Exploring the Terrain of Metaphor Novelty: A Regression-based Approach for Automatically Scoring Metaphors

Natalie Parde and **Rodney D. Nielsen**
Human Intelligence and Language Technologies (HiLT) Laboratory
Department of Computer Science and Engineering
University of North Texas

Abstract

Automatically scoring metaphor novelty has been largely unexplored, but could be of benefit to a wide variety of NLP applications. We introduce a large, publicly available metaphor novelty dataset to stimulate research in this area, and propose a regression-based approach to automatically score the novelty of potential metaphors that are expressed as word pairs. We additionally investigate which types of features are most useful for this task, and show that our approach outperforms baseline metaphor novelty scoring and standard metaphor detection approaches on this task.

Introduction

The vast majority of computational work on figurative language to date has framed metaphor detection as a binary (metaphor/non-metaphor) classification task. However, in reality language lies along a graded continuum, with figurative expressions ranging from highly conventional (or even fossilized) to highly novel[1] or creative (Gibbs 1984). While most conventional metaphors could in theory be handled by word sense disambiguation, novel metaphors are potentially more problematic for NLP applications (Shutova 2015; Haagsma and Bjerva 2016). Automatically scoring the novelty of potential metaphors could allow for conventional and novel metaphors to be processed differently, enabling better-performing language understanding systems. Novel metaphors are also more difficult for humans to process (Lai, Curran, and Menn 2009), and therefore automatic metaphor novelty scoring could additionally be useful for applications such as automatic essay scoring, automatic assessment of cognitive health, and automatic generation of cognitively stimulating discussion topics. Cognitive health applications offer a particularly intriguing downstream use case for metaphor novelty scoring; probable Alzheimer's disease patients, for example, often struggle with comprehending novel metaphors but not conventional metaphors (Amanzio et al. 2008).

[1]We define novelty herein as the frequency with which one may be expected to encounter a metaphor. Thus, if a metaphor is quite common (e.g., "Two *hours passed*.") it has low novelty, whereas if it is surprising or unique (e.g., "Her *laughter waltzed* through the courtyard.") it has high novelty.

The dearth of research on identifying metaphor novelty has at least partially been due to a lack of publicly available data with which to train and evaluate such systems (Haagsma and Bjerva 2016). In this work, our contributions are as follows: we (1) contribute a large (18,000+ instances) dataset annotated for metaphor novelty and make it publicly available to stimulate research in this area.[2] We also (2) conduct an analysis of a wide range of features, both novel and drawn from existing work on standard metaphor detection, to determine what linguistic and conceptual characteristics are most indicative of metaphor novelty. Finally, we (3) contribute a regression-based metaphor novelty scoring approach[3] that outperforms baseline metaphor novelty scoring and standard metaphor detection approaches and establishes a benchmark on this dataset for future work.

Related Work

Detecting Metaphor Novelty

Work that has specifically set out to capture metaphor novelty includes a machine learning approach that made use of selectional preference features (Haagsma and Bjerva 2016) and a rule-based approach based on bigram counts and WordNet hyponymy relations (Krishnakumaran and Zhu 2007). The former group was ultimately unable to pursue their original goals due to a lack of adequate training data and instead applied their approach to general, binary metaphor detection.

The latter distinguished between sentences containing novel and conventional metaphors (a binary classification task). To evaluate their work, they expanded the metaphors in the Berkeley Master Metaphor List (Lakoff 1994) into approximately 1728 simple sentences, and then labeled those sentences as containing either "dead" or "live" metaphors. This differs from our approach in that our system learns from and assigns instances with a continuous value ranging between 0 ("not a metaphor") and 3 ("highly novel metaphor"). Krishnakumaran and Zhu also base their rules strictly on having one of three specific constructions available (subject IS-A object, verb-noun, or adjective-noun), and require that

[2]http://hilt.cse.unt.edu/resources.html

[3]https://github.com/natalieparde/metaphor-novelty-scoring

the words in a pair exist in WordNet (words not in a WordNet hyponymy relation would always result in the sentence being labeled as "live"). These restrictions limit the applicability of the approach to real-world text, in which metaphors may not always contain nouns and some words (literal or metaphoric) may not exist in WordNet.

Detecting Metaphors in General

General metaphor detection (without distinguishing between grades of novelty) has been carried out in numerous ways at the sentence (Dunn 2013; Dunn et al. 2014; Krishnakumaran and Zhu 2007; Mohler et al. 2013), word pair (Shutova, Sun, and Korhonen 2010; Gandy et al. 2013; Tekiroglu, Özbal, and Strapparava 2015; Turney et al. 2011; Gutierrez et al. 2016; Shutova, Kiela, and Maillard 2016), and individual word (Beigman Klebanov, Leong, and Flor 2015; Gargett and Barnden 2015; Beigman Klebanov et al. 2014; 2016; Özbal et al. 2016; Mohammad, Shutova, and Turney 2016; Jang, Wen, and Rosé 2015; Schulder and Hovy 2014; Turney et al. 2011; Jang et al. 2015; 2016) level. We choose to model the metaphor novelty of word pairs because syntactically related pairs of words are the smallest unit for which novelty and meaning may be contextually inferred. Past work on identifying metaphoric pairs has often been constrained to specific types of pairs (e.g., verb-noun or adjective-noun). We loosen this restriction and instead consider the novelty of a wide variety of syntactically related pairs that contain either two content words (noun, verb, adjective, or adverb), or a content word and a personal pronoun. Later, we discuss specific features used in prior metaphor detection approaches from which the feature sets in this work were partially derived.

