Predicting Humorousness and Metaphor Novelty with Gaussian Process Preference Learning

Edwin Simpson[*] and **Erik-Lân Do Dinh**[*] and **Tristan Miller**[*†] and **Iryna Gurevych**[*]

[*]Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt
`https://www.ukp.tu-darmstadt.de/`

[†]Austrian Research Institute for Artificial Intelligence (OFAI)
Freyung 6, 1010 Vienna, Austria
`http://www.ofai.at/`

Abstract

The inability to quantify key aspects of creative language is a frequent obstacle to natural language understanding. To address this, we introduce novel tasks for evaluating the creativeness of language—namely, scoring and ranking text by humorousness and metaphor novelty. To sidestep the difficulty of assigning discrete labels or numeric scores, we learn from pairwise comparisons between texts. We introduce a Bayesian approach for predicting humorousness and metaphor novelty using Gaussian process preference learning (GPPL), which achieves a Spearman's ρ of 0.56 against gold using word embeddings and linguistic features. Our experiments show that given sparse, crowdsourced annotation data, ranking using GPPL outperforms best–worst scaling. We release a new dataset for evaluating humour containing 28,210 pairwise comparisons of 4030 texts, and make our software freely available.

1 Introduction

Creative language, such as humour and metaphor, is an essential part of everyday communication, yet remains a challenge for computational methods. Unlike much literal language, humour and figurative language require complex linguistic and background knowledge to understand, which are difficult to integrate with NLP methods (Hempelmann, 2008; Shutova, 2010).

An important step in processing creative language is to recognise its presence in a piece of text. Humour and metaphors are two of the most frequently used types of creative language whose use most obscures the true meaning of a piece of text from its surface interpretation (Raskin, 1985, pp. 1–5, 100–104; Black, 1955) and whose attributes, such as funniness and novelty, may be present or perceived to varying degrees (Bell, 2017; Dunn, 2010). For example, the level of appreciation (i.e., *humorousness* or equivalently *funniness*) of jokes can vary according to their content and structural features, such as nonsense or disparagement (Carretero-Dios et al., 2010) or, in the case of puns, contextual coherence (Lippman and Dunn, 2000) and the cognitive effort required to recover the target word (Hempelmann, 2003, pp. 123–124).

With metaphors, the literal meaning of frequently used metaphors can drop out of everyday usage, leaving the metaphorical sense as the expected one (Shutova, 2015). For such *conventionalised* metaphors, NLP methods may identify the metaphorical sense from training data or resources such as WordNet, whereas novel metaphors require the ability to recognise the analogy being made.

While previous work (see §2) has considered mainly binary classification approaches to humour or metaphor recognition, this paper focuses on quantifying *humorousness* and *metaphor novelty*. These tasks are important for downstream applications such as conversational agents or machine translation, which must choose the correct tone in response to humour, or find appropriate metaphors or wordplay in a target language. The degree of creativeness may also inform an application whether the semantics of a metaphor or joke can be inferred from similar examples.

The examples in Tables 1 and 2 illustrate the difficulty of classifying text as humorous or metaphorical: in both cases, the examples are at least somewhat humorous or somewhat metaphorical, which makes it harder to assign discrete labels such as "funny"/"not funny" or "metaphor"/"literal". Alternatively, we could assign numerical scores to quantify the humorousness or novelty. However, this can present problems for establishing a gold standard, as human annotators can assign scores inconsistently over time or interpret scores differently to one another (Ovadia, 2004; Yannakakis and Hallam, 2011; Kiritchenko and Mohammad, 2017). For example, if assigning scores between zero and

Money is the Root of All Evil. For more info, send $10.
“Have you seen my collection of ancient Chinese artifacts?” asked Tom charmingly.

Table 1: Examples from the SemEval-2017 Task 7 dataset (Miller et al., 2017). The upper example was among those rated funniest by our annotators, while the lower example was among those rated least funny (presumably due to its very tortured pun on “Ming”).

girls often produce responses like ‘often **go** through a bad patch for a year’
‘when you tried to read the book, there was nothing there, because the words started as a **coat-hanger** to hang pictures on.’

Table 2: Examples of statements from the Metaphor Novelty dataset (Do Dinh et al., 2018) containing highlighted metaphors. The upper example is highly conventionalised, while the lower is more novel and creative.

ten, some annotators may choose middling values while others may prefer extremes.

