, like language models, it is harder to model its aptness. Chiappe et al. (2003) define *aptness* as "the extent to which a comparison captures important features of the topic".

It is possible to express an opinion about some metaphors' and similes' aptness (at least to a degree) without previously knowing what they are trying to convey, or the context in which they appear[1]. For example, we don't need a particular context or frame of reference to construe the simile *She was screaming like a turtle* as strange, and less apt for expressing the quality of a scream, than *She was screaming like a banshee*. In this case, the reason why the simile in the second sentence works better is intuitive. A salient characteristic of a banshee is a powerful scream. Turtles are not known for screaming, and so it is harder to define the quality of a scream through such a comparison, except as a form of irony.[2] Other cases are more complicated. The simile *crying like a fire in the sun* (*It's All Over Now, Baby Blue*, Bob Dylan) is powerfully apt for many readers, but simply odd for others. Fire and sun do not cry in any way. But at the same time the simile can express the association we draw between something strong and intense in other sensory modes, such as vision and touch, on one hand and a loud cry on the other.

Nevertheless, most metaphors and similes need some kind of context, or external reference point to be interpreted. The sentence *The old lady had a heart of stone* is apt if the old lady is cruel or indifferent, but it is unreasonable as a description of a situation in which the old lady is kind and caring. We assume that, to an average reader's sensibility, the sentence models only the first situation appropriately.

[1]While it can be argued that metaphors and similes at some level work differently and cannot always be considered as variations of the same phenomenon (Sam and Catrinel, 2006; Glucksberg, 2008), for this study we treat them as belonging to the same category of figurative language.

[2]It is important not to confuse aptness with transparency. The latter measures how easy it is to understand a comparison. Chiappe et al. (2003) claim, for example, that many literary or poetic metaphors score high on aptness and low on transparency, in that they capture the nature of the topic very well, but it is not always clear why they work.

This is the view of metaphor aptness that we adopt in this paper. Following Bizzoni and Lappin (2018), we treat a metaphor as apt in relation to a literal expression that it paraphrases.[3] If the metaphor is judged to be a good paraphrase, then it closely "models" the core information of the literal sentence through its metaphorical shift. We refer to the prediction of readers' judgments on the aptness candidates for the literal paraphrase of a metaphor as the *metaphor paraphrase aptness task* (MPAT). Bizzoni and Lappin (2018) address the MPAT by using Amazon Mechanical Turk (AMT) to obtain crowd sourced annotations of metaphor-paraphrase candidate pairs. They train a composite Deep Neural Network (DNN) on a portion of their annotated corpus, and test it on the remaining part. Testing involves using the DNN as a binary classifier on paraphrase candidates. They derive predictions of gradient paraphrase aptness for their test set, and assess them by Pearson coefficient correlation to the mean judgments of their crowd sourced annotation of this set. Both training and testing are done independently of any document context for the metaphorical sentence and its literal paraphrase candidates.

In this paper we study the role of context on readers' judgments concerning the aptness of metaphor paraphrase candidates. We look at the accuracy of Bizzoni and Lappin (2018)'s DNN when trained and tested on contextually embedded metaphor-paraphrase pairs for the MPAT. In Section 2 we describe an AMT experiment in which annotators judge metaphors and paraphrases embedded in small document contexts, and in Section 3 we discuss the results of this experiment. In Section 4 we describe our MPAT modeling experiment, and in Section 5 we discuss the results of this experiment. Section 6 surveys some work on metaphor aptness and computational methods to deal with it. In Section 7 we draw conclusions from the studies presented in this paper, and we indicate directions for future work in this area.

2 Annotating Metaphor-Paraphrase Pairs in Contexts

Bizzoni and Lappin (2018) have recently produced a dataset of paraphrases containing metaphors designed to allow both supervised binary classification and gradient rankings. This dataset contains several pairs of sentences, where in each pair the first sentence contains a metaphor, and the second is a literal paraphrase candidate.

This corpus was constructed with a view to representing a large variety of syntactic structures and semantic phenomena in metaphorical sentences. Many of these structures and phenomena do not occur as metaphorical expressions, with any frequency, in natural text and were therefore introduced through hand crafted examples.