Existing Datasets

A number of datasets exist for binary metaphor detection (c.f. Mohler et al. (2016) for a good overview of these). The most popular existing metaphor dataset is the VU Amsterdam Metaphor Corpus (VUAMC) (Steen 2010). The VUAMC is a subset of the BNC Baby corpus, which labels individual words as metaphors in four genres of text: news, conversation, fiction, and academic. The corpus includes all metaphors, regardless of word type or novelty; however, no distinction is made between different grades of novelty in the annotations.

Other existing datasets include Levin et al.'s (2014) collection of conventionalized conceptual metaphors, Birke and Sarkar's (2006) dataset of metaphoric and literal uses of 51 verbs, Tsvetkov et al.'s (2014) dataset of metaphoric and literal adjective-noun pairs and subject-verb-object triples, Mohammad, Shutova, and Turney's (2016) dataset of metaphoric and literal instances of 440 verbs, and the Berkeley Master Metaphor List (Lakoff 1994) which contains 208 conceptual metaphors.

Mohler et al. (2016) recently described a dataset for which the *metaphoricity* of pairs of words is given one of four discrete labels (0="No Metaphoricity," 1="Possible/Weak Metaphor," 2="Likely/Conventional Metaphor," 3="Clear Metaphor"). Of existing corpora, this is most closely suited to our task. However, the pairs have relatively low diversity (there are only 1512 unique target words across all pairs in the free version of the corpus, even when considering "Taxes," "taxes," "tax," "Tax," "taxation," "income taxes," "income tax," etc. all as unique words). Finally, we observed that "clear metaphors" are in some cases quite distinct from novel metaphors. Consider the examples from Mohler et al.'s corpus in Table 1.

Sentence	**Score**
Thank Lyndon Johnson, his Great Society, and the *War* on *Poverty*.	3
A measure of the protection provided to an industry by the entire *structure* of *tariffs*, taking into account the effects of tariffs on inputs as well as on outputs.	3
Regarding the word "fair" - I don't think anybody can comment on "fair" without talking to the marginal value of a dollar, marginal effort to make a dollar, the shift in wealth over the past 30 years, and the overall idea that *money* is the *fuel* of the money machine.	1
The catastrophe and its aftermath displayed in sharp relief the glories and flaws of a city fast becoming the symbol of a nation *drunk* on *democracy*.	1

Table 1: LCC Dataset Sentences and Scores

The first two examples, although certainly metaphoric, are not particularly novel (most readers are likely to have encountered the expression "War on Poverty" quite often). The latter two examples were rated as weakly metaphoric, yet they are far more novel. Thus, it appears that metaphoricity and metaphor novelty, although in some cases correlated, are separate characteristics with nuanced differences.

Data

We built our dataset on top of the VUAMC. To do so, we extracted syntactically-related[4] pairs of either two *content words* (nouns, verbs, adjectives, or adverbs, excluding stopwords, proper nouns, and some auxiliary verbs) or one content word and a personal pronoun, for which at least one of the words was labeled as a metaphor in the VUAMC. We consider these pairs to be potentially metaphoric; although they all contain a word annotated as a metaphor, they do not all necessarily convey the word's metaphoric usage. For example, in the sentence, "Her *laughter* **waltzed** through the *courtyard*," the pair {*laughter*, *waltzed*} is a novel metaphor, whereas {*courtyard*, *waltzed*} is non-metaphoric. Our data is publicly available under the Creative Commons Attribution ShareAlike 3.0 Unported License.

Annotations

We crowdsourced annotations using Amazon Mechanical Turk[5] (AMT) for 18,452 word pairs from the VUAMC, and randomly divided these pairs into training (approximately 80%) and test (all word pairs from the same source document were assigned to the same set). AMT workers were

[4]Stanford CoreNLP (Manning et al. 2014) was used to obtain dependency parses and part-of-speech (POS) tags.

[5]www.mturk.com

asked to label each word pair with a single score based on whether it formed a metaphor in the context of the surrounding sentence (0=not a metaphor) and if so, that metaphor's novelty from low (1) to high (3). Word pairs were grouped into Human Intelligence Tasks (HITs) containing all potentially metaphoric word pairs extracted from 10 sentences. Five worker assignments were requested per HIT. Overall, 479 workers participated in annotating 1004 unique HITs. For each instance in the test set, two annotations were also collected from trained (non-AMT) annotators; these annotations were used to build the *gold standard* to which predictions were compared.