To improve the reliability of annotations, we ask annotators to compare pairs of texts and choose the funniest or most metaphorically novel of the two. Unlike categorical labels, pairwise labels allow a total sorting of the texts since they avoid items having the same value, and can reduce the time taken to label a dataset (Yang and Chen, 2011; Kingsley and Brown, 2010; Kendall, 1948). Pairwise labels can be used to infer scores or rankings using techniques such as learning-to-rank (Joachims, 2002), preference learning (Thurstone, 1927), or best–worst scaling (Flynn and Marley, 2014). A drawback of pairwise labelling is that the number of possible pairs scales with $O(n^2)$, which becomes impractical for large datasets. To reduce annotation costs and enable quicker learning in new domains, it is therefore desirable to learn from sparse datasets rather than exhaustive pairwise labels.

We establish four new tasks for scoring and ranking texts with both sparse and extensive sets of pairwise training labels. We apply these tasks to datasets for humorousness and metaphor novelty, which extend the datasets of Miller et al. (2017) and Do Dinh et al. (2018), respectively, and contain crowdsourced pairwise labels. As a baseline scoring method, we employ the scoring technique for best–worst scaling (BWS; Flynn and Marley, 2014), an established method that can also be applied to pairwise labels to estimate scores very efficiently. Our use of sparse, unreliable crowdsourced data motivates a second, Bayesian approach: Gaussian process preference learning (GPPL; Simpson and Gurevych, 2018), which exploits text features to boost performance when labels are sparse and make predictions for items not compared in the training set.

Our main contributions are (1) four novel tasks for quantifying aspects of creative language, (2) an annotated dataset containing pairwise comparisons of humorousness between sentences, (3) a Bayesian approach for scoring short texts by humorousness and metaphor novelty given sparse pairwise annotations, and (4) an empirical investigation showing that word embeddings and linguistic features can be used to predict humorousness and metaphor novelty, and that GPPL outperforms BWS when faced with sparse data. We publish the datasets and software[1] to encourage further research on these tasks, and to serve the needs of qualitative humanities research into humour and metaphor.

2 Background and Related Work

2.1 Humorousness

The automatic processing of verbal humour has applications in human–computer interaction, machine and machine-assisted translation, and the digital humanities (Miller et al., 2017). To give just one example, an intelligent conversational agent should ideally detect and respond appropriately to comments made in jest. The vast majority of past approaches to the automatic recognition of humour (e.g., Mihalcea and Strapparava, 2006; Purandare and Litman, 2006; Sjöbergh and Araki, 2007; Mihalcea et al., 2010; Zhang and Liu, 2014; Yang et al., 2015; Miller et al., 2017; Mikhalkova and Karyakin, 2017; Chen and Soo, 2018) have framed the problem as a binary classification task, which is sufficient for the detection step of our example. However, the ability to assess the *degree* of humour embodied in an utterance may be necessary for the agent to make a contextually appropriate, human-like response – for example, a groan for a terrible joke, a chuckle for a middling one, or uproarious laughter for a clever one.

Only a few studies have dealt with determining the (relative) funniness of texts. Shahaf et al. (2015) presented a supervised system for determining which of a given pair of cartoon captions is funnier, using features such as sentiment, perplex-

[1] `https://github.com/ukplab/acl2019-GPPL-humour-metaphor`

ity, readability, and keyword descriptions of the cartoon image and its anomalies. While the method achieves promising results (64% accuracy, versus 55% for a bag-of-words baseline), it cannot quantify humorousness on a continuum; multiple captions can be ranked only tournament-style. Moreover, the keyword features are specific to visual rather than verbal humour, and must be manually sourced at great expense, making the method unsuitable for classifying unseen examples. In parallel work, Radev et al. (2016) tested various heuristics for ranking pairs or sets of the same captions by funniness. Such heuristics included tf–idf, *n*-gram frequency, syntactic complexity, and references to objects in the cartoon (which, again, is specific to this multimodal form of humour and depends on manual annotation). The heuristics were evaluated in isolation, rather than as part of a supervised or ensemble classifier. This, combined with the study's unusual evaluation metrics, precludes a meaningful comparison with Shahaf et al. (2015).