Each pair of sentences in the corpus has been rated by AMT annotators for paraphrase aptness on a scale of 1-4, with 4 being the highest degree of aptness. In Bizzoni and Lappin (2018)'s dataset, sentences come in groups of five, where the first element is the "reference element" with a metaphorical expression, and the remaining four sentences are "candidates" that stand in a degree of paraphrasehood to the reference.

Here is an example of a metaphor-paraphrase candidate pair.

1a. The crowd was a roaring river.

b. The crowd was huge and noisy.

[3] Bizzoni and Lappin (2018) apply Bizzoni and Lappin (2017)'s modeling work on general paraphrase to metaphor.

The average AMT paraphrase score for this pair is 4.0, indicating a high degree of aptness.

We extracted 200 sentence pairs from Bizzoni and Lappin (2018)'s dataset and provided each pair with a document context consisting of a preceding and a following sentence,[4] as in the following example.

2a. They had arrived in the capital city. **The crowd was a roaring river**. It was glorious.

b. They had arrived in the capital city. **The crowd was huge and noisy**. It was glorious.

One of the authors constructed most of these contexts by hand. In some cases, it was possible to locate the original metaphor in an existing document. This was the case for

(i) Literary metaphors extracted from poetry or novels, and

(ii) Short conventional metaphors (*The President brushed aside the accusations*, *Time flies*) that can be found, with small variations, in a number of texts.

For these cases, a variant of the existing context was added to both the metaphorical and the literal sentences. We introduced small modifications to keep the context short and clear, and to avoid copyright issues. We lightly modified the contexts of metaphors extracted from corpora when the original context was too long, ie. when the contextual sentences of the selected metaphor were longer than the maximum length we specified for our corpus. This was necessary due to the fact that the original, natural contexts can have an excessive length and include far-reaching references to previous content. In such cases we reduced the length of the sentence and we slightly simplified the text, while sustaining its meaning. We tried to sustain "naturalness" of the context. Since the same context is used for metaphors and their literal candidate paraphrases, we specified short contexts that make sense for both the figurative and the literal sentences, even when the pair had been judged as non-paraphrases. We kept the context as neutral as possible in order to avoid biasing effects on crowd source judgments.

For example, in the following pair of sentences, the literal sentence is *not* a good paraphrase of the figurative one (a simile).

3a. He is grinning like an ape.

b. He is smiling in a charming way. (*average score:* 1.9)

We opted for a context that is natural for both sentences.

4a. Look at him. **He is grinning like an ape.** He feels so confident and self-assured.

b. Look at him. **He is smiling in a charming way.** He feels so confident and self-assured.

We sought to avoid, whenever possible, an incongruous context for one of the sentences that could influence our annotators' ratings.

We collected a sub-corpus of 200 contextually embedded groups of two sentences. We tried to keep our data as balanced as possible, drawing from all four "classes" of paraphrase aptness ratings (between 1 to 4) that Bizzoni and Lappin (2018) obtained. We selected 44 pairs of *1* ratings, 51 pairs of *2*, 43 pairs of *3* and 62 pairs of *4*.

We then used AMT crowd sourcing to rate the contextualized paraphrase pairs, so that we could observe the effect of document context on assessments of metaphor paraphrase aptness.

To test the reproducibility of Bizzoni and Lappin (2018)'s ratings, we launched a pilot study for 10 original, non-contextually embedded pairs, selected from all four "categories" of aptness. We observed that the annotators provided mean ratings very similar to those reported in Bizzoni and Lappin (2018).

[4]Our annotated data set and the code for our model is available at `https://github.com/yuri-bizzoni/Metaphor-Paraphrase`.

The Pearson coefficent correlation between the mean judgments of our out-of-context pilot annotations and Bizzoni and Lappin (2018)'s annotations for the same pair was over 0.9.

We then conducted an AMT annotation task for the 200 contextualized pairs. On average, 20 different annotators rated each pair. We considered as "rogue" those annotators who rated the large majority of pairs with very high or very low scores, and those who responded inconsistently to two "trap" pairs. After filtering out the rogues, we had an average of 14 annotators per pair.