Data Filtering: We automatically filtered annotations as they were received, rejecting HITs that were completed abnormally quickly or by workers whose performance was deemed substandard (i.e., spammers or deliberately malicious workers), using a filtering algorithm based on workers' correlations with one another (Parde and Nielsen 2017). Workers with substandard or questionable performance were disqualified from accepting future HITs.

Adjudication Procedure

The crowdsourced annotations from the training data were automatically aggregated to a continuous label using the regression-based approach developed by Parde and Nielsen (2017). Briefly, this approach trains a random subspace regression model on data that has been labeled by both crowd workers and experts, using features based on annotation distribution and presumed annotator quality, to predict optimal aggregations of crowd labels.

To determine gold standard labels for the test set, the annotations from the two trained annotators for a given instance were averaged, unless the annotators disagreed strongly (e.g., a 0 and a 3) or if one of the annotators did not agree with the score produced by averaging. In those cases (111 total), instances were forwarded to a third-party adjudicator to make the final decision. Inter-annotator agreement (measured using kappa with quadratic weights between the four "classes" of 0, 1, 2, and 3) across all 3162 gold standard instances, prior to any adjudication or annotator discussion, was $\kappa = 0.435$, demonstrating that scoring metaphor novelty is a complex task even for humans. However, most of the annotators' disagreements were minor; when relaxing our agreement constraints and considering scores within a distance of 1 from one another to agree, as is the convention established by Mohler et al. (2016), $\kappa = 0.897$. Our published dataset includes the original crowdsourced annotations for each instance, the aggregated label for each training instance, and the gold standard label for test instance.

Dataset Statistics

As in naturally-occurring text, in our dataset there are many more instances near the lower end of the metaphor novelty spectrum (e.g., "She *spent* five *hours* on that!") than at the higher end (e.g., "She had a *technicolor personality*."). The label distribution across the full dataset is shown in Figure 1 (binned in 0.125 intervals). The word pairs in our dataset include occurrences of 4079 (3990 when case-insensitive) unique metaphoric words (as per original VUAMC annotations); thus, the metaphors in our dataset encompass a more diverse vocabulary than has been seen in prior work.

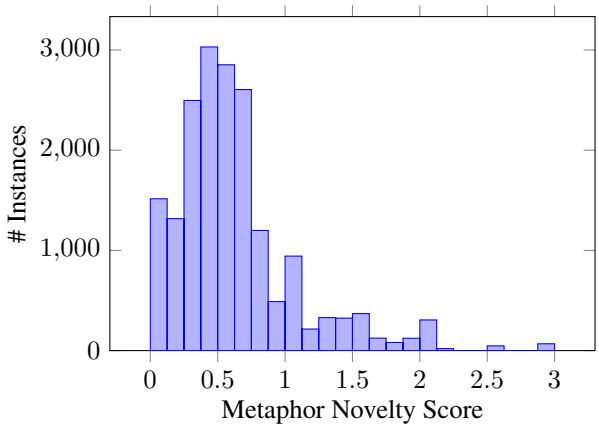

Figure 1: Metaphor Novelty Label Distribution

Method

We frame metaphor novelty scoring as a supervised regression task, in which features extracted from word pairs labeled with continuous values ranging from 0 (not metaphoric) to 3 (highly novel metaphor) are used to train a model to predict the metaphor novelty of unseen test pairs. We introduce some new features expected to be beneficial for scoring metaphor novelty specifically, and examine some existing features (previously found useful for standard metaphor detection) in this new context.

Feature Description

We divide our features into subsets and justify their use in the subsections below, describing them in Tables 2 and 3. Some of these features are frequency-based (e.g., co-occurrence), while others take semantic content into account (e.g., psycholinguistic features, word vectors, and topic probability features, among others).

Psycholinguistic Features: We build several feature sets based on the psycholinguistic characteristics of concreteness, imageability, sentiment, and ambiguity. Specifically, for each psycholinguistic characteristic we compute the features described in Table 2.

Concreteness and Imageability: Many researchers have found concreteness (Tsvetkov et al. 2014; Beigman Klebanov et al. 2014; Beigman Klebanov, Leong, and Flor 2015; Beigman Klebanov et al. 2016; Gargett and Barnden 2015; Tekiroglu, Özbal, and Strapparava 2015; Jang, Wen, and Rosé 2015; Özbal et al. 2016) and imageability scores (Broadwell et al. 2013; Tsvetkov et al. 2014; Gargett and Barnden 2015; Tekiroglu, Özbal, and Strapparava 2015; Özbal et al. 2016) to be useful for general metaphor detection. We wished to explore whether the usefulness of these features transfers to scoring metaphor novelty. We computed each of the features in Table 2 for both concreteness and

Feature	Description
Score	2: The scores (from the resource(s) associated with the specified characteristic) for the governor and modifier in the pair. Missing values are filled with a score of 0.
Score Diff.	1: Absolute difference between the scores.
Score Macro Avg.	2: Let W be the set of content words in the sentence not including either word in the pair, and $\text{SDIFF}(x, y)$ be the score difference between words x and y. $\text{MACROAVG}(g) = (\text{SDIFF}(g, m) + \frac{\sum_{w \in W} \text{SDIFF}(g,w)}{\lvert W \rvert})/2$ for the governor (g), and we also compute $\text{MACROAVG}(m)$ for the modifier (m).