More recently, the #HashtagWars evaluation campaign (Potash et al., 2017) defined two humour ranking tasks for Twitter data. The organisers compiled data from a TV game show whose producers solicit funny tweets for a given hashtag and then partition them into three sets: the funniest tweet, nine runners-up, and the remainder. The campaign had two computational tasks: (a) given a pair of tweets from different sets, determine which tweet is funnier; and (b) classify all tweets according to their set. As with Shahaf et al. (2015), the determination of humour here was coarse-grained, with no attempt to quantify it. A similar corpus (but no classification experiment) was presented by Castro et al. (2018b) and later developed into a shared task (Castro et al., 2018a). The dataset's crowd annotators were asked to classify the humorousness of tweets on a Likert scale, grouping them into five sets versus Potash et al.'s (2017) three. Mindful of psychological studies on subjective evaluations (Thurstone, 1927), Shahaf et al. (2015) reject the idea that such ordinal rating data can be treated as interval data, and argue that direct comparisons are preferable for humour judgements.

2.2 Metaphor Novelty

Most previous work on metaphor detection has been conducted with a binary classification in mind (metaphor vs. literal). This dichotomy is reflected in more widely used datasets, such as the VU Amsterdam Metaphor Corpus (VUAMC; Steen et al., 2010) or the datasets in multiple languages created by Tsvetkov et al. (2014). Advantages include the wide variety of approaches that can be (and have been) employed for automatic detection and a rather straightforward annotation process. This usually also entails a high interannotator agreement, meaning that the annotations are reliable. In the case of VUAMC, this amounts to a Cohen's κ of 0.80. However, the two-class modelling of metaphor has certain limits. These become obvious when looking at examples from the aforementioned datasets (see Table 2, which includes an example from VUAMC). In particular, many metaphors annotated in the binary datasets differ widely in their metaphoricity – i.e., their degree of being a metaphor. Thus, while the annotations might be reliable, they might not be very meaningful. A graded approach to metaphor better accommodates its subjective and fuzzy nature, but previous work taking such a fine-grained approach is less common.

Dunn (2014) conducted experiments regarding the notion of *metaphoricity* on a sentence basis. Using crowdsourcing, he obtained a small corpus of 60 sentences with metaphoricity scores between 0 (non-metaphoric) and 1 (highly metaphoric). This dataset was then used to determine various features from which a metaphoricity measure could be computed. Due to the lack of a large, graded evaluation corpus, the measure was tested on VUAMC along with a threshold relative to the number of contained metaphors. Haagsma and Bjerva (2016) employed clustering and neural network approaches using selectional preferences to detect novel metaphors. While the violation of selectional preferences had been used in general metaphor detection before, Haagsma and Bjerva (2016) argue that they are specifically indicative of novel metaphors as opposed to conventionalised ones. However, the authors also struggled with the lack of graded annotations to test their approach.

More recently, Parde and Nielsen (2018) and Do Dinh et al. (2018) created graded metaphoricity layers for VUAMC using crowdsourcing, with the former approach labelling grammatical constructions and the latter labelling tokens. However, manually labelling larger amounts of data is costly, even with crowdsourcing. Further, while VUAMC covers multiple domains, it is still limited in scope, size, and language. Thus, an approach is needed to generalise from few graded or ranked metaphor annotations to a larger corpus or different domains.

2.3 Learning from Pairwise Comparisons

Pairwise comparisons can be used to infer rankings or ratings by assuming a *random utility model* (Thurstone, 1927), meaning that the annotator chooses an instance with probability p, where p is a function of the *utility* of the instance. Therefore, when instances in a pair have similar utilities, the annotator selects one with a probability close to 0.5, while for instances with very different utilities, the instance with higher utility will be chosen consistently. The random utility model forms the core of two popular preference learning models, the Bradley–Terry model (Bradley and Terry, 1952; Luce, 1959; Plackett, 1975), and the Thurstone–Mosteller model (Thurstone, 1927; Mosteller, 1951). Given this model and a set of pairwise annotations, probabilistic inference can be used to retrieve the latent utilities of the instances.

Besides pairwise comparisons, a random utility model is also employed by MaxDiff (Marley and Louviere, 2005), a model for best–worst scaling (BWS), in which the annotator chooses the best and worst instances from a set. While the term "best–worst scaling" originally applied to the data collection technique (Finn and Louviere, 1992), it now also refers to models such as MaxDiff that describe how annotators make discrete choices. Empirical work on BWS has shown that MaxDiff scores (instance utilities) can be inferred using either maximum likelihood or a simple counting procedure that produces linearly scaled approximations of the maximum likelihood scores (Flynn and Marley, 2014). The counting procedure defines the score for an instance as the fraction of times the instance was chosen as best, minus the fraction of times the instance was chosen as worst, out of all comparisons including that instance (Kiritchenko and Mohammad, 2016). From this point on, we refer to the counting procedure as BWS, and apply it to the tasks of inferring scores from both best–worst scaling annotations for metaphor novelty and pairwise annotations for funniness.

