Weeding out Conventionalized Metaphors: A Corpus of Novel Metaphor Annotations

Erik-Lân Do Dinh, Hannah Wieland, and Iryna Gurevych
Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt
http://www.ukp.tu-darmstadt.de

Abstract

We encounter metaphors every day, but only a few jump out on us and make us stumble. However, little effort has been devoted to investigating more novel metaphors in comparison to general metaphor detection efforts. We attribute this gap primarily to the lack of larger datasets that distinguish between conventionalized, i.e., very common, and novel metaphors. The goal of this paper is to alleviate this situation by introducing a crowdsourced novel metaphor annotation layer for an existing metaphor corpus. Further, we analyze our corpus and investigate correlations between novelty and features that are typically used in metaphor detection, such as concreteness ratings and more semantic features like the *Potential for Metaphoricity*. Finally, we present a baseline approach to assess novelty in metaphors based on our annotations.

1 Introduction

Metaphors have received considerable interest in NLP in recent years (see Shutova (2015)). Research questions range from direct detection of metaphors in text (*linguistic metaphors*) to finding mappings between conceptual source and target domains (*conceptual metaphors*).

However, an important aspect of metaphors—*novelty*—is often overlooked, or intentionally disregarded. Consider the metaphors (bold) in the following examples:

(1) We all live on **tight budgets**, but we still need to have some fun.

(2) They were beginning to attract a **penumbra** of gallery-goers, as though they were offering a guided tour.

The metaphor *tight budgets* in (1) is an often used collocation and therefore highly conventionalized. While the basic senses of *tight*—e.g., being physically close together or firmly attached—conflict with the more abstract *budget*, the metaphoric use as meaning *limited* can be readily understood. In contrast, the use of *penumbra* in (2) is more creative and novel. Its literal meaning is "an area covered by the outer part of a shadow."[1] Its metaphoric meaning is seldom encountered: Shadows follow objects that cast them, and especially penumbras can be perceived as having fuzzy outlines; attributes which are picked up by the metaphorical sense of a rather unspecified group of people following someone in differing vicinity.

Common linguistic metaphor definitions used in NLP (Steen et al., 2010; Tsvetkov et al., 2014) do not differentiate between *conventionalized* and *novel* metaphors. Some even allow for auxiliary verbs and prepositions to be annotated as metaphors when they are not used in their original sense (e.g., the non-spatially used *on* in "She wrote a study **on** metaphors"). While such cases can be filtered out rather easily from any given corpus—e.g., by using POS tag and lemma filters—many conventionalized metaphors persist. Existing work avoids this problem partially by only annotating certain grammatical constructions, such as adjective–noun or verb–object relations (Shutova et al., 2016; Rei et al., 2017). However, these too usually do not distinguish between conventionalized and novel metaphors.

Following Shutova (2015), we deem the distinction between conventionalized and novel metaphors important, because the meaning of conventionalized metaphors can usually be found in dictionaries or other resources like WordNet—novel metaphors on the other hand pose a more difficult challenge. But the lack of resources incorporating this distinction leads to few researchers

[1] https://www.macmillandictionary.com/dictionary/american/penumbra

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1412–1424
Brussels, Belgium, October 31 - November 4, 2018.

investigating novel metaphors, or the related measure of metaphoricity. They use small datasets that have been manually annotated by experts (Del Tredici and Bel, 2016), or focus crowdsourcing studies on a small number of instances (Dunn, 2014). It is only recently that any work has introduced larger-scale novel metaphor annotations (Parde and Nielsen, 2018). In contrast to our approach, they collect annotations on a relation level (see also Section 2).

Annotating metaphors is not an easy task, due to the inherent ambiguity and subjectivity. Therefore, we investigate approaches for annotating novelty in metaphors, before closing the resource gap for token-based annotations by creating a layer of novelty scores.

Our contributions are the following:

(1) We augment an existing metaphor corpus by assessing metaphor novelty on a token level using crowdsourcing, enabling larger research on novel metaphors,

(2) we analyze our corpus for correlation with features used for general metaphor detection or metaphoricity prediction, and

(3) we show that a baseline approach based on features usually used for (binary) metaphor detection can be useful for distinguishing novel and conventionalized metaphors.

In Section 2, we first discuss annotation guidelines that have been used for existing datasets, as well as prior work on novel metaphor detection and crowdsourcing in NLP. This discussion is followed by Section 3, where we detail the base corpus and the crowdsourced creation of our annotation layer. In Section 4, we describe our baseline for detecting novel metaphors and its results. We conclude in Section 5 with a summary of our contributions, and an outlook of what our newly introduced layer of annotations can enable.

2 Related Work

Annotating metaphors is a difficult task, because there is no single definition that can be adhered to. Instead, different researchers formulate their own versions, which can vary quite substantially.

The Pragglejaz Group (2007) created influential metaphor annotation guidelines, the *Metaphor Identification Procedure* (MIP). After reading a text, an annotator assesses each token as being used metaphorically or not; i.e., if the token has a more *basic* meaning that can be understood in comparison with the one it expresses in its current context. More basic is described as being:

- More concrete; what they evoke is easier to imagine, see, hear, feel, smell, and taste.
- Related to bodily action.
- More precise (as opposed to vague).
- Historically older.

They note that this basic meaning does not necessarily have to be the most frequent one. An extended version, MIPVU, was used to annotate parts of the British National Corpus (BNC Consortium, 2007, BNC), resulting in the VU Amsterdam Metaphor Corpus (Steen et al., 2010, VUAMC). Shutova and Teufel (2010) adapt the MIP as a prerequisite to annotating source and target domains of metaphor—the metaphoric mapping—in parts of the BNC.