Table 2: Psycholinguistic Features

imageability, using concreteness scores from the Brysbaert concreteness dataset (2014) (scaled to a 0-1 range) and from the MRC Psycholinguistic Database (Wilson 1988) (MRCPD), and imageability scores from the expanded MRC+ dataset (Liu et al. 2016).

Sentiment: Metaphors generally convey more emotion than literal statements, and when presented with pairs of sentences, humans often rate the more metaphoric sentence as also being more emotional (Mohammad, Shutova, and Turney 2016). Thus, it may be hypothesized that the level of emotion associated with one or both of the words in a pair is indicative of the pair's metaphor novelty. Jang et al. (2016) explored sentiment-based features for general metaphor detection by including counts of words that fell into the LIWC affective processes categories (positive emotion, negative emotion, anxiety, anger, and sadness) in the context of discussion posts. Gargett and Barnden (2015) employed valence, arousal, and dominance scores from the ANEW dataset as features in their binary metaphor detection approach. We compute the features in Table 2 using SentiWordNet (Baccianella, Esuli, and Sebastiani 2010) to explore the role of sentiment in predicting metaphor novelty.

Ambiguity: Although ambiguity has not previously been used as a feature for general metaphor detection, it may be useful for predicting metaphor novelty; the words comprising conventional metaphors may have higher ambiguity scores if their metaphoric sense is common enough to be considered a case of regular polysemy. We use MRCPD's ambiguity scores to compute the features in Table 2 to test this theory.

Co-Occurrence: We also compute features based on co-occurrence statistics for each word pair. We describe these features in Table 3 (features prefaced by *PMI*). Although the specific features we implement have not been used to detect metaphor in the past, other researchers have incorporated somewhat similar strategies. Tekiroglu, Özbal, and Strapparava (2015) sought to determine whether sensorial words were likelier to be metaphors by computing the co-occurrence frequency of potentially metaphoric words with words representing each of the five senses, and Schulder and Hovy (2014) also found term relevance to be a useful feature for binary word-level metaphor detection. We expect word co-occurrence may be useful for scoring metaphor novelty simply because the words in highly novel metaphors are unlikely to have occurred together in text frequently, whereas the words forming conventional metaphors likely have. We compute our co-occurrence features using n-gram counts from Google Web1T (Brants and Franz 2006).

Syntax-based and/or Positional (SynPos): We include several features based on syntax or word position in our feature set, from which we expect the classifier to learn simple rules. The most common of these in prior work have been part-of-speech tags, which Schulder and Hovy (2014), Özbal et al. (2016), and Beigman Klebanov, Leong, and Flor (2014; 2015) all included in their binary metaphor detection approaches with success. Our syntax-based and positional features are *Word Distance*, *Dependency Type*, and *POS* in Table 3.

Conceptual Features: Tsvetkov et al. (2014), Shutova, Kiela, and Maillard (2016), and Gutierrez et al. (2016) have previously incorporated *word embeddings* in their binary metaphor detection work. These could aid in predicting metaphor novelty by providing some conceptual insight regarding the words' semantic properties, as well as additional co-occurrence information. We use Google's pretrained Google News embeddings to generate 300-dimensional vectors for the governor and modifier, and include these as features. We also compute the *cosine similarity* between the two feature vectors.

Researchers have also had success employing *topic models* for standard metaphor detection in the past. Jang et al. (2016) considered the topic of the sentence containing the word being classified, and how that topic differed from surrounding sentences. Beigman Klebanov et al. (2014), Jang, Wen, and Rosé (2015), and Özbal et al. (2016) include features based on a word's probability of belonging to different topic models, and Broadwell et al. (2013) use topic chains (a noun phrase and all subsequent noun phrases that refer to the original noun phrase) as part of their approach to binary metaphor detection. We hypothesize that some conceptual mappings may be more likely to produce novel/conventional metaphors than others; topic modeling allows us to test this theory by automatically predicting the conceptual domains to which the words in the pair belong. We build topic models from two sources: Project Gutenberg books,[6] and English Wikipedia.[7] For each set, we compute *Topic Probability* and *Probability Word in Top Topic* features, described in Table 3.

SynSet: Finally, we build several features incorporating lexical information from WordNet (Fellbaum 1998). These features are designed to capture the number of senses associated with a word. Similarly to the ambiguity features, the synset features are driven by the premise that words with a larger number of commonly-known senses are likelier to be present in conventional metaphors than novel metaphors.