To make predictions for unlabelled instances and cope better with sparse pairwise labels, Chu and Ghahramani (2005) proposed Gaussian process preference learning (GPPL), a Thurstone–Mosteller–based model that accounts for the features of the instances when inferring their scores. GPPL uses Bayesian inference, which has been shown to cope better with sparse and noisy data (Xiong et al., 2011; Titov and Klementiev, 2012; Beck et al., 2014; Lampos et al., 2014), including disagreements between multiple annotators (Cohn and Specia, 2013; Simpson et al., 2015; Felt et al., 2016; Kido and Okamoto, 2017). Through the random utility model, GPPL is able to handle disagreements between annotators as noise, since no label has a probability of one of being selected.

Given a set of pairwise labels, and the features of labelled instances, GPPL can estimate the posterior distribution over the utilities of any instances given their features. Relationships between instances are modelled by a Gaussian process (GP), which computes the covariance between instance utilities as a function of their features (see Rasmussen and Williams, 2006). Since typical methods for posterior inference (Nickisch and Rasmussen, 2008) are not scalable ($O(n^3)$, where n is the number of instances), Simpson and Gurevych (2018) introduced a scalable method for GPPL that permits arbitrarily large numbers of instances and pairs. This method uses stochastic variational inference (Hoffman et al., 2013), which limits computational complexity by substituting the instances for a fixed number of *inducing points* during inference.

Simpson and Gurevych (2018) applied GPPL to ranking arguments by convincingness, which, like funniness and metaphor novelty, is an abstract linguistic property that is hard to quantify directly. They found that GPPL outperformed SVM and BiLSTM regression models that were trained directly on gold-standard scores. Regression approaches are also unsuitable for our scenario, since utilities for training the regression model would first need to be estimated from pairwise labels using, for example, BWS. This type of pipeline approach often suffers from error propagation, which integrated methods such as GPPL avoid (Finkel et al., 2006). We therefore propose the use of GPPL for our creative language tasks to provide a strong baseline that, unlike BWS, can exploit textual features as well as pairwise labels.

3 Data

Humour dataset. Our humour dataset is an extension of the data provided for the SemEval-2017 pun recognition challenge (Miller et al., 2017). Several factors motivated our selection of this dataset: (1) Unlike the multimodal datasets of Shahaf et al.

(2015) and Radev et al. (2016), the humour in Miller et al. (2017) is purely verbal. (2) Unlike the cartoon caption and Twitter datasets used in previous studies, the SemEval-2017 jokes were sourced largely from professional humorists and curated joke collections, providing a better a priori expectation of their quality and use of standard language. (3) The dataset has seen use even outside the original shared task (e.g., Mikhalkova and Karyakin, 2017; Cai et al., 2018; Poliak et al., 2018). (4) The jokes have been pre-classified according to their type (homographic puns, heterographic puns, and non-puns), so our extension of it could serve the needs of future qualitative research into humour.

The original dataset consists of 4030 short texts averaging about 11 words in length. Of the texts, 3398 contain humour (mostly, but not exclusively, punning jokes) and 632 do not (proverbs and aphorisms). Our examination of the data revealed three duplicate instances in the humour class; to preserve the size of the dataset, we replaced these with three new punning jokes provided to us by the dataset's original compilers. We applied humorousness annotations using a crowdsourcing setup. First, we randomly paired the texts such that each text appeared in exactly 14 unique pairs. Each of these 28,210 unique pairs was then presented to five annotators who were asked to judge which text (if either) was funnier. Annotators were recruited from American users of the Amazon Mechanical Turk crowdsourcing platform and paid at a rate commensurate with the US federal minimum wage.

To generate gold-standard scores, we apply BWS to the complete dataset. To evaluate whether the number of annotations is sufficient to produce a reliable gold standard, we randomly subsampled the annotations to produce subsamples with one to four annotators per pair. We then computed Spearman's rank correlation coefficient, ρ, between the gold-standard ranking and BWS scores computed for each subsample. The results averaged over ten random repeats (see Table 3) show that the rankings are very similar even when fewer annotators label each pair. We also computed the mean interannotator agreement (Krippendorff's α) across instances. The result, 0.80, indicates a satisfactory level of agreement among the crowd workers (Artstein and Poesio, 2008). Taken together, these results suggest that five annotators per pair is more than sufficient to reach a consensus ranking using BWS.