3 Annotation Results

We found a Pearson correlation of 0.81 between the in-context and out-of-context mean human paraphrase ratings for our two corpora. This correlation is virtually identical to the one that Bernardy et al. (2018) report for mean acceptability ratings of out-of-context to in-context sentences in their crowd source experiment. It is interesting that a relatively high level of ranking correspondence should occur in mean judgments for sentences presented out of and within document contexts, for two entirely distinct tasks.

Our main result concerns the effect of context on mean paraphrase judgment. We observed that it tends to flatten aptness ratings towards the centre of the rating scale.

Of the metaphors that had been considered highly apt (average rounded score of *4*) in the context-less pairs, 71.1% received a more moderate judgment (average rounded score of *3*). On the other hand, the reverse movement was rare: only 5% of pairs rated *3* out of context (2 pairs) was boosted to a mean rating of *4* in context.

At the other end of the scale, 68.2% of the metaphors judged at *1* category of aptness out of context were raised to a mean of *2* in context, while only the 3.9% of pairs rated *2* out of context were lowered to *1* in context.

Ratings at the middle of the scale - *2* (defined as semantically related non-paraphrases) and *3* (imperfect or loose paraphrases) - remained largely stable, with little movement in either direction. 9.8% of pairs rated *2* were re-ranked as *3* when presented in context, and 10% of pairs ranked at *3* changed to *2*.

It seems that context tends to "improve" metaphors with a low level of aptness, but lowers the judgments on metaphors with a high level of aptness.

The division between *2* and *3* separates paraphrases from non-paraphrases. Our results suggest that this binary rating of paraphrase aptness was not strongly affected by context. Context operates at the extremes of our scale, raising low aptness ratings and lowering high aptness ratings. This effect is clearly indicated in the regression chart in Fig 1.

This effect of context on human ratings is very similar to the one reported in Bernardy et al. (2018). They find that sentences rated as ill formed out of context are in part improved when they are presented in their document contexts. However the mean ratings for sentences judged to be highly acceptable out of context declined when assessed in context. Bernardy et al. (2018)'s linear regression chart for the correlation between out-of-context and in-context acceptability judgments as collected in their survey looks remarkably like our Fig 1. There is, then, a striking parallel in the compression pattern that context appears to exert on human judgments for two entirely different linguistic properties.

This pattern requires an explanation. Bernardy et al. (2018) suggest that adding context causes speakers to focus on broader semantic and pragmatic issues of discourse coherence, rather than simply judging syntactic well formedness (measured as naturalness) when a sentence is considered in isolation. On this view, compression of rating results from a pressure to construct a plausible interpretation for any sentence within its context. If this is the case, an analogous process may generate the same compression effect for metaphor aptness assessment of sentence pairs in context. Speakers may attempt to achieve broader discourse coherence when assessing the metaphor-paraphrase aptness relation in a document context. Out of context they focus more narrowly on the semantic relations between a metaphorical sentence and its paraphrase candidate. Therefore, this relation is the centre of a speaker's concern and receives more fine-grained assessment when considered out of context than in context.

However, a second possibility is that adding context to the aptness task increases the general cognitive

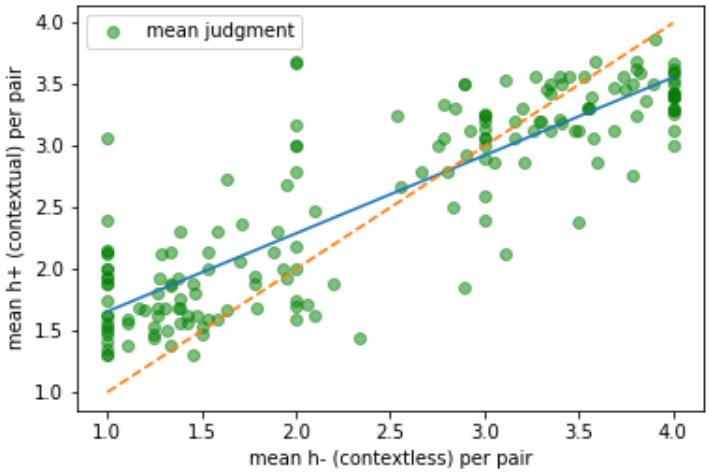

Figure 1: In-context and out-of-context mean ratings. Points above the broken diagonal line represent sentence pairs which received a higher rating when presented in context. The total least-square linear regression is shown as the second line.

load involved in processing the sentence. This effect may also cause hearers/readers to focus less on the properties of the sentences for which a judgment is solicited, and more on processing the entire discourse unit. Such a shift in focus might also produce the observed compression effect, but for different reasons than those that the pragmatic discourse coherence explanation proposes. This issue clearly requires further research. We discuss these two possible interpretations in more in detail in Section 7.

4 Modelling Paraphrase Judgments in Context

We use the DNN model described in Bizzoni and Lappin (2018) to predict aptness judgments for in context paraphrase pairs. It has three main components:

1. Two encoders that learn the representations of two sentences separately
2. A unified layer that merges the output of the encoders
3. A final set of fully connected layers that operate on the merged representation of the two sentences to generate a score. Our pairs are evaluated through this final score.

The encoder for each pair of sentences taken as input is composed of two parallel "Atrous" Convolutional Neural Networks (CNNs) and LSTM RNNs, feeding two sequenced fully connected layers.

The encoder is preloaded with the lexical embeddings from Word2vec Mikolov et al. (2013). The sequences of word embeddings that we use as input provides the model with dense word-level information, while the model tries to generalize over these embedding patterns.

The combination of a CNN and an LSTM allows us to capture both long-distance syntactic and semantic relations, best identified by a CNN, and the sequential nature of the input, most efficiently identified by an LSTM. Several existing studies, cited in Bizzoni and Lappin (2017), demonstrate the advantages of combining CNNs and LSTMs to process texts, and show that using these two architectures together has a positive effect on language processing.

The model produces a single classifier value between 0 and 1. We transform this score into a binary output of 0 or 1 by applying a threshold of 0.5 for assigning 1. In this way, we can use the model's output for two evaluation methodologies: classification and ranking.

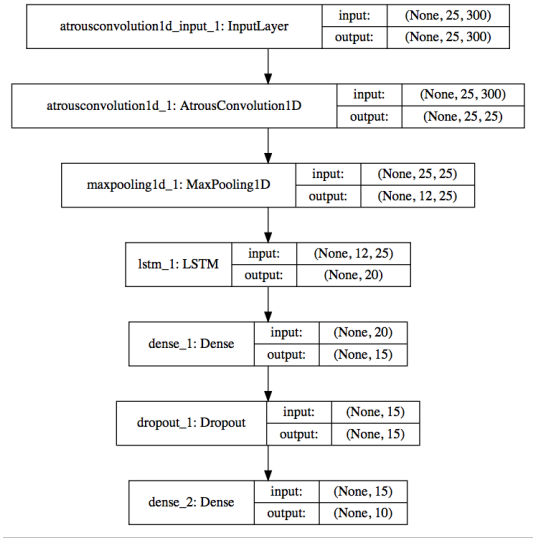

Figure 2: DNN encoder for predicting metaphorical paraphrase aptness from Bizzoni and Lappin (2018). Each encoder represents a sentence as a 10-dimensional vector. These vectors are concatenated to compute a single score for the pair of input sentences.

The architecture of the model is given in Fig 2.

We use the same general protocol as Bizzoni and Lappin (2018) for training with supervised learning, and testing the model.

Following the methodology applied in Bernardy et al. (2018), the input to the encoders is the concatenation of the word embeddings of the whole paragraph (context and focus sentence).

Using Bizzoni and Lappin (2018)'s out-of-context metaphor dataset and our contextualized extension of this set, we apply four variants of the training and testing protocol.

1. Training and testing on the in-context dataset.
2. Training on the out-of-context dataset, and testing on the in-context dataset.
3. Training on the in-context dataset, and testing on the out-of-context dataset.
4. Training and testing on the out-of-context dataset (Bizzoni and Lappin (2018)'s original experiment provides the results for out-of-context training and testing).

When we train or test the model on the out-of-context dataset, we use Bizzoni and Lappin (2018)'s original annotated corpus of 800 metaphor-paraphrase pairs. The in-context dataset contains 200 annotated pairs. As for the baseline, we rely on Bizzoni and Lappin (2018)'s earlier work on paraphrase, where, together with several alternative versions of the neural model, a baseline relying on vector cosine similarity between sentences is provided, and outperformed by the model.

5 MPAT Modelling Results

We use the model both to predict binary classification of a metaphor paraphrase candidate, and to generate gradient aptness ratings on the 4 category scale (see Bizzoni and Lappin (2018) for details). A positive binary classification is accurate if it is $\geq$ a 2.5 mean human rating. The gradient predictions are derived from the softmax distribution of the output layer of the model. The results of our modelling experiments are given in Table 1.

The main result that we obtain from these experiments is that the model learns binary classification to a reasonable extent on the *in-context* dataset, both when trained on the same kind of data (in-context

Training set	Test set	F-score	Correlation
With-context*	With-context*	0.68	-0.01
Without-context	With-context	**0.72**	**0.3**
With-context	Without-context	0.6	0.02
Without-context	Without-context	0.74	0.75

Table 1: F-score binary classification accuracy and Pearson correlation for three different regimens of supervised learning. The * indicates results for a set of 10-fold cross-validation runs. This was necessary in the first case, when training and testing are both on our small corpus of in-context pairs. In the second and third rows, since we are using the full out-of-context and in-context dataset, we report single-run results. The fourth row is Bizzoni and Lappin (2018)'s best run result. (Our single-run best result for the first row is an F-score of 0.8 and a Pearson correlation 0.16).

pairs), and when trained on Bizzoni and Lappin (2018)'s original dataset (out-of-context pairs). However, the model does not perform well in predicting gradient in-context judgments when trained on in-context pairs. It improves slightly for this task when trained on out-of-context pairs.

By contrast, it does well in predicting both binary and gradient ratings when trained and tested on out-of-context data sets.

Bernardy et al. (2018) also note a decline in Pearson correlation for their DNN models on the task of predicting human in-context acceptability judgments, but it is less drastic.

They attribute this decline to the fact that the compression effect renders the gradient judgments less separable, and thus harder to predict. A similar, but more pronounced version of this effect may account for the difficulty that our model encounters in predicting gradient in-context ratings. The binary classifier achieves greater success for these cases because its training tends to polarise the data in one direction or the other.

We also observe that the best combination seems to consist in training our model on the original out-of-context dataset and testing it on the in-context pairs. In this configuration we reach an F-score (0.72) only slightly lower than the one reported in Bizzoni and Lappin (2018) (0.74), and we record the highest Pearson correlation, 0.3 (which is still not strong, compared to Bizzoni and Lappin (2018)'s best run, 0.75[5]). This result may partly be an artifact of the the larger amount of training data provided by the out-of-context pairs.

We can use this variant (out-of-context training and in-context testing) to perform a fine-grained comparison of the model's predicted ratings for the same sentences in and out of context. When we do this, we observe that out of 200 sentence pairs, our model scores the majority (130 pairs) higher when processed in context than out of context. A smaller but significant group (70 pairs) receives a lower score when processed in context. The first group's average score *before adding context* (0.48) is consistently lower than that of the second group (0.68). Also, as Table 2 indicates, the pairs that our model rated, *out of context*, with a score lower than 0.5 (on the model's softmax distribution), received on average a higher rating *in context*, while the opposite is true for the pairs rated with a score higher than 0.5. In general, sentence pairs that were rated highly out of context receive a lower score in context, and vice versa. When we did linear regression on the DNNs in and out of context predicted scores, we observed substantially the same compression pattern exhibited by our AMT mean human judgments. Figure 3 plots this regression graph.

6 Related Cognitive Work on Metaphor Aptness

Tourangeau and Sternberg (1981) present ratings of aptness and comprehensibility for 64 metaphors from two groups of subjects. They note that metaphors were perceived as more apt and more comprehensible to the extent that their terms occupied similar positions within dissimilar domains. Interestingly,

[5]It is also important to consider that their ranking scheme is different from our design: the Pearson correlation reported there is the average of the correlations over all groups of 5 sentences present in the dataset.

OOC score	Number of elements	OOC Mean	OOC Std	IC Mean	IC Std
0.0-0.5	112	0.42	0.09	0.54	0.1
0.5-1.0	88	0.67	0.07	0.64	0.07

Table 2: We show the number of pairs that received a low score out of context (first row) and the number of pairs that received a high score out of context (second row). We report the mean score and standard deviation (Std) of the two groups when judged out of context (OOC) and when judged in context (IC) by our model. The model's scores range between 0 and 1. As can be seen, the mean of the low-scoring group rises in context, and the mean of the high-scoring group decreases in context.

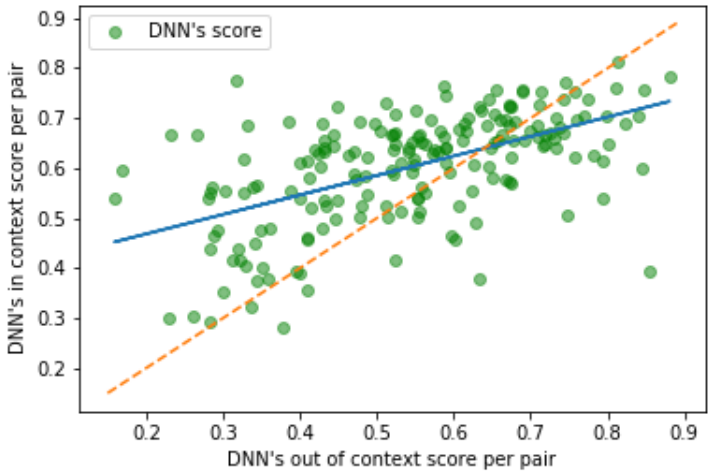

Figure 3: In-context and out-of-context ratings assigned by our trained model. Points above the broken diagonal line represent sentence pairs which received a higher rating when presented in context. The total least-square linear regression is shown as the second line.

Fainsilber and Kogan (1984) present experimental results in support of the claim that imagery does not clearly correlate with metaphor aptness. Aptness judgments are also subject to individual differences.

Blasko (1999) points to such individual differences in metaphor processing. She asked 27 participants to rate 37 metaphors for difficulty, aptness and familiarity, and to write one or more interpretations of the metaphor. Subjects with higher working memory span were able to give more detailed and elaborate interpretations of metaphors. Familiarity and aptness correlated with both high and low span subjects. For high span subjects aptness of metaphor positively correlated with number of interpretations, while for low span subjects the opposite was true.

McCabe (1983) analyses the aptness of metaphors with and without extended contexts. She finds that domain similarity correlates with aptness judgments in isolated metaphors, but not in *contextualized* metaphors. She also reports that there is no clear correlation between metaphor aptness ratings in isolated and in contextualized examples.

Chiappe et al. (2003) study the relation between aptness and comprehensibility in metaphors and similes. They provide experimental results indicating that aptness is a better predictor than comprehensibility for the "transformation" of a simile into a metaphor. Subjects tended to remember similes as metaphors (i.e. remember *the dancer's arms moved like startled rattlesnakes* as *the dancer's arms **were** startled rattlesnakes*) if they were judged to be particularly apt, rather than particularly comprehensible. They claim that context might play an important role in this process. They suggest that context should ease the transparency and increase the aptness of both metaphors and similes.

Tourangeau and Rips (1991) report a series of experiments indicating that metaphors tend to be

interpreted through emergent features that were not rated as particularly relevant, either for the tenor or for the vehicle of the metaphor. The number of emergent features that subjects were able to draw from a metaphor seems to correlate with their aptness judgments.

Bambini et al. (2018) use Event-Related Brain Potentials (ERPs) to study the temporal dynamics of metaphor processing in reading literary texts. They emphasize the influence of context on the ability of a reader to smoothly interpret an unusual metaphor.

Bambini et al. (2016) use electrophysiological experiments to try to disentangle the effect of a metaphor from that of its context. They find that de-contextualized metaphors elicited two different brain responses, $N400$ and $P600$, while contextualized metaphors only produced the $P600$ effect. They attribute the $N400$ effect, often observed in neurological studies of metaphors, to expectations about upcoming words in the absence of a predictive context that "prepares" the reader for the metaphor. They suggest that the $P600$ effect reflects the actual interpretative processing of the metaphor.

This view is supported by several neurological studies showing that the $N400$ effect arises with unexpected elements. This happens, for example, when new presuppositions are introduced into a text in a way not implied by the context (Masia et al. (2017)). It can also occur because of unexpected associations with a noun-verb combination, not indicated by previous context, as when it is preceded by a neutral context (Cosentino et al. (2017)).

7 Conclusions and Future Work

We have observed that embedding metaphorical sentences and their paraphrase candidates in a document context generates a compression effect in human metaphor aptness ratings. Context seems to mitigate the perceived aptness of metaphors in two ways. Those metaphor-paraphrase pairs that were given a very low score out of context tend to receive an increased score in context, while those with very high scores out of context decline in rating when presented in context. At the same time, the demarcation line between paraphrase and non-paraphrase is not particularly blurred by the introduction of extended context around the expression.

As previously observed by McCabe (1983), we found that context has an influence on humans' aptness ratings for metaphors, although, unlike them, we did find a correlation between the two sets of ratings. Chiappe et al. (2003)'s expectation that context should facilitate a metaphor's aptness was supported only in one sense. Aptness increases for low-rated pairs. But it decreases for high-rated pairs.

We applied Bizzoni and Lappin (2018)'s DNN for the MPAT to an in-context test set, experimenting with both out-of-context and in-context training corpora. We obtained reasonable results for binary classification of paraphrase candidates for aptness, but the performance of the model declined sharply for the prediction of human gradient aptness judgments, relative to its performance on a corresponding out-of-context test set. This appears to be the result of the increased difficulty in separating rating categories introduced by the compression effect.

Strikingly, the linear regression analyses of human aptness judgments for in- and out-of-context paraphrase pairs, and of Bizzoni and Lappin (2018)'s DNN predictions for these pairs reveal similar compression patterns. These patterns produce ratings that cannot be clearly separated along a linear ranking scale.

To the best of our knowledge ours is the first study of the effect of context on metaphor aptness on a corpus of this dimension, using crowd sourced human judgments as the gold standard for assessing the predictions of a computational model of paraphrase. We also present the first comparative study of both human and model judgments of metaphor paraphrase for in-context and out-of-context variants of metaphorical sentences.

Finally, the compression effect that context induces on paraphrase judgments corresponds closely to the one observed independently in another task, which is reported in Bernardy et al. (2018). We regard this effect as a significant discovery that increases the plausibility and the interest of our results. The fact that it appears clearly with two tasks involving different sorts of DNNs and distinct learning regimes (unsupervised learning with neural network language models for the acceptability prediction

task, as opposed to supervised learning with our composite DNN for paraphrase prediction) reduces the likelihood that this effect is an artefact of our experimental design.

It is important to note that this shift towards the centre of the scale, recorded both for humans and for our model, is *not* consistent with a simple homogenization effect for the compared items. If the addition of identical context to both sentences just made it harder for the network to see the differences between the two items, we would expect the shift in aptness judgment to go in one direction on the scale. All contextualized pairs should be rated as better paraphrases than their decontextualized equivalents. The same effect should hold for human annotators.

As we suggested earlier, two explanations for the compression effect come to mind. On the first compression is the result of a specifically linguistic phenomenon. In the presence of a larger textual context speakers concentrate on the pragmatic coherence of the discourse, and so they pay less attention to the properties of the sentence for which assessment is solicited. This is the approach that Bernardy et al. (2018) propose. On the second explanation compression is the result of the increase in cognitive load that processing the context imposes.

To distinguish between these accounts it would be interesting to experiment with two different kinds of contexts: a natural one for each sentence, and a random context that is unrelated in content to the sentence. If the cognitive load hypothesis is correct, the compression effect should be present with both types of context, as they each increase processing. However, if the effect appears only with natural contexts, then this result would lend support to the pragmatic coherence hypothesis. Random contexts do not generally facilitate coherent discourse interpretations, and so we would expect speakers to exhibit a tendency to focus on the naturalness of the test sentence in isolation. This should reduce or cancel the observed compression effect. One of our main concerns in future research will be to achieve a better understanding of the compression effect of context on human judgments and DNN models.

While our dataset is still small, we are presenting an initial investigation of a phenomenon which is, to date, little studied. We are working to enlarge the dataset. In future work we will expand both our in- and out-of-context annotated metaphor-paraphrase corpora. While the corpus we used contains a number of hand crafted examples, it would be preferable to find these example types in natural text, and we are working on this. We are seeking to expand the size of the data set. It will also be useful to conduct qualitative analyses on the kinds of metaphors and similes that are more prone to a context-induced rating switch. We intend to improve the reliability of our modelling experiments by using alternative DNN architectures for the MPAT.

Acknowledgments

We are grateful to our colleagues in the Centre for Linguistic Theory and Studies in Probability (CLASP), FLoV, at the University of Gothenburg for useful discussion of some of the ideas presented in this paper. We would also like to thank two anonymous reviewers and Matthew Purver for their insightful and detailed comments on an earlier draft. The research reported here was done at CLASP, which is supported by a 10 year research grant (grant 2014-39) from the Swedish Research Council.

References

Bambini, V., C. Bertini, W. Schaeken, A. Stella, and F. Di Russo (2016). Disentangling metaphor from context: an erp study. *Frontiers in psychology 7*, 559.

Bambini, V., P. Canal, D. Resta, and M. Grimaldi (2018). Time course and neurophysiological underpinnings of metaphor in literary context. *Discourse Processes*, 1–21.

Bernardy, J.-P., S. Lappin, and J. H. Lau (2018). The influence of context on sentence acceptability judgmements. *Proceedings of ACL 2018, Melbourne, Australia*, 456–461.

Bizzoni, Y. and S. Lappin (2017). Deep learning of binary and gradient judgements for semantic paraphrase. In *IWCS 2017 - 12th International Conference on Computational Semantics - Short papers, Montpellier, France, September 19 - 22, 2017*.

Bizzoni, Y. and S. Lappin (2018). Predicting human metaphor paraphrase judgments with deep neural networks. *Proceedings of The Workshop on Figurative Language Processing, NAACL 2018, New Orleans LA*, 45–55.

Blasko, D. G. (1999). Only the tip of the iceberg: Who understands what about metaphor? *Journal of Pragmatics 31*(12), 1675–1683.

Chiappe, D. L., J. M. Kennedy, and P. Chiappe (2003). Aptness is more important than comprehensibility in preference for metaphors and similes. *Poetics 31*(1), 51–68.

Cosentino, E., G. Baggio, J. Kontinen, and M. Werning (2017). The time-course of sentence meaning composition. n400 effects of the interaction between context-induced and lexically stored affordances. *Frontiers in psychology 8*, 813.

Fainsilber, L. and N. Kogan (1984). Does imagery contribute to metaphoric quality? *Journal of psycholinguistic research 13*(5), 383–391.

Glucksberg, S. (2008). How metaphors create categories–quickly. *The Cambridge handbook of metaphor and thought*, 67–83.

Masia, V., P. Canal, I. Ricci, E. L. Vallauri, and V. Bambini (2017). Presupposition of new information as a pragmatic garden path: Evidence from event-related brain potentials. *Journal of Neurolinguistics 42*, 31–48.

McCabe, A. (1983). Conceptual similarity and the quality of metaphor in isolated sentences versus extended contexts. *Journal of Psycholinguistic Research 12*(1), 41–68.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 26*, pp. 3111–3119. Curran Associates, Inc.

Sam, G. and H. Catrinel (2006). On the relation between metaphor and simile: When comparison fails. *Mind & Language 21*(3), 360–378.

Tourangeau, R. and L. Rips (1991). Interpreting and evaluating metaphors. *Journal of Memory and Language 30*(4), 452–472.

Tourangeau, R. and R. J. Sternberg (1981). Aptness in metaphor. *Cognitive psychology 13*(1), 27–55.