[6] https://www.gutenberg.org/

[7] https://www.wikipedia.org/

Feature	Description
PMI	1: The PMI between the two words. Distance between words was considered when computing frequencies; thus, if g was the fifth word in a sentence and m was the seventh word, $\text{FREQ}(g)$ would be the number of trigrams in Web1T beginning with g, $\text{FREQ}(m)$ would be the number of trigrams ending with m, and $\text{FREQ}(g, m)$ would be the number of trigrams beginning with g and ending with m. The probability of g occurring, $\text{P}(g)$, was then equivalent to $\text{FREQ}(g)$ divided by the summed frequency of all trigrams in Web1T. PMI between the two words was therefore: $\text{PMI}(g, m) = \log \frac{\text{P}(g,m)}{\text{P}(g) \times \text{P}(m)}$.
PMI with Sentence	2: The average PMI between each word in the pair and the other content words in the surrounding sentence, excluding the other word in the pair.
PMI Span	2: The PMI between the governor g (or modifier) and the span s of text from g through the modifier m (or governor) (e.g., in "*frowning* like a *thunderstorm*" where m is *thunderstorm* and g is *frowning*, PMI is computed between "frowning like a" and "thunderstorm"). Computed as: $\text{PMI}(s, g) = \log \frac{\text{P}(s,g)}{\text{P}(s) \times \text{P}(g)}$.
PMI w/Sentence − PMI	2: The difference between each of the values computed in PMI with Sentence and the PMI between g and m.
Word Distance	1: Absolute distance between g and m, by number of words.
Dependency Type	24: One-hot encoded vector representing the syntactic dependency type relating g and m.
POS	30: One-hot encoded vectors representing the parts of speech for both g (15) and m (15).
Word Vector	600: For each word in the pair, a 300-dimensional Word2Vec (Mikolov et al. 2013) embedding.
Cosine Similarity	1: The cosine similarity between the two embeddings.
Topic Probability	400: The probability that each word belongs to each of 200 topics (100 learned from Project Gutenberg and 100 learned from Wikipedia). Topic models were trained using latent dirichlet allocation.[8]
Probability Word in Top Topic	4: For each set of topic models, for both g and m, the probability that the word belongs to the topic to which the other word had the highest probability of belonging.
Max (Min) SynSets	2: The maximum (minimum) of two values: the number of synsets for g, and the number of synsets for m.
SynSet Avg.	1: The average number of synsets for g and m.
SynSet Diff.	1: The absolute value of the difference between the number of synsets for g and m.

Table 3: Other Features

Regression Approach

We implemented our approach using a deep neural network. To tune parameters, including the number of hidden layers, we randomly split the training set into 75% training and 25% validation such that all instances originating from the same document remained in the same subset. Optimal parameters as determined via tuning on the validation set are presented in Table 4. The activation and dropout noted in the table were applied to all layers excluding the output layer. The neural network was implemented using Keras[9] with TensorFlow[10].

Parameter	Value
Layers	5
Inputs	1091
Units in Hidden Layer 1	256
Units in Hidden Layer 2	32
Units in Hidden Layer 3	16
Units in Hidden Layer 4	8
Units in Layer 5 (Output Layer)	1
Activation	SoftSign
Dropout	0.1
Kernel Initializer	Glorot Normal
Optimizer	Nadam
Loss Function	Mean Squared Error
Epochs	5
Train Batch Size	32
Test Batch Size	16

Table 4: Neural Network Parameters

Evaluation

Each experimental case was trained on the training data (15,290 instances with aggregated labels), and tested on the test data (3162 instances with gold labels). Since our system's output is continuous, we report the correlation coefficient (*r*) and root mean squared error (RMSE) for each case. We detail our overall performance evaluation and provide a feature analysis in the subsections below.

Metaphor Novelty Experiments

Baseline Approaches: As noted earlier, Haagsma and Bjerva (2016) were ultimately unable to build their metaphor novelty detector due to lack of adequate data. It is additionally impossible for us to compare one-to-one with Krishnakumaran and Zhu's (2007) approach because theirs was designed to classify sentences (not word pairs) and was not equipped to handle potential non-metaphors. Their rules also relied entirely on WordNet, which was not an issue with their dataset; when analyzing a random sample of their data

[8] https://radimrehurek.com/gensim/

[9] https://keras.io/

[10] https://www.tensorflow.org/

we found that all words that their algorithm could have considered were on WordNet. However, our data comes from less-convenient real-world texts so this is not true of our instances.[11] The results that we provide herein will establish a benchmark to accelerate future research in metaphor novelty scoring. Since it is impossible to compare directly to prior work on this task at the present, we compare to the following here:

Random: A simple baseline that outputs a random continuous variable ranging between 0 and 3 for each instance.

Distribution-Aware Random: Learns a probability density function[12] from the training set labels and outputs a random continuous variable following that distribution for each instance.

Mean Value: Predicts the mean training value for each instance.

Tsvetkov et al.: Trains and tests Tsvetkov et al.'s (2014) metaphor detection approach on our dataset. Tsvetkov et al.'s approach learns a random forest classifier from conceptual semantic features (including abstractness and imageability, WordNet supersenses, and vector space word representations) to predict whether subject-verb-object triples or adjective-noun pairs are metaphoric. Their approach was designed to learn from discrete, binary classes; to use it with our dataset, we modified their source code such that it trains scikit-learn's Random Forest Regressor (which learns from and predicts continuous values) instead of scikit-learn's Random Forest Classifier (which requires discrete labels). We compare to this approach to validate that *scoring metaphor novelty* is a distinct task from *regression-based metaphor detection*, and thus simply applying even extremely high-performing metaphor detection approaches may not yield outstanding results (Tsvetkov et al.'s approach achieved an accuracy of 86% for general metaphor detection on adjective-noun pairs).