# annotators	1	2	3	4
Spearman's ρ	0.81	0.92	0.97	0.99

Table 3: Agreement measures for the humour dataset.

	humour	metaphor
# instances	4,030	15,181
# unique pairs	28,210	65,323
# unique pairs for each instance	14	(avg) 8.6
annotations/pair	5	(avg) 1.55

Table 4: Statistics for the humour and metaphor novelty datasets.

Metaphor Novelty Dataset. We use the metaphor novelty dataset of Do Dinh et al. (2018), which contains novelty scores for metaphors (i.e., metaphoric tokens) from the VU Amsterdam Metaphor Corpus (Steen et al., 2010) across four genres: news, fiction, conversation transcripts, and academic texts. The metaphors were compared by crowd workers using best–worst scaling tuples of four randomly chosen metaphors – that is to say, annotators were presented with random selections of four sentences with the metaphoric tokens highlighted, and they selected the most novel and most conventionalised metaphors from this set. The tuples were chosen such that each metaphor appeared in six different comparisons, and each comparison was labelled by three annotators.

For the new tasks proposed in this paper, we extract from each of these four-tuples, for each annotator, the pair comparing the most novel to the most conventionalised metaphor token in context. Since we create only those pairs containing the most and least novel instances in each tuple, each tuple generates only one pairwise comparison per worker. Because not all pairs are unique, and different pairs were extracted for different annotators, the number of unique pairs decreases, and the number of annotations per unique pair is less than three. We also use the gold standard provided by Do Dinh et al. (2018), which was obtained by applying BWS to the complete dataset.

Table 4 presents some statistics on the humour and metaphor novelty datasets.

4 Task Definitions

We introduce tasks to evaluate models for ranking instances by humorousness and metaphor novelty given pairwise comparisons. For the humorousness dataset, an instance is represented by a short text (typically 1–2 sentences) that possibly forms a joke. For the metaphor novelty dataset, an instance is represented by a metaphoric token and its sentential context. The tasks are designed to test the following hypotheses regarding our proposed Bayesian approach, GPPL, and other ranking models proposed in future: (a) given a sufficient number of pairwise labels, the proposed model converges close to the gold standard; (b) the proposed model is able to generalise to unseen instances using a combination of embeddings and linguistic features; (c) with a sparser set of pairwise training labels, the proposed model can exploit feature data to produce more accurate predictions than BWS; and (d) obtaining the same number of annotations for each pair to mitigate annotator disagreement is less effective than randomly choosing pairs to be annotated. To test these hypotheses, we devise a number of tasks that can be tested on both datasets.

Task 1: Test (a) the convergence of the proposed model to the gold standard. First, train the model on all available annotations without using any feature data – that is, learn a ranking from pairwise comparisons only. Using this model, estimate scores for all instances and rank the instances according to these scores. Compare this ranking to the gold BWS ranking using Spearman's rank correlation coefficient (ρ).

Task 2: Evaluate (b) the predictive ability of the proposed model. Randomly select 60% of the instances as a training set. Train the model on only those annotations that compare instances in the training set, then predict scores for instances in the test set (20%). Rank the test instances according to those scores and evaluate the ranking against BWS gold using ρ.

Task 3: Test (c) predictions for test instances when annotation data is sparse. Subsample the training set from Task 2 by randomly selecting 5%, 10%, 20%, 33%, and 66% of the original training annotations. To test hypothesis (d), we compare two subsampling methods: *annotation* subsampling (choose a random subset of pairwise annotations) and *pair* subsampling (first choose unique random pairs of instances, then take all annotations associated with those pairs). Pair sampling ensures that all selected pairs have multiple annotations from different annotators, which may help to mitigate noise, while annotation subsampling provides a more diverse coverage of possible pairs of instances. For each subsample, train the model and rank the instances in the test set. Evaluate against the gold-standard ranking using ρ.

Task 4: Test (c) the estimated scores for training instances when the pairwise annotation data is sparse. Repeat the same setup as Task 3, but evaluate the rankings for instances in the training set. This allows us to evaluate how many annotations are required to reliably rank a set of instances with each scoring method and subsampling method (d).

5 Experiments

5.1 Experimental Setup

We use the tasks defined in the previous section to evaluate the suitability of our proposed Bayesian approach, GPPL. For both datasets, the GPPL model is tested with 300-dimensional average word embeddings, using the word2vec model trained on Google News (Mikolov et al., 2013). For the metaphor task, the embedding for the token used metaphorically is concatenated with the average word embeddings that represent the subsuming context sentence.