Others use rather relaxed guidelines. Tsvetkov et al. (2014) rely on intuitive definitions by their annotators, not specifying metaphor more closely. They ask their annotators to mark words that "are used non-literally in the following sentences." Jang et al. (2015) provide a Wikipedia definition of metaphor to users in a crowdsourcing study. Subsequently, the users are tasked with annotating forum posts by deciding "whether the highlighted word is used metaphorically or literally."

A common trait of the listed works is that they do not distinguish between nuances of metaphoricity. Instead, they impose a binary scheme (metaphoric/literal) on the rather diffuse language phenomenon of metaphor. In contrast, Dunn (2014) introduces a scalar measurement of *metaphoricity* on a sentence level. He created a corpus of 60 genre-diverse sentences with varying levels of metaphoricity, which are rated in three crowdsourcing experiments: a binary selection, a ternary selection, and an ordering approach. The mean value for each sentence is then used as the metaphoricity value. Dunn (2014) uses this data to derive a computational measure of metaphoricity, which he then employs in an unsupervised system to label metaphoric sentences in the VUAMC. The evaluation, however, is only done on the binary labels provided therein.

In a similar direction, Del Tredici and Bel (2016) present their concept of *Potential of*

genre	tokens	metaphors	content tokens	content metaphors	novel metaphors
acprose	75,272	9,170	42,544	5,505	102
convrsn	57,249	3,841	22,019	1,774	25
fiction	54,115	5,349	26,935	3,174	94
news	51,672	7,580	29,346	4,727	132
total	238,308	25,940	120,844	15,180	353

Table 1: VUAMC corpus statistics. Content tokens include adjectives, adverbs, verbs (without *have*, *be*, *do*), and nouns. For the purpose of this table, metaphors with a novelty score higher than $T = 0.5$ are considered novel (the possible range is [–1,1]).

Metaphoricity (POM) of verbs. The POM describes the inherent potential of a verb to take on a metaphoric meaning, derived from its distributional behavior. They infer that low-POM verbs are only able to have low degrees of metaphoricity, thus can only evoke conventionalized metaphors. Therefore, they propose to exclude such low-POM verbs from novel metaphor detection systems.

Haagsma and Bjerva (2016) use violations of selectional preferences (Wilks, 1978) to find novel metaphors. Selectional preferences describe which semantic classes a verb prefers as its direct object. Since they are often mined from large corpora and based on frequency, they argue that this feature is more suited for novel metaphor detection than for general detection of (also conventionalized) metaphors. They evaluate their approach on the VUAMC; however, they acknowledge that their usage of this corpus is not optimal because it contains many conventionalized metaphors.

Parde and Nielsen (2018) create a corpus of novel metaphor annotations. For their crowdsourced annotation, they use a scale with four options. Unlike our approach, they annotate relations between words as novel or conventionalized. On the one hand, this is a sensible approach because generally the context of a word determines its metaphoricity (and indeed, its novelty in case of metaphoric use). On the other hand, such annotations lack the flexibility and ease of use of token-based annotations for which the context is not defined a priori.

We tackle the lack of data by annotating an existing corpus using crowdsourcing; i.e., by splitting up the task in many small chunks which different, non-expert annotators are instructed to complete. Crowdsourcing has been used for a variety of annotation tasks in NLP, often using different study designs. Snow et al. (2008) obtain good annotation results for five tasks with different setups: two tasks ask for numerical values (affect recognition, word similarity), two other tasks require a binary decision (textual entailment recognition, event ordering), and a final task provides three options for the annotators (word sense disambiguation). Sukhareva et al. (2016) utilize crowdsourcing to annotate semantic frames. They design their task as a decision tree, with annotators moving down the tree when annotating. Mohammad et al. (2016) use crowdsourcing for metaphor and emotion annotation in order to investigate their correlation. They employ an ordering approach to annotation, and only consider verbs that already contain a metaphoric sense in WordNet. Kiritchenko and Mohammad (2016) obtain annotations for sentiment associations via crowdsourcing. They use *Best–Worst Scaling* (Louviere and Woodworth, 1990), an annotation approach which creates scores from ranking annotations.

3 Corpus

To obtain a corpus of novel metaphor annotations, we employ an existing metaphor corpus. This can potentially reduce ambiguity for annotators and allows us to focus on the creation of novelty scores. We use the VU Amsterdam Metaphor Corpus (Steen et al., 2010) as the base corpus for our novelty annotations due to its comparably large size and genre diversity. It is comprised of over 200,000 tokens from four genres: *academic*, *fiction*, and *news* texts, and *conversation* transcripts (Table 1). Further, reusing existing annotations enables us to only query annotators for novelty of already annotated metaphors, instead of having them analyze every token.

Using crowdsourcing (Amazon Mechanical Turk), we first conduct a pilot study to choose among four different annotations methods. We

then employ the best method to collect annotations and create an additional *novelty score* between 1 (novel) and −1 (conventionalized) for each token labeled as metaphor in the VUAMC. Note that non-content words like prepositions and auxiliary verbs (*have*, *be*, *do*) are filtered out beforehand. Our annotations/scores can be integrated into the original VUAMC resource. We make the annotations and scripts to embed them into the original corpus publicly available.[2]

3.1 Pilot Study

Similar to Dunn (2014), we consider multiple annotation approaches. We compare them in a pilot study using 210 metaphor tokens with their sentence context, randomly chosen from the *fiction* subcorpus. For the sake of a meaningful evaluation, we ensured that 25% of these metaphorically used tokens were novel metaphors. Before choosing one approach for the entire corpus, we compare the following four annotation approaches:

- **binary annotation**: crowd workers decide if a given metaphoric token is used in a novel or conventionalized way;
- **scale annotation**: crowd workers decide on the novelty of a given metaphoric token on a four-point scale, from *very novel* to *very conventionalized*;
- **scale annotation (no metaphor)**: crowd workers decide on the "unusualness" of a given token in its context on a four-point scale, from *I've never heard it before* to *I'm using it everyday*; without giving information that the tokens represent metaphors;
- **best–worst scaling**: crowd workers pick the most novel and the most conventionalized metaphor from four samples.

Conceptually, the *binary annotation* should put the least cognitive load on the annotator, resulting in fast annotation times and efficient completion of the task. However, this method does not allow for nuances in annotation. Dunn (2014) counters this problem by assigning as the score the percentage of "metaphoric" labels for an instance. But this solution is only feasible for smaller datasets, as it requires many annotations per instance to yield nuanced scores.

[2] https://github.com/UKPLab/emnlp2018-novel-metaphors

	IAA	F_1	avg assignment completion time
binary	0.38	0.75	1:39 min
scale	0.32	0.75	1:04 min
scale w/o met.	0.16	0.67	2:10 min
BWS	—	0.84	1:58 min

Table 2: Comparison of the approaches investigated in the pilot study. Shown are inter-annotator agreement (Krippendorff's α, after mapping the scale annotations to binary labels), evaluation against our silver standard (F_1), and average completion time for an assignment (in case of binary, scale, and scale without metaphor this amounts to one decision, for BWS to two). Note that there is no IAA for the BWS approach because no two tuples are the same.

The next two approaches, *scale annotation* and *scale annotation (no metaphor)*, try to mitigate this problem by introducing four options to choose from. We choose a scale of four instead of three options to force the annotators to indicate a preference, rather than allowing a "neutral" answer. The difference between both scale approaches is in the guideline descriptions; for the second scale approach (no metaphor), we remove any mention of metaphor, and instead use paraphrases (e.g., instead of "novel metaphor" we use "I have not seen it before"). Our motivation behind this rephrasing is that we want to avoid confusion especially with regards to very conventionalized metaphors. By strictly asking for novelty of the expression/usage, we potentially simplify the task for the annotator.

The last approach uses *best–worst scaling* (BWS, see also Section 2). It has the advantage of not explicitly asking the annotator for a decision on a singular token/metaphor. Instead, they are asked to compare (four) different metaphors and select the most novel and the most conventionalized. A disadvantage is the higher workload for the annotator, since they have to read four sentences for one assignment.

In Table 2, we give a short overview of inter-annotator agreement (Krippendorff's α), the average completion time of an assignment, and a comparison with our semi-automatically created silver standard (F_1). The latter is built by using the majority vote of all four methods; ties are resolved manually by looking into the individual an-

notations. To obtain binary labels from the two scale approaches for majority voting, we map the two "more novel" options to novel, and the other two options to conventionalized. For the BWS approach, we first average the number of novel metaphors from the other three methods. From a sorted, decreasing list of BWS scores we then mark this many metaphors as novel. We compare against our own annotations as a sanity check.

Regarding completion time, we observe that the scale method is the fastest by a wide margin. This discrepancy is somewhat surprising, as the scale method introduces two more options to consider compared to the binary method. Apparently, these additional, intermediate options make it easier for annotators to come to a decision, especially for edge cases. On the other hand, scale without metaphors takes twice as long as the scale method. The latter only differs from the former in its inclusion of *metaphor* in the task description and the labels, which we thus interpret as creating a setting for the annotators where they may expect to have some kind of intuition about the task, and thus are more confident (and faster) in completing. Given the long completion time in conjunction with the lower F_1 score and the very low IAA, it seems clear that omitting the metaphor information and using a more colloquial task description make the scale annotation (no metaphor) by far the most difficult for workers to complete, resulting in the least usable annotations. While best–worst scaling is time consuming as well, it yields the best results with regards to the silver standard ($F_1 = 0.84$) by a large margin. We thus choose BWS for annotating the full corpus. Further details on the pilot study can be found in Wieland (2018).

3.2 Corpus Creation

We design our guidelines (see Appendix A) to be simple, but include redundancy in the description to address frequent misunderstandings and ambiguities. We also explicitly mention idioms and unrecognizable metaphors as cases of conventionalized metaphors, because these were sources of confusion in our pilot study. Unrecognizable here means that a word is used in an established, commonly understood—but not the most concrete—sense; e.g., *hard* in "She fought **hard**." Additionally, the annotators are provided with example metaphors of differing novelty.

After filtering out prepositions and auxiliary verbs (*have*, *be*, *do*) using the POS tags supplied by the VUAMC, we collect annotations covering 15,180 metaphors in total (Table 1). We only include workers located in the US. For creation of the best–worst scaling tuples, and for aggregation of the annotations, we use the scripts provided by Kiritchenko and Mohammad (2016).[3] We use a best-worst scaling factor of 1.5 and four items per tuple. Thus, each metaphor appears in six different best-worst scaling comparisons. This results in 22,770 best–worst scaling items to be annotated.

3.3 Analysis

Overall statistics about our created annotations are shown in Table 1 (along with the already existing annotations). For the sake of this overview, we introduce a threshold $T = 0.5$, and treat metaphors with a BWS score equal to or above this threshold as novel, metaphors with a BWS score below the threshold as conventionalized. For example, in this way the metaphor "[...] the artistic temperament which kept her **tight-coiled** as a spring [...]" (0.514) is treated as novel, while "To **quench** [thirst] is more than to refresh [...]" (0.424) is treated as conventionalized. However, since we provide the scores, this threshold can be adjusted to suit a given application. Note that, while novel metaphors are arguably much more scarce than conventionalized ones, BWS creates scores which are approximately normally distributed, supporting our threshold choice.

Before we conduct a more in-depth analysis of the annotated metaphors, we show some examples. In Table 3, we list four novel and four conventionalized metaphors (as annotated). A good example for a novel metaphor is the description of "words [...] as a **coat-hanger**" in Table 3 (3). This usage cannot be found in dictionaries, and clearly constitutes creative language use. In contrast, the meaning *to experience something bad* of *to go through [a situation]* (ibid., (7)), or the sense *to do/conduct* of *to get [something] done* (ibid., (5)), are strongly conventionalized, as indicated by their inclusion in dictionaries.[4]

We also group the tokens by lemma to examine if certain words are more likely to be used in a novel metaphoric way. Inspecting the mean

[3] http://saifmohammad.com/WebPages/BestWorst.html

[4] e.g., https://www.macmillandictionary.com/dictionary/american/go-through#go-through__7, https://www.macmillandictionary.com/dictionary/american/get#get__60

no	score	metaphor in context
(1)	0.765	Ron Todd [...] warned that party leaders could not expect everybody to '**goose-step**' in the same direction once the policy had been carried.
(2)	0.750	Westerns have a **gladiatorial**, timeless quality.
(3)	0.735	Allan Ahlberg says: 'In the past, a lot of children's books seemed to be the work of talented illustrators whose pictures looked brilliant framed in a gallery, but when you tried to read the book, there was nothing there, because the words started as a **coat-hanger** to hang pictures on.'
(4)	0.727	thus one can and must say ... that each fight is the singularisation of all the circumstances of the social whole in movement and that by this singularisation, it **incarnates** the enveloping totalization which the historical process is.
(5)	−0.765	If the complaint is proved, a nuisance order is made requiring the defendant to **get** the necessary work done.
(6)	−0.765	Apart from some dark patches on the wall that he hadn't noticed **before**, there was nothing to see.
(7)	−0.774	In relation to the sentence stem 'A girl and her mother ...', girls often produce responses like 'often **go** through a bad patch for a year but once they learn to understand each other, become the best of friends' or 'can help each other with their problems'.
(8)	−0.871	The analyst is then forced on the defensive, explaining why **new** features can not be included because they are technically difficult or prohibitively expensive.

Table 3: Example uses of very novel and very conventionalized annotated metaphorical tokens (bold) in sentence context, according to our aggregated annotations. Scores near 1 denote strong novelty, scores near −1 indicate very conventionalized metaphors.

scores, we see the large conventionalization of words such as *get* (−0.37), *see* (−0.35), or *new* (−0.35). Examples for words used in a mostly novel way are *envelop* (0.53), *incarnate* (0.50), and *thrust* (0.46). In the following sections, we examine the correlations with frequency and POM, which give further weight to the examples.

Further, we investigate how novelty is distributed over the different subcorpora (also see Table 1). Somewhat expectedly, the metaphors used in the academic subcorpus are mostly conventionalized (1.9% novel metaphors). A larger percentage of novel metaphors can be found in the news (2.8%) and fiction (3.0%) subcorpora. The conversation subcorpus shows the least amount of novel metaphors (1.3%), which we interpret as natural—more novel metaphors require some amount of creativity to create, which is arguably easier in writing. We also show the distribution of novel metaphors across POS tags (Table 4). It is striking that the verbs are least represented among the novel metaphors, especially compared to their overall metaphoric occurrence. In contrast, adjectives/adverbs and nouns are more likely to be used in a novel way.

POS	tokens	metaphors	novel metaphors
nouns	47,171	5,513	145
verbs	27,831	6,513	88
adj/adv	45,842	3,154	120

Table 4: Metaphor annotations grouped by POS tags. For this overview, we treat metaphors with a score above $T = 0.5$ as novel.

Correlation with Token Frequency We compare how the token frequency is related to novelty. Intuitively, words that are seldom used should show a higher (average) novelty. In turn, we expect often used words to exhibit a wider range of (already conventionalized) senses, thus having lower (average) novelty. We obtain token frequencies from a Wikipedia dump, and correlate them with the mean novelty scores from our annotations (disregarding part-of-speech tags). The relation in terms of Spearman's rank correlation is $\rho = -0.60$, which indicates a moderate anti-correlation. This can also be observed in Figure 1: while high frequency seems to hint at conventionalized metaphors, low frequency is not necessarily

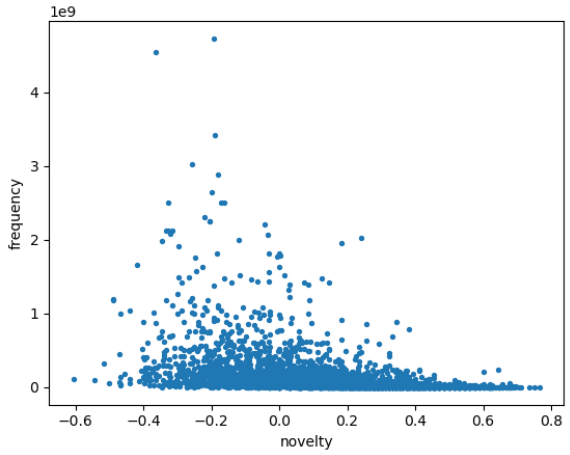


Figure 1: **Frequency.** Relation between average novelty score of metaphoric tokens and their frequency in Wikipedia (correlation of $\rho = -0.60$). Since we use the automatically created POS tags from the VUAMC to filter out non-content tokens, this can include erroneously tagged tokens. For an improved overview, we manually filtered out the five most frequent, incorrectly tagged tokens from this plot (*to*, *as*, *on*, *that*, and *this*).

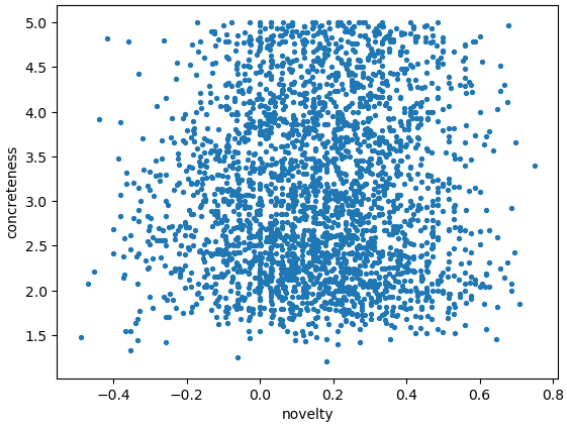


Figure 2: **Concreteness.** Relation between novelty score of metaphoric tokens and their concreteness, showing no discernible correlation ($\rho = 0.03$). A similar picture emerges if we only consider the manually annotated tokens included in the original concreteness list by Brysbaert et al. (2014).

an indication of novelty of metaphoric use.

Two prominent exceptions are the tokens *national* and *united* (Figure 1, upper right). While they show comparatively high novelty (both labeled as metaphorical only once in the VUAMC), they appear surprisingly often in Wikipedia. Another artifact of using Wikipedia as the background corpus can be seen on the left: *try* is only annotated as metaphoric in infinitive-compounds that are decidedly conventionalized (e.g., "trying to look"), yet it appears comparatively seldom in Wikipedia. However, we chose to use Wikipedia instead of the BNC in order to have an out-of-domain comparison with a more contemporary, larger background corpus.

Correlation with Concreteness The use of concreteness as a feature in automatic metaphor detection grounds in the Conceptual Metaphor Theory (Lakoff and Johnson, 1980). In short, metaphors are modeled as cognitive mappings between an often concrete source domain, and a usually more abstract target domain. For example, in "He **shot down** my arguments," the more concrete domain of ARMED CONFLICT (*shot*) is mapped to the rather abstract domain DISCUSSION (*argument*).

To analyze the relation between novelty and concreteness, we first extend the concreteness list by Brysbaert et al. (2014) using a technique similar to Mohler et al. (2014). For a given token t, we extract 20 approximate nearest neighbors *nn(t)* from Google News Embeddings (Mikolov et al., 2013) using Annoy.[5] The concreteness value for t is then computed by averaging its neighbors' concreteness values from the concreteness list.

Subsequently, we calculate the correlation between the average novelty and the concreteness of the lemmas. Both Pearson correlation ($r = 0.04$) and Spearman's rank correlation ($\rho = 0.03$) are close to zero and indicate no correlation (Figure 2). Thus, while concreteness has been shown to work well as a feature to distinguish between literal and non-literal language in general (Beigman Klebanov et al., 2014; Tsvetkov et al., 2014), it does not seem useful for discerning between novel and conventionalized metaphoric usage in particular. For example, the rather abstract *now* (1.48) and the very concrete *people* (4.82) are assigned similarly low novelty scores. On the other hand, *justice* has a quite high novelty score (0.65), while also being as abstract as *now*. One reason for this non-correlation between concreteness and novelty might be that the automatic induction of concreteness ratings introduces too much noise. However, an experiment where we only use tokens occurring in the manually composed list (Brysbaert et al., 2014) shows similarly low correlation. This could be influenced by artifacts in the concrete-

[5] https://github.com/spotify/annoy

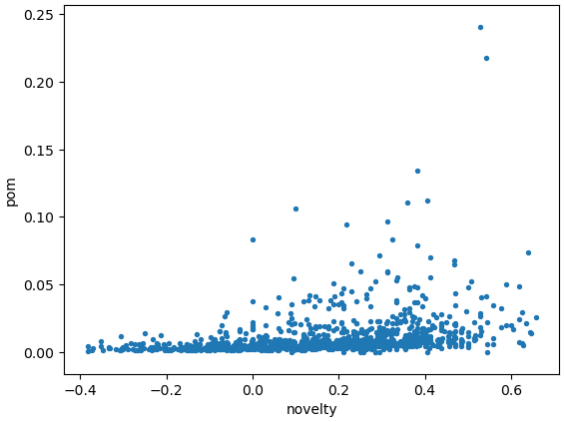


Figure 3: **POM.** Relation between average novelty score of verb lemmas and POM (correlation of $\rho = 0.52$).

ness list: as laid out by Beigman Klebanov et al. (2015), it exhibits some problems. For example, it shows high variance in the annotated concreteness scores for various non-concrete adjectives. But this does not explain the extent of the non-correlation. We thus believe that our results indeed indicate no relation between concreteness and metaphor novelty. And indeed, if a difference of concreteness between components of an expression hints at a metaphor, as is proposed by the conceptual metaphor theory, then it is plausible that it does not hint at novelty of the metaphoric expression at the same time. Consequently, we investigate a feature with more semantic capacity in the next subsection.

Correlation with POM The *Potential for Metaphoricity* (POM) of verbs was introduced by Del Tredici and Bel (2016) (Section 2). It denotes the a priori chance of a verb to occur in highly metaphoric contexts; in essence, it measures the contextual flexibility of a verb. Generally, very novel metaphors also display a high metaphoricity. Therefore, we expect that low-POM verbs (i.e., verbs that occur similarly often in many different contexts) exhibit a low novelty score on average and low variance, while high-POM verbs should show a higher average novelty score. The POM can be regarded as a variant of selectional preference strength, which measures how strongly a verb constrains its direct object in terms of semantic classes. As such, we forgo an analysis of selectional preference violations in favor of examining the POM. Hovy et al. (2013) generalize the notion of selectional preferences to other forms of grammatical relations. However, instead of generating scores, they use dependency trees in an SVM with tree kernels. The POM could be similarly generalized to all POS tags, e.g., by including head and dependent tokens as context.

We create the POM for all annotated verbs using the same procedure as Del Tredici and Bel (2016). First, we extract the context (i.e., subject and object) for each occurrence of a verb from a large, parsed corpus (Wikipedia). To compute context vectors, the word embeddings (Levy and Goldberg, 2014) of subject and object are averaged (if only one of the two is available, the embedding for this token serves as the context). For each verb, the context vectors are then clustered using Birch clustering (Zhang et al., 1996). Finally, the standard deviation between the sizes of the context clusters denotes the POM of the verb.

As with our previous experiments, we compute Spearman's rank correlation between the mean novelty scores of the verb lemmas and the corresponding POMs. We arrive at Spearman's $\rho = 0.52$, which indicates moderate correlation (Figure 3). Verbs like *pique* (POM: 0.218) and *slit* (0.134) are more often used in a novel metaphoric way, while low-POM verbs *show* and *see* (both: 0.01) are only used as conventional metaphors. Even though the correlation is only moderate, it supports the WordNet-based evaluation by Del Tredici and Bel (2016), who found that high-POM verbs generally induced novel metaphoric sentences. Note that their POM values are higher because they optimize the clustering parameters, which we leave at the default setting.

4 Baseline

While the main focus of this work is the analysis of our new annotation layer, we also create a simple baseline regression system for predicting the novelty scores. We run the system using two configurations: first with only word embeddings as input, then augmented with frequency and POM scores.

4.1 System

We implement a single-layer BiLSTM for predicting the novelty score. As input, we use a padded 11-token window (five before, five after the token) of dependency-based word embeddings by Levy and Goldberg (2014) that we also employed for the POM computation. The BiLSTM layer has 50 dimensions and ReLu activation. Training is done

in maximally 20 epochs, but can halt earlier due to early stopping on the development set. The data is split into training (50%), development (25%), and test set (25%). We only conduct experiments on verbs, so that we can compare the performance when including additional features.

Following our analysis of typically used features for metaphor detection in relation to novelty, we incorporate the relative frequency of a token into the model, to investigate if the substantial correlation observed in Section 3.3 has an impact on our regression experiments. Further, we include the POM, as it also showed a moderate correlation with novelty. Both additional features are concatenated with the respective word embeddings and the resulting vectors are fed to the BiLSTM.

4.2 Results

We evaluate our results using mean absolute error (MAE), and average it over 10 runs with different random seeds. Recall that the possible values lie between −1 and 1, leading to a possible MAE between 0 (best) and 2 (worst). The baseline results over different runs are stable, we show the regression plot for one configuration in Figure 4.

The mean absolute error for the configuration using only the embeddings is MAE = 0.166. In contrast, we obtain MAE = 0.163 for the same configuration when we add the frequency and the POM features. Thus, while small, the latter model shows improvements over the word embedding baseline. As can be seen in the plot, there is still much room for improvement (e.g., we did not conduct extensive hyper-parameter optimization). The network makes errors both in underestimating ("they can be seen as **riddled** with holes") and overestimating novelty ("veins branching off it to **form** a network"). These errors are in many cases independent from POM and frequency features. Thus, while a better optimization (e.g., in the clustering step when creating the POM) might reduce estimation errors, we also need to consider further features for metaphor novelty estimation.

5 Conclusion

We presented a new layer of novelty scores for the VU Amsterdam Metaphor Corpus, created using crowdsourcing. To this end, we conducted a pilot study to choose an appropriate method for metaphor novelty annotation and found that best–worst scaling outperformed binary and scale

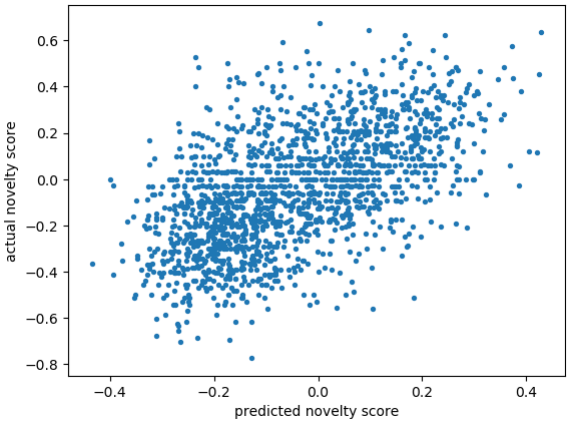


Figure 4: Predicted and actual novelty score of metaphoric tokens in the VUAMC for a baseline configuration (word embeddings only, without adding frequency and POM information).

methods. Our corpus analysis of typically used features for metaphor detection showed no correlation of novelty with concreteness. However, we found substantial correlation with frequency of tokens in a background corpus and with potential for metaphoricity, a context-based a priori metaphoricity measure. Further, we created a baseline to distinguish novel from conventionalized metaphors. For our approach, the latter two features could improve results only slightly, indicating a need for more sophisticated features.

Previous work in automatic metaphor processing has largely focused on general detection of linguistic and conceptual metaphors, mostly disregarding the subject of novelty. Our corpus enables new evaluation and training possibilities for detecting the latter. In future work, we want to develop more sophisticated methods to detect and distinguish novel metaphors. For example, we want to extend the notion of POM to nouns and adjectives, and investigate other a priori measures for metaphor novelty. Further, we want to jointly detect metaphors and score their novelty. Another interesting direction is to investigate the correlation between perceived novelty and the existence of dictionary definitions for metaphoric senses of a token or expression.

Acknowledgements

This work has been supported by the German Federal Ministry of Education and Research (BMBF) under the promotional reference 01UG1816B (CEDIFOR).

References

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

Beata Beigman Klebanov, Chee Wee Leong, Michael Heilman, and Michael Flor. 2014. Different Texts, Same Metaphors: Unigrams and Beyond. In *Proceedings of the Second Workshop on Metaphor in NLP*, pages 11–17, Baltimore, MD, USA. Association for Computational Linguistics.

BNC Consortium. 2007. The British National Corpus, version 3 (BNC XML Edition).

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness Ratings for 40 Thousand Generally Known English Word Lemmas. *Behavior Research Methods*, 46(3):904–11.

Marco Del Tredici and Núria Bel. 2016. Assessing the Potential of Metaphoricity of Verbs Using Corpus Data. In *Proceedings of LREC 2016*, pages 4573–4577, Portorož, Slovenia. European Language Resources Association.

Jonathan Dunn. 2014. Measuring Metaphoricity. In *Proceedings of ACL 2014*, pages 745–751, Baltimore, MD, USA. Association for Computational Linguistics.

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, CA, USA. Association for Computational Linguistics.

Dirk Hovy, Shashank Srivastava, Sujay Kumar Jauhar, Mrinmaya Sachan, Kartik Goyal, Huiying Li, Whitney Sanders, and Eduard Hovy. 2013. Identifying Metaphorical Word Use with Tree Kernels. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 52–57, Atlanta, GA, USA. Association for Computational Linguistics.

Hyeju Jang, Seunghwan Moon, Yohan Jo, and Carolyn Penstein Rosé. 2015. Metaphor Detection in Discourse. In *Proceedings of SIGDIAL 2015*, pages 384–392, Prague, Czech Republic. Association for Computational Linguistics.

Svetlana Kiritchenko and Saif M. Mohammad. 2016. Capturing Reliable Fine-Grained Sentiment Associations by Crowdsourcing and Best–Worst Scaling. In *Proceeding of NAACL-HLT 2016*, pages 811–817, San Diego, CA, USA. Association for Computational Linguistics.

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

Omer Levy and Yoav Goldberg. 2014. Dependency-Based Word Embeddings. In *Proceedings of ACL 2014*, pages 302–308, Baltimore, MD, USA.

Jordan J. Louviere and George G. Woodworth. 1990. Best–worst Analysis. Working paper. Department of Marketing and Economic Analysis, University of Alberta.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of NIPS 2013*, pages 3111–3119, Stateline, NV, USA. Curran Associates Inc.

Saif M. Mohammad, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a Medium for Emotion: An Empirical Study. In *Proceedings of *SEM 2016*, pages 23–33, Berlin, Germany. Association for Computational Linguistics.

Michael Mohler, Marc Tomlinson, David Bracewell, and Bryan Rink. 2014. Semi-Supervised Methods for Expanding Psycholinguistics Norms by Integrating Distributional Similarity with the Structure of WordNet. In *Proceedings of LREC 2014*, Reykjavik, Iceland. European Language Resources Association.

Nathalie Parde and Rodney D. Nielsen. 2018. A Corpus of Metaphor Novelty Scores for Syntactically-Related Word Pairs. In *Proceedings of LREC 2018*, pages 1535–1540, Miyazaki, Japan. European Language Resources Association.

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

Marek Rei, Luana Bulat, Douwe Kiela, and Ekaterina Shutova. 2017. Grasping the Finer Point: A Supervised Similarity Network for Metaphor Detection. In *Proceedings of EMNLP 2017*, pages 1538–1547, Copenhagen, Denmark. Association for Computational Linguistics.

Ekaterina Shutova. 2015. Design and Evaluation of Metaphor Processing Systems. *Computational Linguistics*, 41(4):579–623.

Ekaterina Shutova, Douwe Kiela, and Jean Maillard. 2016. Black Holes and White Rabbits: Metaphor Identification with Visual Features. In *Proceedings of NAACL-HLT 2016*, pages 160–170, San Diego, CA, USA. Association for Computational Linguistics.

Ekaterina Shutova and Simone Teufel. 2010. Metaphor Corpus Annotated for Source–Target Domain Mappings. In *Proceedings of LREC 2010*, pages 3255–3261, Valetta, Malta. European Language Resources Association.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. 2008. Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In *Proceedings of EMNLP*

2008, pages 254–263, Honolulu, HI, USA. Association for Computational Linguistics.

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

Maria Sukhareva, Judith Eckle-Kohler, Ivan Habernal, and Iryna Gurevych. 2016. Crowdsourcing a Large Dataset of Domain-Specific Context-Sensitive Semantic Verb Relations. In *Proceedings of LREC 2016*, pages 2131–2137, Portorož, Slovenia. European Language Resources Association.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor Detection with Cross-Lingual Model Transfer. In *Proceedings of ACL 2014*, pages 248–258, Baltimore, MD, USA. Association for Computational Linguistics.

Hannah Wieland. 2018. Crowd-Sourcing Novel Metaphor Annotations. Bachelor Thesis, TU Darmstadt.

Yorick Wilks. 1978. Making Preferences More Active. *Artificial Intelligence*, 11(3):197–223.

Tian Zhang, Raghu Ramakrishnan, and Miron Livny. 1996. BIRCH: An Efficient Data Clustering Method for Very Large Databases. In *Proceedings of SIGMOD 1996*, pages 103–114, New York, NY, USA. Association for Computing Machinery.

A Guidelines

This task is about metaphors. A metaphor is a figure of speech that describes one thing by mentioning another thing. You can divide metaphors into two types: conventional and novel ones. In the following texts some words are marked as metaphorical.
You will be given four metaphors and your task is to decide which metaphor is the most conventional and which metaphor is the most novel.
Conventional metaphors are metaphors which are often used in everyday language.
In contrast novel metaphors which are usually not used in everyday language.
Please check the instructions and examples below before starting with the HIT!

Instructions and hints:

- Please read the whole context around the metaphors before deciding which metaphor is most novel or most conventional.
- You have to answer two questions per task:

  The first question is which metaphor is most *conventional*. That means you have to select the metaphor which is the most common in everyday language.

  The second question is which metaphor is the most *novel*. That means you have to select the metaphor that is the most uncommon in everyday language.
- Expressions that are so common that you cannot recognize them as a metaphor are candidates for the most conventional metaphor.
- Idioms are rather conventional metaphors.
- Each task is about four metaphors which are marked in red.
- The sentence which contain the metaphor are highlighted in bold letters.
- You have to fully complete all three tasks to get paid.

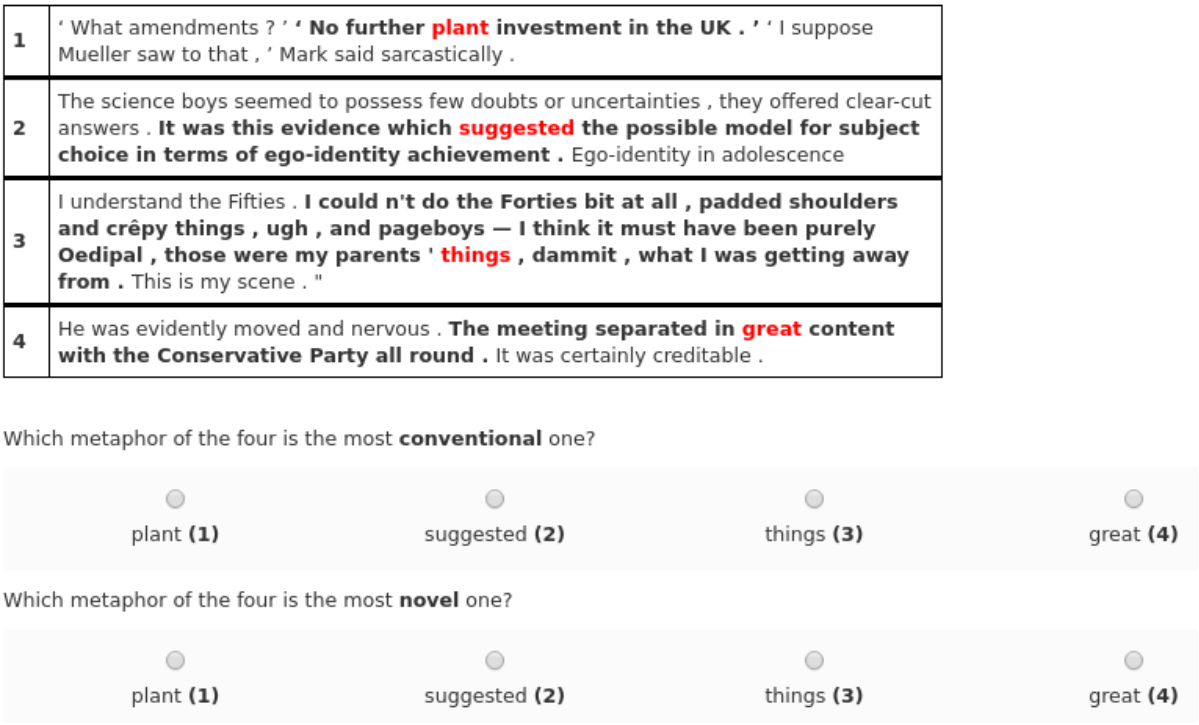
Examples:

1: "She **gave** him that idea."
2: "I see the **point**."
3: "Some books have to be **tasted**."
4: "Time **flies.**"

- Answer: most conventional metaphor: **gave** (1); most novel metaphor: **tasted** (4).
- Explanation: "gave" is in the expression "to give someone an idea" so common that you can barely see the metaphor, that means it is extremely conventional. "to *taste* books" in contrast is a very uncommon use of the word "taste" and thus a novel metaphor. The other two metaphors are idioms which are conventional but not as conventional as "to give someone an idea.".

Figure 5: Annotation guidelines for best-worst scaling HITs.

B Example of a Best-Worst Scaling HIT



1	‘ What amendments ? ’ **‘ No further plant investment in the UK . ’** ’ I suppose Mueller saw to that , ’ Mark said sarcastically .
2	The science boys seemed to possess few doubts or uncertainties , they offered clear-cut answers . **It was this evidence which suggested the possible model for subject choice in terms of ego-identity achievement .** Ego-identity in adolescence
3	I understand the Fifties . **I could n't do the Forties bit at all , padded shoulders and crêpy things , ugh , and pageboys — I think it must have been purely Oedipal , those were my parents ' things , dammit , what I was getting away from .** This is my scene . "
4	He was evidently moved and nervous . **The meeting separated in great content with the Conservative Party all round .** It was certainly creditable .

Figure 6: Example of a best-worst scaling HIT, where annotators were asked to choose the most novel and the most conventionalized metaphor among 4 instances.