Results: We compare our method (training and testing on all valid types of word pairs) with *Random*, *Distribution-Aware Random*, and *Mean Value* in Table 5, and our method with *Tsvetkov et al.*'s metaphor detection approach (training and testing only on adjective-noun pairs since their approach was not designed for other types of word pairs) in Table 6. The adjective-noun subset of our dataset included 3151 instances, and its labels were distributed similarly to the dataset as a whole. We ran our approach 10 times[13] and reported the average r and RMSE.

As seen in Tables 5 and 6, our approach outperforms the alternatives in terms of both r and RMSE. Improvement in r for our approach is orders of magnitude higher relative to *Random*, *Distribution-Aware Random*, and *Mean Value*, and a 65.1% improvement over *Tsvetkov et al*. Decreases in RMSE for our approach are 55.7%, 20.2%, 9.6%, and 6.3% relative to *Random*, *Distribution-Aware Random*, *Mean Value*, and *Tsvetkov et al.*, respectively. All results reported are statistically significant ($p < 0.0001$).

[11]Note that it would likewise be problematic for us to run our approach on their dataset because we could not guarantee that we'd be considering the same pairs as them (the pairs considered by their algorithm are not explicitly specified in the dataset), and we would have to label pairs based on their sentence's label since pair-level annotations are not provided. The presence of a potentially large number of mislabels could significantly affect results.

[12]We do this using SciPy's Gaussian kernel density estimator.

[13]Weights are initialized with a random seed.

Method	***r***	**RMSE**
Random	0.0048	1.4658
Distribution-Aware Random	0.0007	0.8143
Mean Value	0.0000	0.7192
Ours	**0.4600**	**0.6502**

Table 5: Comparison with Baseline Approaches

Method	***r***	**RMSE**
Tsvetkov et al.	0.2716	0.7804
Ours	**0.4483**	**0.7312**

Table 6: Comparison with Tsvetkov et al.'s Approach (only adjective-noun pairs per their design)

Feature Analysis

Feature Sets: We conduct a comparative analysis of the feature sets described earlier (CONCRETENESS, IMAGEABILITY, SENTIMENT, AMBIGUITY, CO-OCCURRENCE, SYNPOS, TOPIC, EMBEDDING, and SYNSET) to provide a general understanding of which types of features contribute most effectively toward scoring metaphor novelty. We first perform an ablation study by removing one feature set at a time from the set of all features. Then, with the highest-performing (i.e., its removal leads to the largest drop in r) feature set, we conduct a bottom-up feature evaluation by adding one feature set at a time to it to further tease apart the contributions of individual feature sets.

Results: The results of our ablation study are presented in Table 7. The highest-performing individual feature set is SYNPOS, the syntax-based and positional features. The removal of those features leads to a 21.8% drop in correlation relative to ALL. EMBEDDINGS are a close second; their removal leads to a 17.8% drop in correlation. Since the highest-performing feature set in the ablation study is SYNPOS, we use it as the basis of our bottom-up feature analysis (see Table 8). We find that combining EMBEDDINGS with SYNPOS yields the highest performance. SYNPOS and EMBEDDINGS were also the only two types of features whose removal led to significant differences in the predicted scores versus using all features, and only SYNPOS + EMBEDDING was significantly different from SYNPOS († indicates statistical significance ($p < 0.001$) in Tables 7 and 8).

Discussion

The results unearth some interesting findings. The success of our approach relative to Tsvetkov et al.'s, a 65.1% improvement when training and testing on the same instances, substantiates our position that highly effective metaphor detection approaches cannot simply be transferred to scoring

Feature Set	*r*	RMSE
ALL	0.4600	0.6502
ALL - CONCRETENESS	0.4597	0.6501
ALL - IMAGEABILITY	0.4568	0.6529
ALL - SENTIMENT	0.4581	0.6514
ALL - AMBIGUITY	0.4551	0.6551
ALL - CO-OCCURRENCE	0.4530	0.6470
ALL - SYNPOS	**0.3595**†	**0.6792**†
ALL - TOPIC	0.4563	0.6511
ALL - EMBEDDING	0.3779†	0.6696†
ALL - SYNSET	0.4579	0.6515

Table 7: Feature Ablation

Feature Set	*r*	RMSE
SYNPOS	0.3566	0.6733
SYNPOS + CONCRETENESS	0.3623	0.6722
SYNPOS + IMAGEABILITY	0.3610	0.6734
SYNPOS + SENTIMENT	0.3623	0.6728
SYNPOS + AMBIGUITY	0.3578	0.6729
SYNPOS + CO-OCCURRENCE	0.3596	0.6729
SYNPOS + TOPIC	0.3581	0.6728
SYNPOS + EMBEDDING	**0.4523**†	**0.6484**†
SYNPOS + SYNSET	0.3568	0.6731

Table 8: Additional Feature Analysis

metaphor novelty with the expectation of similarly high performance.

We discover that SYNPOS features are particularly discriminative with our data. Since our dataset contains many instances that straddle the boundary between non-metaphoric and highly conventionalized metaphors, this indicates that learned rules based on part-of-speech combinations, dependency types, and word distance are relatively effective at delineating that boundary without the introduction of additional semantic or co-occurrence features. Word embeddings were the second-highest performing feature set and also led to the largest performance increase when paired with SYNPOS in the bottom-up feature experiment. The usefulness of EMBEDDINGS indicates that alongside shallower syntactic cues, deeper semantic context is an important clue in determining metaphor novelty. In contrast, psycholinguistic features used commonly for general metaphor detection proved to be only mildly discriminative between shades of metaphor novelty. Likewise, topic-based features were not particularly useful, suggesting that both novel and conventional metaphors arise from a wide and often overlapping range of cross-domain mappings.

Conclusion

In this work, we introduce and make publicly available a new, large dataset in an underdeveloped terrain: automatically scoring metaphor novelty. We perform a comparative feature analysis to study the performance of both novel features and features originating from general metaphor detection for this task. Finally, we contribute a regression approach that learns from these features to automatically score metaphor novelty, finding that the approach outperforms a strong general metaphor detection-based approach by 65.1%. This provides evidence that scoring metaphor novelty is a distinct task from metaphor detection with its own unique nuances. We hope the dataset, benchmarks, and code released here will significantly stimulate and advance related metaphor research.

Acknowledgments

This material is based upon work supported by the NSF Graduate Research Fellowship Program under Grant 1144248, and the NSF under Grant 1262860. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Amanzio, M.; Geminiani, G.; Leotta, D.; and Cappa, S. 2008. Metaphor comprehension in alzheimer's disease: Novelty matters. *Brain and Language* 107(1):1–10.

Baccianella, S.; Esuli, A.; and Sebastiani, F. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proc. of the 7th Intl. Conf. on Language Resources and Evaluation*, 2200–2204.

Beigman Klebanov, B.; Leong, B.; Heilman, M.; and Flor, M. 2014. Different texts, same metaphors: Unigrams and beyond. In *Proc. of the Second Workshop on Metaphor in NLP*, 11–17.

Beigman Klebanov, B.; Leong, C. W.; Gutierrez, E. D.; Shutova, E.; and Flor, M. 2016. Semantic classifications for detection of verb metaphors. In *Proc. of the 54th Annual Meeting of the ACL*, 101–106.

Beigman Klebanov, B.; Leong, C. W.; and Flor, M. 2015. Supervised word-level metaphor detection: Experiments with concreteness and reweighting of examples. In *Proc.of the Third Workshop on Metaphor in NLP*, 11–20.

Birke, J., and Sarkar, A. 2006. A clustering approach for the unsupervised recognition of nonliteral language. In *Proc. of the 11th Conf. of the European Chapter of the ACL*.

Brants, T., and Franz, A. 2006. Web 1t 5-gram v1.

Broadwell, G. A.; Boz, U.; Cases, I.; Strzalkowski, T.; Feldman, L.; Taylor, S.; Shaikh, S.; Liu, T.; Cho, K.; and Webb, N. 2013. Using imageability and topic chaining to locate metaphors in linguistic corpora. In *Proc. of the 6th Intl. Conf. on Social Computing, Behavioral-Cultural Modeling and Prediction*, 102–110.

Brysbaert, M.; Warriner, A. B.; and Kuperman, V. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods* 46(3):904–911.

Dunn, J.; de Heredia, J. B.; Burke, M.; Gandy, L.; Kanareykin, S.; Kapah, O.; Taylor, M.; Hines, D.; Frieder, O.; Grossman, D.; Howard, N.; Koppel, M.; Morris, S.; Ortony, A.; and Argamon, S. 2014. Language-independent

ensemble approaches to metaphor identication. In *Proc. of the AAAI Workshop on Cognitive Computed for Augmented Human Intelligence*, 6–12.

Dunn, J. 2013. What metaphor identification systems can tell us about metaphor-in-language. In *Proc. of the First Workshop on Metaphor in NLP*, 1–10.

Fellbaum, C., ed. 1998. *WordNet: An Electronic Lexical Database*. Language, speech, and communication. MIT Press.

Gandy, L.; Allan, N.; Atallah, M.; Frieder, O.; Howard, N.; Kanareykin, S.; Koppel, M.; Last, M.; Neuman, Y.; and Argamon, S. 2013. Automatic identification of conceptual metaphors with limited knowledge. In *Proc. of the AAAI Conf. on Artificial Intelligence*, 328–334.

Gargett, A., and Barnden, J. 2015. Modeling the interaction between sensory and affective meanings for detecting metaphor. In *Proc. of the Third Workshop on Metaphor in NLP*, 21–30.

Gibbs, R. W. 1984. Literal meaning and psychological theory. *Cognitive Science* 8(3):275–304.

Gutierrez, E. D.; Shutova, E.; Marghetis, T.; and Bergen, B. 2016. Literal and metaphorical senses in compositional distributional semantic models. In *Proc. of the 54th Annual Meeting of the ACL*, 183–193.

Haagsma, H., and Bjerva, J. 2016. Detecting novel metaphor using selectional preference information. In *Proc. of the Fourth Workshop on Metaphor in NLP*, 10–17.

Jang, H.; Moon, S.; Jo, Y.; and Rosé, C. 2015. Metaphor detection in discourse. In *Proc. of the SIGDIAL 2015 Conf.*, 384–392.

Jang, H.; Jo, Y.; Shen, Q.; Miller, M.; Moon, S.; and Rosé, C. 2016. Metaphor detection with topic transition, emotion and cognition in context. In *Proc. of the 54th Annual Meeting of the ACL*, 216–225.

Jang, H.; Wen, M.; and Rosé, C. 2015. Effects of situational factors on metaphor detection in an online discussion forum. In *Proc. of the Third Workshop on Metaphor in NLP*, 1–10.

Krishnakumaran, S., and Zhu, X. 2007. Hunting elusive metaphors using lexical resources. In *Proc. of the Workshop on Comput. Approaches to Figurative Language*, 13–20.

Lai, V. T.; Curran, T.; and Menn, L. 2009. Comprehending conventional and novel metaphors: An ERP study. *Brain Research* 1284:145–155.

Lakoff, G. 1994. *Master Metaphor List*. Univ. of California.

Levin, L.; Mitamura, T.; Macwhinney, B.; Fromm, D.; Carbonell, J.; Feely, W.; Frederking, R.; Gershman, A.; and Ramirez, C. 2014. Resources for the detection of conventionalized metaphors in four languages. In *Proc. of the 9th Intl. Conf. on Language Resources and Evaluation*.

Liu, T.; Cho, K.; Strzalkowski, T.; Shaikh, S.; and Mirzaei, M. 2016. The validation of mrcpd cross-language expansions on imageability ratings. In *Proc. of the 10th Intl. Conf. on Language Resources and Evaluation*, 3748–3751.

Manning, C. D.; Surdeanu, M.; Bauer, J.; Finkel, J.; Bethard, S. J.; and McClosky, D. 2014. The Stanford CoreNLP natural language processing toolkit. In *ACL System Demonstrations*, 55–60.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Proc. of the 26th Intl. Conf. on Neural Information Processing Systems*, 3111–3119.

Mohammad, S. M.; Shutova, E.; and Turney, P. D. 2016. Metaphor as a medium for emotion: An empirical study. In *Proc. of the 5th Joint Conf. on Lexical and Computational Semantics*.

Mohler, M.; Bracewell, D.; Tomlinson, M.; and Hinote, D. 2013. Semantic signatures for example-based linguistic metaphor detection. In *Proc. of the First Workshop on Metaphor in NLP*, 27–35.

Mohler, M.; Brunson, M.; Rink, B.; and Tomlinson, M. 2016. Introducing the lcc metaphor datasets. In *Proc. of the 10th Intl. Conf. on Language Resources and Evaluation*.

Özbal, G.; Strapparava, C.; Tekiroglu, S. S.; and Pighin, D. 2016. Learning to identify metaphors from a corpus of proverbs. In *Proc. of the Conf. on Empirical Methods in Natural Language Processing*, 2060–2065.

Parde, N., and Nielsen, R. D. 2017. Finding patterns in noisy crowds: Regression-based annotation aggregation for crowdsourced data. In *Proc. of the Conf. on Empirical Methods in Natural Language Processing*, 1908–1913.

Schulder, M., and Hovy, E. 2014. Metaphor detection through term relevance. In *Proc. of the Second Workshop on Metaphor in NLP*, 18–26.

Shutova, E.; Kiela, D.; and Maillard, J. 2016. Black holes and white rabbits: Metaphor identification with visual features. In *Proc. of the Conf. of the North American Chapter of the ACL*, 160–170.

Shutova, E.; Sun, L.; and Korhonen, A. 2010. Metaphor identification using verb and noun clustering. In *Proc. of the 23rd Intl. Conf. on Computational Linguistics*, 1002–1010.

Shutova, E. 2015. Design and evaluation of metaphor processing systems. *Comput. Linguist.* 41(4):579–623.

Steen, G. 2010. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. Converging evidence in language and communication research. John Benjamins Publishing Company.

Tekiroglu, S. S.; Özbal, G.; and Strapparava, C. 2015. Exploring sensorial features for metaphor identification. In *Proc. of the Third Workshop on Metaphor in NLP*, 31–39.

Tsvetkov, Y.; Boytsov, L.; Gershman, A.; Nyberg, E.; and Dyer, C. 2014. Metaphor detection with cross-lingual model transfer. In *Proc. of the 52nd Annual Meeting of the ACL*, 248–258.

Turney, P.; Neuman, Y.; Assaf, D.; and Cohen, Y. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proc. of the Conf. on Empirical Methods in Natural Language Processing*, 680–690.

Wilson, M. 1988. Mrc psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers* 20(1):6–10.