For Task 2 on both datasets, we augment the average word embeddings with linguistic features: average token frequency (taken from a 2017 Wikipedia dump), a polysemy measure represented by the average number of synsets (taken from WordNet 3.0), and average bigram frequency (taken from Google Books Ngrams). Again for the metaphor task, we additionally append the metaphor token frequency if the frequency feature is selected. We repeat Task 2 with different subsets of these features to determine the most effective combination. The token frequency feature has previously been shown to distinguish between metaphoric and literal use (Beigman Klebanov et al., 2014), but also to be indicative of metaphor novelty (Do Dinh et al., 2018). By incorporating the polysemy feature we seek to increase performance especially for the funniness dataset, which includes many puns. The bigram feature reinforces the frequency feature by highlighting instances that include rare bigrams.

For best–worst scaling, we use the implementation provided by Kiritchenko and Mohammad (2016). We use the GPPL implementation provided by Simpson and Gurevych (2018). To ensure a reasonable computation time, we follow the authors' recommendations for hyperparameters and set the number of inducing points to $M = 500$ and the length-scales using the median heuristic. In future work, it may be possible to tune these hyperparameters further; however, M is a trade-off

instances	humour	metaphor
all	0.917	0.736
no tied BWS scores	0.951	0.737

Table 5: Task 1. Spearman's ρ between GPPL and gold-standard scores produced by BWS when trained without features.

between computation time and model accuracy, as the training time scales with $O(M^3)$ computational cost. With our current setup, the combined training and prediction time was approximately 2 hours for the metaphor novelty dataset and 2.5 hours for the funniness dataset running on a 24-core cluster with 2 GHz CPU cores.

5.2 Results

Task 1. We compare the BWS gold-standard ranking to the GPPL ranking produced when trained on all available pairwise annotations. We ignore feature data, representing instances solely by an ID instead of a feature vector. This is feasible because we train and test on the same instances, and so do not need features to generalise from training to test instances.

The resulting correlations are shown in the first line of Table 5. While the rankings for the humorousness dataset have high correlation, there is still some discrepancy for metaphor novelty. We note that the BWS scoring method means that multiple instances receive the same scores, while GPPL assigns unique values to all instances. To investigate whether these ties affect the rank correlations, we computed new rankings without ties by randomly sampling one instance for each tie, then computing Spearman's ρ for the subsampled instances. The mean over ten subsamples is shown in the second row of Table 5. For the humorousness dataset, the correlation increases when ties are excluded, suggesting that ties contribute to the difference between the BWS and GPPL rankings. The differences caused by tied BWS scores do not indicate errors but show a small difference due to the nature of BWS and GPPL scores.

However, for metaphor novelty, the difference when tied scores are removed is negligible. Instead, the lower correlation compared to the humour dataset hints at the more uneven annotation of the metaphors – that is, there are many very conventionalised instances, so each one was chosen less frequently as the least novel instance in a four-tuple, whereas the smaller number of novel metaphors means that each one is selected multiple times as the most novel instance in a four-tuple. This results in few pairs containing the highly-conventionalised instances, which introduces noise into the BWS and GPPL rankings. In contrast to the humour dataset, which is roughly balanced between funny and non-funny texts, the metaphor dataset is much more skewed towards one class, the conventionalised metaphors.

Unlike GPPL, the BWS score for a given instance does not take into account the scores of the instances that it was compared against. We investigate this effect by computing, for each instance s, the total rank c_s of instances compared against s, where c_s is the sum of GPPL ranks of instances that were annotated as funnier or more novel than s, minus the sum of ranks of instances that were annotated as less funny or novel than s. We then compute correlations between c_s and the difference in ranking between GPPL and BWS, obtaining both Spearman's ρ and Pearson's $r = 0.21$ for the humorousness dataset, and ρ and $r = 0.22$ for metaphor novelty. This indicates that the choice of instances to compare against contributed to the difference between GPPL and BWS rankings: the GPPL score for an instance is estimated relative to the scores of instances that it was compared against, while BWS scores are not. This difference may be greater for the metaphor dataset, since there are fewer pairs per instance and hence potentially noisier rankings.

The distributions of differences between rankings are shown in Figure 1, showing that the majority of differences are small for both datasets. This indicates that our proposed GPPL model can capture the gold-standard ranking adequately given a sufficient amount of pairwise training data.

For the humour dataset, we also used the original classifications from Miller et al. (2017) to evaluate how well the BWS and GPPL rankings separate non-pun instances from puns using the area under the receiver operating characteristic curve (AUROC; Fawcett, 2006). This area represents the probability that a randomly chosen pun will be ranked higher than a randomly chosen non-pun. Note, however, that some non-puns may contain other types of humour, so we do not expect to achieve a perfect score. We find that both BWS and GPPL achieve AUROC = 0.8, which reflects a good separation of the two classes.

Task 2. The results for predicting unseen instances in Task 2 are shown in Table 6. For both datasets, the combination of word2vec embeddings

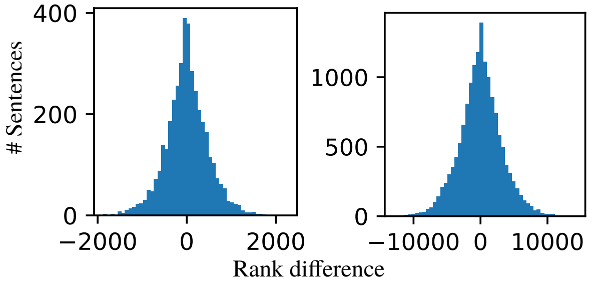

Figure 1: Task 1. Distribution of rank differences between BWS and GPPL scores for humorousness (left) and metaphor novelty (right).

features	humour	metaphor
w2v	0.531	0.551
w2v, freq., polysemy	0.552	0.540
w2v, freq., bigrams	**0.561**	**0.562**
w2v, polysemy, bigrams	0.537	0.523
w2v, freq., polysemy, bigrams	0.542	0.516

Table 6: Task 2. Predicting rankings on unseen test instances: Spearman's ρ against BWS gold standard ($p \lll 0.01$).

(*w2v*), average token frequency (*freq.*), and average bigram frequency performs best. Additionally including the polysemy feature generally decreased performance for the metaphor novelty dataset, but improved performance on the funniness dataset when compared to the word2vec-only experiment. The improvement due to token and bigram frequency suggests that the average word embeddings do not capture all word-level information.

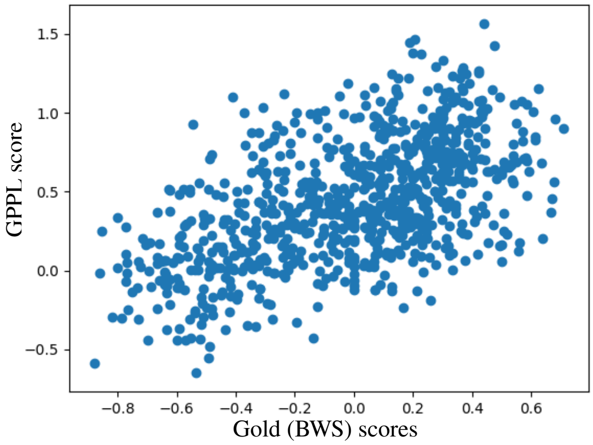

Figure 2: Gold vs. GPPL scores for the best Task 2 model for humour.

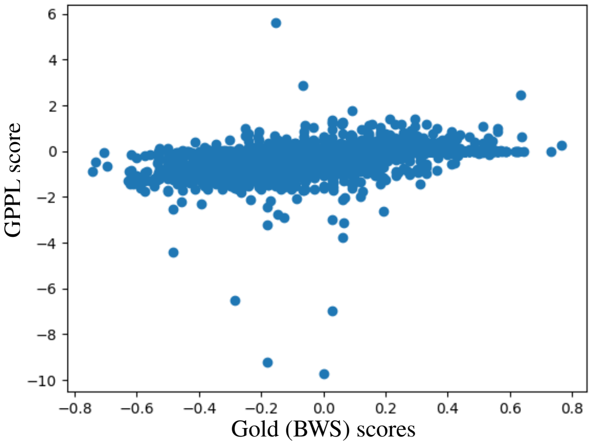

Figure 3: Gold vs. GPPL scores for the best Task 2 model for metaphor novelty.

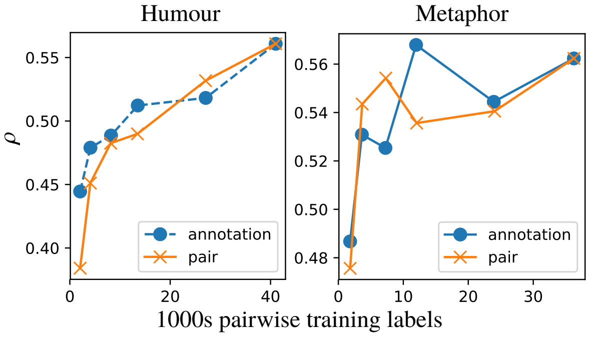

Figure 4: Task 3. Spearman's ρ for rank prediction on test instances (subsampled by *pair* or by *annotation*) with decreasing data sparsity ($p \lll 0.01$).

We compare the scores produced by BWS and GPPL for the best feature combination in Figures 2 and 3. In the metaphor novelty dataset, the GPPL scores are contained mainly in the range −2 to 2, with a few extreme outliers. In contrast, the BWS scores are all between −0.8 and 0.8. The ten largest outliers include two occurrences each of the metaphor tokens "fit" and "let", which are both rated correctly as highly conventionalised (e.g., in the sentence "How many times must I tell you that if you *let* things go too far, nobody can stop what will undoubtedly happen?"). The extreme outliers for GPPL scores are, however, not present in the humorousness dataset. In GPPL, the scores reflect confidence: the larger number of pairwise annotations in the metaphor dataset may increase the range of scores; smaller values may also correspond to noisier or more contradicting annotations.

Task 3. Figure 4 shows the results of Task 3, with the rightmost points corresponding to the Task 2 results. The results show that GPPL handles smaller training set sizes down to 5% with a much smaller decrease in performance compared to BWS. The *annotation* sampling strategy appears to be

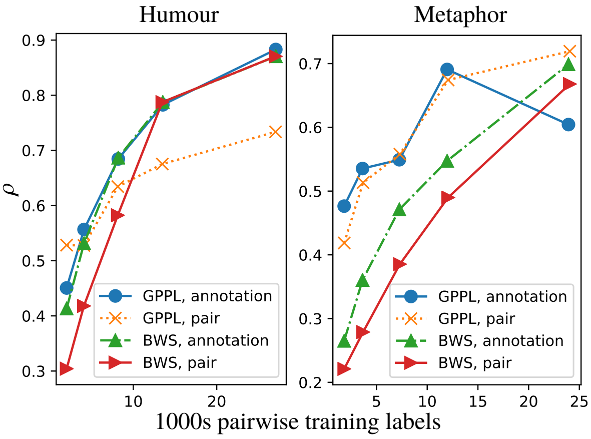

Figure 5: Task 4. Spearman's ρ for rank prediction on training instances (subsampled by *pair* or by *annotation*) with decreasing data sparsity ($p \lll 0.01$).

beneficial when data is sparse: it provides a greater diversity of pairs, so may provide better coverage over the set of instances, and therefore the feature space.

Task 4. In Figure 5, we show the results for Task 4, comparing GPPL against BWS for instances in the training set. Gold-standard rankings were not used in training, and the ranks were inferred by BWS and GPPL from the pairwise labels; hence, reducing the amount of pairwise data available reduces the quality of the rankings. For GPPL, we see that the ranking performance with sparse data is substantially higher than BWS. This is particularly notable for metaphor novelty, while for funniness, using the *annotation* strategy, the performance of BWS converges to that of GPPL as the dataset is increased. While GPPL performance with the *pair* strategy is highest with the small training set size for humour, it falls below that of BWS as the dataset increases. The results further suggest that the *annotation* strategy is preferable, which may inform future crowdsourcing efforts, and that while GPPL performs best with small training data, there are situations where BWS may have an advantage.

6 Conclusion

This paper has introduced new tasks for evaluating the degree of humorousness of a short text and the novelty of a metaphor within a short text. For humorousness, we have provided a new set of crowdsourced pairwise comparisons, while for metaphor novelty we extracted pairwise labels from existing best–worst scaling data. We have introduced a Bayesian approach, Gaussian process preference learning, that can use sparse pairwise annotations to estimate humorousness or novelty scores given word embeddings and linguistic features. Our experiments showed that GPPL outperforms BWS at ranking instances in the training set when few pairwise labels are available, and generalises well to ranking test instances that were not compared in the training set.

Given that our model achieves good results with rudimentary, task-agnostic linguistic features, in future work we plan to investigate the use of humour- and metaphor-specific features, including some of those used in past work (see §2) as well as those inspired by the prevailing linguistic theories of humour (Attardo, 1994) and metaphor (Black, 1955; Lakoff and Johnson, 1980). The benefits of including word and bigram frequency also point to possible further improvements using *n*-grams, tf–idf, or other task-agnostic linguistic features. