When is a Metaphor Actually Novel? Annotating Metaphor Novelty in the Context of Automatic Metaphor Detection

Sebastian Reimann and **Tatjana Scheffler**
Ruhr University Bochum
Department for German Language and Literature
Bochum, Germany
{sebastian.reimann,tatjana.scheffler}@rub.de

Abstract

We present an in-depth analysis of metaphor novelty, a relatively overlooked phenomenon in NLP. Novel metaphors have been analyzed via scores derived from crowdsourcing in NLP, while in theoretical work they are often defined by comparison to senses in dictionary entries. We reannotate metaphorically used words in the large VU Amsterdam Metaphor Corpus based on whether their metaphoric meaning is present in the dictionary. Based on this, we find that perceived metaphor novelty often clash with the dictionary based definition. We use the new labels to evaluate the performance of state-of-the-art language models for automatic metaphor detection and notice that novel metaphors according to our dictionary-based definition are easier to identify than novel metaphors according to crowdsourced novelty scores. In a subsequent analysis, we study the correlation between high novelty scores and word frequencies in the pretraining and finetuning corpora, as well as potential problems with rare words for pre-trained language models. In line with previous works, we find a negative correlation between word frequency in the training data and novelty scores and we link these aspects to problems with the tokenization of BERT and RoBERTa.

1 Introduction

Most data for training and evaluating automatic metaphor detection systems contains binary labeling that only distinguishes between metaphoric and literal tokens. The distinction between novel and conventionalized metaphor has received some, albeit little attention in the context of annotating data for automatic metaphor detection (Do Dinh et al., 2018; Neidlein et al., 2020; Djokic et al., 2021). There is however still a lack of publicly available, large-scale annotations that make this distinction. So far, all existing work on novel metaphor in NLP has used crowdsourced annotations of metaphors from the VU Amsterdam Metaphor Corpus (VUAMC; Steen et al., 2010).

This lack of resources and research in general is problematic in several ways: Neidlein et al. (2020) suggest that considerable numbers of novel metaphors remain undetected by metaphor detection systems. What is more, the notion of when a metaphor can be considered *novel* varies in the literature. Besides crowdsourced novelty scores, dictionary-based approaches are frequently used. For example, Reijnierse et al. (2018) and Egg and Kordoni (2022) consider a metaphor to be not conventionalized, if the sense in which it is used cannot be found in a dictionary. Krennmayr (2006) already argued for the use of corpus-based dicitionaries in metaphor analysis since they provide transparency and replicability, which would also be an advantage in annotation of metaphor novelty. Do Dinh et al. (2018) even stress the need to compare their novelty scores with dictionary entries.

Identifying novel metaphors is especially important for computational approaches to figurative language, since other forms of metaphor can often be easily captured by distributional approaches to meaning. For example, depending on the subsection for the respective register, between 33% and 45% of the prepositions in the VUAMC were used in a metaphorical way (Steen et al., 2010). These are often temporal prepositions such as *in July*, which are seen as spatial words such as *in* transferred into a temporal sense. These highly conventionalized metaphors are not usually of interest in computational approaches to figurative language.

In this study, we evaluate crowdsourced novelty scores by investigating how well they align with dictionary-based definitions of *novel metaphor* (RQ1). We show that there are systematic discrepancies between these two definitions. Based on our findings, we develop a new label for novel metaphor and use this label to evaluate current

Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII), pages 87–97

state-of-the-art metaphor detection systems specifically on novel metaphors (RQ2). We will make these new annotations on metaphor novelty publicly available. We find that the systems appear to find a higher share of novel metaphors defined by our dictionary-based label, compared to a distinction based on the crowdsourced novelty score and a threshold. We link our results to the findings of Neidlein et al. (2020) and raise the concern that crowdsourced novelty scores may mainly trace the overall rarity of the words and should be replaced by deeper estimates of the unconventionality of metaphors in future work.

2 Previous Work

2.1 Metaphor Annotation

The Metaphor Identifcation Procedure Vrjie Universiteit Amsterdam (MIPVU; Steen et al., 2010) was widely used to obtain binary metaphor annotations. MIPVU identifies so-called metaphor related words (MRWs) and distinguishes between indirect and direct MRWs. Indirect MRWs are identified by comparing the contextual meaning of a word with available senses in the dictionary: if one semantically related meaning in the dictionary can be considered more "basic" (more concrete or human-related) than the contextual interpretation, the word is seen as potentially metaphoric. In (1), the meaning of the word *brilliant* equates to the sense *extremely clever or skillful* in the Longman Dictionary of Contemporary English (LDOCE) (Longman, 2023). The more concrete meaning is *brilliant light or colour is very bright and strong*. As both brightness and intelligence are seen as positive, we can conclude that *brilliant* is an MRW according to MIPVU.

(1) This was a brilliant move.

In direct MRWs, there is no contrast between the contextual and a more basic meaning of a word but the word still is part of a mapping between two domains. This is for example the case in metaphoric comparisons, like (2). Here *proud* and *man* are technically used in their most basic meaning. However, by comparison and lexical signals (*like*) the domain TREE is mapped onto the domain HUMAN/MAN.

(2) This tree stands like a proud man.

Steen et al. (2010) applied MIPVU to the BNC-Baby Corpus in order to create the VUAMC, which was then used as training and test data in the Metaphor Detection Shared Tasks 2018 and 2020 (Leong et al., 2018, 2020) and other studies on automatic metaphor detection.

There exist, however, approaches to metaphor annotation that go beyond a mere distinction between metaphoric and literal. The LCC dataset (Mohler et al., 2016) contains word pairs in four languages (English, Spanish, Russian and Farsi) annotated on a four-point-scale according to their metaphoricity. The judgement on *metaphoricity* includes how easy the source domain can be perceived by the senses, how vivid the used language is, how frequently the metaphor may be encountered. The degree of conventionalization is thus to some extent taken into account in this annotation of metaphoricity. However, it is only one of several factors that influence metaphoricity and Mohler et al. (2016) do not present annotations on degree of conventionalization of a metaphor isolated from the other aspects of metaphoricity.

Another fine-grained distinction can be drawn between deliberate metaphors, which are meant to be understood as metaphors (Reijnierse et al., 2018), and non-deliberate ones. The Deliberate Metaphor Identification Procedure (DMIP) of Reijnierse et al. (2018) is a way to systematically annotate an MRW (previously identified via MIPVU) on potential deliberateness by checking whether its source domain is needed to actually understand the metaphor in its context, which, according to Reijnierse et al. (2018), is always the case for novel and unconventionalized metaphors. Reijnierse et al. (2018) define *novel metaphor* based on whether the metaphoric sense of a word is represented in the dictionary. In Reijnierse et al. (2019), the entire VUAMC was annotated for potentially deliberateness of an MRW according to DMIP. The labels in Reijnierse et al. (2019) only present a binary distinction between *potentially deliberate* and *non-deliberate*.

The first approach to provide annotations on perceived metaphor novelty in the VUAMC was by Parde and Nielsen (2018), who obtained novelty annotations for syntactically related word pairs from the VUAMC. Here, on the one hand, a smaller dataset of about 3,000 pairs was annotated by trained annotators and a larger dataset (about 18,000 pairs) was annotated by crowdworkers. In both cases, annotators needed to rate the word pairs on a scale from 0 to 3, where 0 marks non-metaphoric instances and 3 highly novel instances. The disagreements for the trained annota-

tors were resolved by discussion and a third annotator, whereas the crowdworkers' annotations were automatically aggregated to a final annotation on the same scale.

Do Dinh et al. (2018) also provided annotations on metaphor novelty for the VUAMC. They focused however on annotations for each token labeled as MRW and not on syntactic pairs. Here, crowdworkers were asked to rank MRWs from the VUAMC according to how novel they are. These annotations were aggregated and transformed into scores ranging from -1 (very conventionalized) to 1 (very novel). The authors moreover explored how their novelty scores correlate with word frequency, concreteness scores and potential for metaphoricity (POM) (Del Tredici and Bel, 2016), where they observed a correlation of novelty annotations with frequency and POM but not with concreteness.

For the distinction between metaphor and nonsense, Pedinotti et al. (2021) released a dataset of 300 items, 100 metaphoric sentences, 100 literal and 100 nonsensical statements. The metaphors in their dataset were also grouped into creative (i.e. novel) and conventional metaphors. Unfortunately, Pedinotti et al. (2021) did not explain further how they exactly defined the terms *creative* and *conventional*. Additionally, they provided annotations by crowdworkers on semantic plausibility, that is how meaningful a sentence is, and metaphoricity. Here, novel metaphors were considered less plausible than conventional metaphors by human judges and were rated more metaphorical than conventional metaphors.

2.2 Automatic Metaphor Detection and Novel Metaphor

In recent years, large pre-trained language models were dominating the field of automatic metaphor detection. This is exemplified by the results of the 2020 shared task on metaphor detection (Leong et al., 2020), where the five best-performing approaches all used some variation of BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019). Approaches such as DeepMet by Su et al. (2020) that intends to simulate a reading comprehension with two RoBERTa encoder layers and linguistic features such as POS-tags, and MelBERT (Choi et al., 2021), which emulates two theoretical methods for identifying metaphors in text, Metaphor Identification Procedure (MIP) (Pragglejaz Group, 2007), a predecessor of MIPVU and Selectional Preference Violation (SPV) (Wilks, 1975), where metaphors are identified by looking at whether a word semantically matches its context, achieve impressive F1-scores of more then 75 on a binary distinction between MRW and literal use.

When it comes to finding novel metaphors, much less work has been carried out. One early attempt to detect novel metaphors was conducted by Haagsma and Bjerva (2016). They employ selectional preference violations (Wilks, 1975) by extracting the frequencies of verb-noun pairs from a large Wikipedia corpus, semantically clustering them and calculating selectional preference metrics. These serve as inputs for a logistic regression classifier that was then tested on detecting metaphors in VUAMC data. In the evaluation, they however conclude that, back then, they were not able to clearly state the effectiveness of their system for novel metaphors since labels on metaphor novelty were not available for the VUAMC.

Besides providing annotations on metaphor novelty, Do Dinh et al. (2018) also built a system to predict metaphor novelty scores. For this, they used a BiLSTM with dependency-based word embeddings that achieved a mean absolute error of 0.166. Adding the features with which novelty scores correlate (frequency and POM) however only led to an improved MAE of 0.163.

Neidlein et al. (2020) conducted an extensive analysis of metaphor detection systems based on large language models. One focus was on how well these models were able to handle novel metaphors. For this, they set various thresholds for the scores of Do Dinh et al. (2018) and looked at the accuracy for MRWs with a score higher than that threshold. They observed that the higher the threshold, the lower the accuracy, and concluded that novel metaphors are more difficult than conventionalized metaphors. They moreover looked at word frequencies in the training set and found that the performance is lower on metaphoric words that have rarely been seen in fine-tuning and that a high number of conventionalized metaphors have high word frequencies in the training data. Moreover, the models evaluated in (Neidlein et al., 2020) performed better if derivational variants (such as *warm* and *warmth*) have been seen in training before.

To see if BERT can distinguish between metaphors and nonsense, Pedinotti et al. (2021), on the one hand calculated the pseudo-log-likelihood (PLL) score (Wang and Cho, 2019) of metaphoric, literal and nonsensical sentences and the cosine similarity with so-called *landmarks*, words that are

either from the same or a different semantic space than the metaphor in question (Kintsch, 2000). Based on the PLL scores, the model considers novel metaphor significantly more plausible than nonsensical sentences, however less plausible than conventional metaphors. Moreover, it struggles harder to interpret novel metaphor by comparing them with landmarks than to interpret conventional metaphors.

The only effort so far that actually made use of metaphor novelty in training was presented by Djokic et al. (2021). They used a BERT-based classifier that predicted novelty scores in a joint manner together with general binary labels on metaphor (MRW or not MRW). It was trained on data from the VUAMC and then applied to an unlabeled corpus of short stories, which unfortunately does not allow for a systematic evaluation. They only tested the score prediction on the VUAMC test data, which led to a slightly improved MAE of 0.142 compared to the baseline of Do Dinh et al. (2018).

3 Methodology

We systematically evaluate the crowdsourced novelty scores of Do Dinh et al. (2018) by comparing them with our own binary annotation of MRWs that uses a linguistic definition of novel metaphor, according to which a metaphor is considered novel if the contextual meaning of the MRW is not included in a standard dictionary for the language under investigation. Steen et al. (2010) and Reijnierse et al. (2018) used the MacMillan dictionary (Rundell, 2002) as a primary lexical resource in their metaphor annotation. Unfortunately, the online version of the MacMillan dictionary has been shut down in July 2023. In our annotation, we thus use the Longman Dictionary of Contemporary English (LDOCE), which was used by Steen et al. (2010) as a secondary source and which is also corpus-based, as our main resource for checking if the contextual meaning of a word is represented in the dictionary, either as a sense description in the entry of a word or as entry for a fixed expression.

DMIP, introduced in the previous section, considers metaphor novelty (via the availability of sense descriptions in dictionaries) as one criterion for potential deliberateness of a metaphor. It was applied to all MRWs in the VUAMC by Reijnierse et al. (2019). Unfortunately, the authors did not provide any further information on why they consider a metaphor potentially deliberate. Nevertheless, Reijnierse et al. (2019) present novel metaphors according to a dictionary-based definition as a subset of potentially deliberate MRWs. This provides a good heuristic to find MRWs that are potentially novel. We consequently check the contextual meaning in the dictionary for:

- MRWs that were annotated as potentially deliberate in Reijnierse et al. (2019) with the exception of direct MRWs, as their contextual meaning is not different from a basic meaning in the dictionary
- MRWs marked as non-deliberate but which received scores over the previously used threshold of 0.45 (a modification by Djokic et al. (2021) of the originally used threshold of 0.5 in Do Dinh et al. (2018)) for novelty. Since the annotations on potential deliberateness are supposed to include all novel metaphors, these cases already represent an interesting clash in annotations since the scores here suggest high novelty but the annotations of Reijnierse et al. (2019), which treat them as non-deliberate, suggest otherwise.

We consider MRWs with low novelty scores and which are not marked potentially deliberate to be not novel. Given the annotations of Reijnierse et al. (2019) we can assume that the availability of a dictionary entry was already considered for labeling them non-deliberate and thus also conventionalized.

All listed cases are manually checked by two annotators: a student assistant trained in metaphor annotation and the first author of this paper. If the contextual meaning of the MRW is not found in the dictionary, the MRW receives the label *novel*, otherwise it is considered *conventionalized*.

After our additional dictionary-based novelty annotation, we conduct a survey of model performance in a similar fashion to Neidlein et al. (2020). We conduct reruns of metaphor detection systems and then compare their performance. For this, we chose DeepMet (Su et al., 2020), because of its strong performance in the 2020 Metaphor Detection Shared Task. We moreover selected MelBERT (Choi et al., 2021) since it achieves competitive performance with DeepMet, while its architecture is more strongly motivated by linguistic theories on metaphor. Here, the layer inspired by SPV is particularly interesting, as already Haagsma and

Bjerva (2016) attempted to use SPV for the detection of novel metaphor. For MelBERT we therefore test the entire architecture as well as both layers in isolation. Finally, we considered the model used in Djokic et al. (2021), since it was designed with the specific goal of finding novel metaphors. We evaluate their model with both the joint objective as well as with only the metaphor detection task in training, in the following referred to as Djokic (joint pred.) and Djokic (met. only), respectively.

The models are trained on the binary classification task (metaphoric vs. literal), with the data from the VUAMC as in the 2020 Metaphor Detection Shared Task (including the same training-test splits) and the same hyperparameters as in the respective original papers. All models use the BERT and RoBERTa implementations from the HuggingFace Transformer library (Wolf et al., 2020): DeepMet and MelBERT use *roberta-base* and the models by Djokic et al. (2021) use *bert-base-cased*. In the evaluation, following Neidlein et al. (2020), we then look at the share of novel metaphors (both according to our definition and according to a novelty score threshold of 0.45) that was detected by the model.

4 Results

4.1 Dictionary-Based Annotation of Novel Metaphors

In total we re-annotated 1160 MRWs with our dictionary based definition of *novel metaphor*. When deciding on whether a dictionary entry for a specific contextual meaning exists or not, the two annotators reached relatively robust agreement of Cohen's $\kappa = 0.73$. Instances for which we disagreed were revisited on a case-to-case basis and a consensus decision was reached.

Table 1 shows the detailed results of our comparison. Overall, we can see that a substantial number of MRWs whose contextual meaning is not represented in the dictionary would be ignored if we applied a threshold of 0.45. We can however see that the share of novel metaphors according to the dictionary-based definition rises with higher crowdsourced novelty since the vast majority of MRWs with scores lower than 0.1 have a conventionalized sense description in the dictionary. This picture however changes with higher scores and for scores only slightly below the threshold, the majority may already be considered novel according to a dictionary-based definition. This suggests that, while there may be some correlation between the two ways of annotating metaphor novelty, defining novelty via crowdsourced scores and a set threshold ignores a wide range of metaphorically used words without a corresponding sense in the dictionary. Table 2 shows three such examples, where a dictionary entry for the respective meaning in the sentence was not found but which eventually would not be considered novel when only looking at novelty scores and the threshold.

The entry in the LDOCE for *pollution* refers either to the process or the substances that make the water, the air or the soil dirty and to the fixed expressions of *noise pollution* and *light pollution*. The author of the example sentence uses pollution to refer to something they perceive as immoral. The sense descriptions for *gulp* explicitly refer to a human activity, either to swallowing or taking in breaths. Here on the other hand, one of these activities is in a novel way ascribed to the *soil*. *Somersault* is in the dictionary described as a bodily movement by a person and not by an organ. *Soupy* only has one meaning (*having a thick liquid quality like soup*), which is obviously not fit to describe music as in the example.

The largest clash between the two definitions of metaphor novelty may however be observed through MRWs that were marked as non-deliberate in Reijnierse et al. (2019) but received novelty scores of over 0.45. Applying this threshold would consider them novel, but the overwhelming majority actually has a sense descriptions in the dictionary. This is illustrated by Example 3, which received a novelty score of 0.545 but for which it can be argued that the use of *gripped* is equivalent to the second entry in the LDOCE dictionary for *to grip* (power and control over someone or something), which would render it conventionalized.

(3) He rejects charges that he was partly responsible for the ‘ casino atmosphere ’ that **gripped** US corporate life in the early 1980s .

A final observation is that a dictionary-based definition leads to a lower number of novel metaphors, compared to defining novelty via crowdsourced scores and a threshold. The former would lead to 421 (318 in the training set and 103 in the test set), compared to 536 novel MRWs (385 in the training and 151 in the test set) according to scores only.

novelty score	potentially delib.	total	in dictionary	not in dictionary (%)
<0.1	yes	244	189	55 (22.54%)
0.1–0.2	yes	88	49	39 (44.31%)
0.2–0.45	yes	292	104	188 (64.38%)
>0.45	yes	113	27	86 (76.10%)
>0.45	no	423	370	53 (12.53%)

Table 1: Overview over MRWs that were annotated by us and if they were found in the dictionary.

Example Sentence	Novelty
The wastes include lindane [...] and even pornography (a different kind of **pollution**).	0.103
You can almost hear the soil **gulping**.	0.303
Paula 's stomach turned a **somersault** .	0.412
The voice of rock'n'roll, in contrast, is almost unrelievedly **soupy**.	0.441

Table 2: Examples of novel MRWs and their respective novelty scores, with metaphorically used words in bold.

4.2 Analysis of Model Performance on Novel Metaphors

Table 3 shows the results of our metaphor detection experiments. We first observe that novel metaphors according to our dictionary-based definition appear to be easier to find than novel metaphors based on crowdsourced scores. The recall for novel metaphors is still worse than the recall for all metaphors but higher that the recall for novel metaphors according to the crowdsourcing threshold.

This would suggest, on the one hand, that novel metaphors are less of a problem for metaphor detection systems than previously assumed by Neidlein et al. (2020). On the other hand, they still remain harder to find than conventionalized metaphors and especially the continued, mostly poor, results for words with high novelty scores hint at other problems. We discuss them in the next section.

Comparing the different model architectures to each other, we can see that DeepMet outperformed the other approaches. Interestingly, the joint prediction of novelty scores and metaphoricity did not help in finding metaphors since adding the loss from the novelty score prediction even led to a minor drop in overall performance for the model of Djokic et al. (2021). Moreover, contrary to previous assumptions, the linguistically motivated architecture of Choi et al. (2021) performed worse than the other models on novel metaphors. Despite previous assumptions that SPV might be suitable to detect metaphor novelty (Haagsma and Bjerva, 2016), MelBERT with only the SPV layer found the lowest share of novel metaphors.

One hypothesis for the poor recall of MelBERT's SPV layer when it comes to novel metaphors may lie in the particular implementation of SPV. It compares the representation of the word in its context with the embedding of the [CLS]-token, representing the entire sentence. This raises doubts about whether it is enough to represent a semantic clash between the word and the context in which it is used.

5 Discussion

5.1 Subjectivity

Our results have shown that perceived novelty and availability of dictionary entries indeed diverge. We now discuss reasons that may cause this difference. On the one hand, we raise the possibility that the perception of untrained annotators might still be too subjective to be solely taken into account when drawing conclusions on the novelty of a metaphor. This can be exemplified by the two instances of *block* in 4. MIPVU considers these two tokens to be separate MRWs and it can be assumed that they were treated separately in the crowdsourcing annotation of Do Dinh et al. (2018) and were annotated by separate annotators. The first *block* has received a novelty score of 0.176 whereas the second *block* received a novelty score of -0.029. This difference seems counterintuitive as they occur in the same context and are used with the same contextual meaning. This example shows that annotations by crowd workers may diverge wildly even for similar instances.

Model	F1	Recall (all metaphor)	Recall (novel/label)	Recall (novel/threshold)
DeepMet	**73.53**	**74.54**	**63.10**	**59.60**
MelBERT (all)	73.07	70.00	56.31	54.96
MelBERT (MIP)	71.77	68.58	57.28	52.98
MelBERT (SPV)	71.28	67.88	53.39	50.33
Djokic (joint pred.)	69.50	72.50	60.19	50.00
Djokic (met. only)	70.60	72.30	62.13	50.67

Table 3: Performance of the selected models, best performance for each metric in bold.

(4) In general, our policy should be to proceed with building our state block by block[...]

5.2 Word Frequency

Word frequency is another factor to be considered when explaining the discrepancy between crowdsourced and dictionary based definitions of metaphor novelty. Do Dinh et al. (2018) have demonstrated a negative correlation of the crowdsourced novelty score with word frequency in a Wikipedia dump. This suggests that a higher novelty score often indicates that a word is rare. In turn, rare words are rarely seen in the pre-training or fine-tuning process of metaphor detection systems. Neidlein et al. (2020) have already shown that pre-trained language models appear to have problems with rare words in the context of automatic metaphor detection. Their findings also indicate that words with low novelty scores have a tendency to occur frequently and, vice versa, words with high novelty scores occur rarely in the training data. Now, assuming that metaphoric words with novelty scores over 0.45 indeed have low frequencies in the training corpora, it seems logical that fewer of them are identified as metaphoric.

The Spearman correlation between frequencies in the training data from the VUAMC, used in fine-tuning the models, and the crowdsourced novelty scores indeed is $\rho = -0.612$, clearly supporting previous findings of Do Dinh et al. (2018) and Neidlein et al. (2020). Looking at the data used in pre-training, we conduct a similar analysis with the BookCorpus (Zhu et al., 2015), used for pre-training BERT and RoBERTa in addition to Wikipedia data. The Spearman correlation for both is similar, with $\rho = -0.601$ and the plot in Figure 1 is in line with the findings of Neidlein et al. (2020) and shows that metaphors with high novelty scores almost exclusively occur infrequently in the pre-training data. On the one hand, this suggests that annotators in crowdsourcing may have a higher

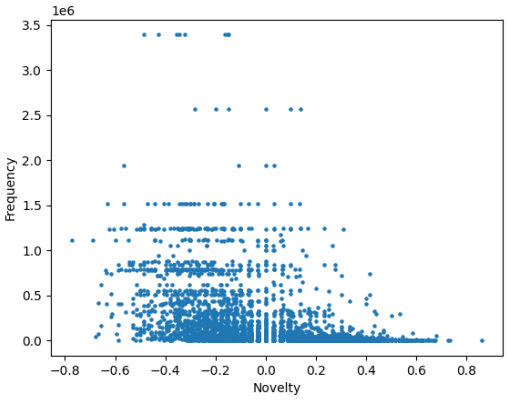

Figure 1: Plot showing the relation between novelty score of a metaphoric word and its frequency in the BookCorpus.

tendency to classify metaphors that involve words they do not frequently encounter as novel, which may be problematic since the example of *pollution* in Table 1 shows that also relatively common words may be used with novel, metaphoric senses. A further investigation of this would however go beyond the scope of this paper.

On the other hand, this relation might have direct implications for model performance. To investigate how rare or unseen words in either pre-training or fine-tuning influence the task of automatic metaphor detection, we perform an error analysis on three examples with crowd sourced novelty scores over 0.45, which did not occur in the pre-training data and which were misclassified by all models. They are shown in Table 4. Indeed, none of these three words were present in the vocabulary files of the used BERT and RoBERTa models.

Moreover, the way how the respective tokenizers split these rare words up into subword tokens may be problematic. As Table 5 shows, the subword tok-

MRW	Sentence	Score
indistinguishable	Whatever else such a strategy may achieve, it certainly does not manage to produce a situation in which children are politically indistinguishable from adults and it rests on premises which, unless they can be defended, gain nothing for any defence to the charge of arbitrariness.	0.485
millipedes	After all, Mancunians and visitors to the Manchester conurbation are going to have to look at these mechanical millipedes for well into the twenty-first century.	0.531
Wriggling	Wriggling across country on the D216 to Port-d'Envaux, you come to two more chateaux : 18th-century Panloy, flaking romantically away on its hillock overlooking a bend in the Charente and, almost next door, the much older, moated Crazannes, half-smothered in amazing flamboyant Gothic carving.	0.625

Table 4: Wrongly classified examples by all model architectures.

MRW	Subwords (BERT)	Subwords (RoBERTa)
indistinguishable	in ##dis ##ting ##ui ##sha ##ble	ind ##ist ##inguishable
millipedes	mill ##ip ##ede ##s	mill ##ip ##edes
wriggling	w ##rig ##gling	w ##rig ##gling

Table 5: Output of the BERT and RoBERTa tokenizer for the MRWs presented in Table 4.

enization of the previously presented cases is not at all in line with the actual morphology of the words. *Distinguish* in *indistinguishable* for example is unrecognizable in the way it has been split into subword units. Neidlein et al. (2020) have shown that it may help if models have seen derivational variants of unseen words in training. However, if they are split like in this example, it is doubtful whether such knowledge can be transferred.

Nayak et al. (2020) have raised further doubts on how well the semantics of a word are represented if the subword tokenization does not fall in line with the actual constituents of the word. For instance, the cosine similarity between *unsaturated* (wrongly tokenized by BERT) and *saturated* is at only 0.30, whereas the cosine distance between *un saturated* (with a space and actually correctly tokenized) and *saturated* is at 0.81. Similar issues may have hurt the viability of the semantic representation for metaphor detection for the examples in Table 5.

In contrast, the MRW *drunkenly* in Example (5) received a relatively high novelty score (0.559) but was still correctly recognized by all models as metaphoric. The tokenizers of both BERT and RoBERTa split the word as *drunken ##ly*, thus retaining its derivational bases and suggesting that the more natural splitting may have played a role in the correct classification.

(5) The plane climbs reluctantly, one set of wings dipping **drunkenly**.

6 Conclusion and Future Work

We systematically compared the crowdsourced novelty scores of Do Dinh et al. (2018) with sense entries available in the dictionary. We evaluated state of the art systems for automatic metaphor detection on their performance on novel metaphor wrt. both score-based and dictionary-based labels. Finally, we discussed these results by taking word frequency as well as the underlying subword tokenization of BERT and RoBERTa into account.

We found that measuring novelty purely by crowdsourced scores and a set threshold ignores a wide range of metaphors for which no conventionalized sense descriptions are available, and in addition considers words used in actually conventionalized senses to be novel. As many theoretical approaches to metaphor use dictionaries as a tool for measuring the degree of conventionalization,

we thus present new annotations of metaphor novelty for the widely used VUAMC that are more in line with these theoretical concepts. We moreover argue that our dictionary-based annotations are more transparent, compared to crowdsourced scores, where the perception of annotators sometimes appears to diverge greatly and where the overall rarity of a word may have a great impact on the annotator's perception of novelty.

Our evaluation of metaphor detection models suggests that rare words may present a larger problem than words used in unusual contexts or with novel meaning. This is reflected in a higher percentage of recognized novel metaphors according to our linguistically grounded label, compared with novelty defined via crowdsourced scores only. Moreover, we found that the underlying word representations of BERT and RoBERTa are often formed from sub-word units that do not reflect the actual morphology of derived words and consequently might not be fit for semantically complex tasks such as metaphor detection, especially when the words are used in unusual, novel contexts.

One potential line for future research would be a closer look at the perception of annotators in crowdsourcing on whether unusual words have a higher tendency to be perceived as novel metaphors, even though they are used in a relatively conventionalized way. Moreover, since the number of truly novel metaphors in the VUA corpus is quite small, further data sets that contain a higher share of novel metaphors and, consequently, evaluation on these data sets is necessary to better judge the performance of metaphor detection systems. Finally, we would like to propose to extend the task of automatic metaphor detection from a binary classification task to a three-way classification by further distinguishing between novel and conventionalized metaphors.

Limitations

One limitation of our study is that the number of novel metaphors in the test set is relatively small, especially for our own dictionary based definition of metaphor novelty. While our model evaluation shows a tendency when it comes to the performance on novel metaphors, a test set containing a larger number of novel metaphors would be needed in order to draw more reliable conclusions on the performance of current language models on detecting novel metaphors.

Finally, such a definition of metaphor novelty relies heavily on the availability of well-structured dictionaries. While this is not a problem for English, it may be difficult to obtain such resources for other languages, especially low-resource languages. Other ways to measure metaphor novelty need to be considered in those cases.

Ethics Statement

In our work we only used corpora that were already freely available. The student research assistant conducted the annotation work within a fixed work contract and was paid according to public pay scales.

Acknowledgements

We thank the reviewer for the valuable comments and Simon Kreutz for his annotation efforts. We moreover thank the PC^2(Paderborn Center for Parallel Computing) for granting us compute time on the Noctua 1 cluster. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1475 – Project ID 441126958.

References

Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. MelBERT: Metaphor detection via contextualized late interaction using metaphorical identification theories. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1763–1773, Online. Association for Computational Linguistics.

Marco Del Tredici and Núria Bel. 2016. Assessing the potential of metaphoricity of verbs using corpus data. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4573–4577, Portorož, Slovenia. European Language Resources Association (ELRA).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Vesna G. Djokic, Ekaterina Shutova, and Verna Dankers. 2021. Episodic memory demands modulate novel metaphor use during event narration. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43.

Erik-Lân Do Dinh, Hannah Wieland, and Iryna Gurevych. 2018. Weeding out conventionalized metaphors: A corpus of novel metaphor annotations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1424, Brussels, Belgium. Association for Computational Linguistics.

Markus Egg and Valia Kordoni. 2022. Metaphor annotation for German. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2556–2562, Marseille, France. European Language Resources Association.

Hessel Haagsma and Johannes Bjerva. 2016. Detecting novel metaphor using selectional preference information. In *Proceedings of the Fourth Workshop on Metaphor in NLP*, pages 10–17, San Diego, California. Association for Computational Linguistics.

Walter Kintsch. 2000. Metaphor comprehension: A computational theory. *Psychonomic bulletin & review*, 7(2):257–266.

Tina Krennmayr. 2006. Using dictionaries in linguistic metaphor identification. *Selected Papers from the 2006 and 2007 Stockholm Metaphor Festivals*, page 95.

Chee Wee (Ben) Leong, Beata Beigman Klebanov, Chris Hamill, Egon Stemle, Rutuja Ubale, and Xianyang Chen. 2020. A report on the 2020 VUA and TOEFL metaphor detection shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 18–29, Online. Association for Computational Linguistics.

Chee Wee (Ben) Leong, Beata Beigman Klebanov, and Ekaterina Shutova. 2018. A report on the 2018 VUA metaphor detection shared task. In *Proceedings of the Workshop on Figurative Language Processing*, pages 56–66, New Orleans, Louisiana. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Longman. 2023. *Longman Dictionary of Contemporary English (Online Edition)*. Pearson Education Limited.

Michael Mohler, Mary Brunson, Bryan Rink, and Marc Tomlinson. 2016. Introducing the LCC metaphor datasets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4221–4227, Portorož, Slovenia. European Language Resources Association (ELRA).

Anmol Nayak, Hariprasad Timmapathini, Karthikeyan Ponnalagu, and Vijendran Gopalan Venkoparao. 2020. Domain adaptation challenges of BERT in tokenization and sub-word representations of out-of-vocabulary words. In *Proceedings of the First Workshop on Insights from Negative Results in NLP*, pages 1–5, Online. Association for Computational Linguistics.

Arthur Neidlein, Philip Wiesenbach, and Katja Markert. 2020. An analysis of language models for metaphor recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3722–3736, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Natalie Parde and Rodney Nielsen. 2018. A corpus of metaphor novelty scores for syntactically-related word pairs. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Paolo Pedinotti, Eliana Di Palma, Ludovica Cerini, and Alessandro Lenci. 2021. A howling success or a working sea? testing what BERT knows about metaphors. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 192–204, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pragglejaz Group. 2007. MIP: A Method for Identifying Metaphorically Used Words in Discourse. *Metaphor and Symbol*, 22(1):1–39.

W. Gudrun Reijnierse, Christian Burgers, Tina Krennmayr, and Gerard J. Steen. 2018. Dmip: A method for identifying potentially deliberate metaphor in language use. *Corpus Pragmatics*, 2(2):129–147.

W. Gudrun Reijnierse, Christian Burgers, Tina Krennmayr, and Gerard J. Steen. 2019. Metaphor in communication: the distribution of potentially deliberate metaphor across register and word class. *Corpora*, 14(3):301–326.

Michael Rundell, editor. 2002. *Macmillan English Dictionary for Advanced Learners*. Macmillan, Oxford.

Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna A. Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*, volume 14 of *Converging Evidence in Language and Communication Research*. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Chuandong Su, Fumiyo Fukumoto, Xiaoxi Huang, Jiyi Li, Rongbo Wang, and Zhiqun Chen. 2020. DeepMet: A reading comprehension paradigm for token-level metaphor detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 30–39, Online. Association for Computational Linguistics.

Alex Wang and Kyunghyun Cho. 2019. BERT has a mouth, and it must speak: BERT as a Markov random field language model. In *Proceedings of the*

Workshop on Methods for Optimizing and Evaluating Neural Language Generation, pages 30–36, Minneapolis, Minnesota. Association for Computational Linguistics.

Yorick Wilks. 1975. A preferential, pattern-seeking, semantics for natural language inference. *Artificial intelligence*, 6(1):53–74.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *arXiv preprint arXiv:1506.06724*.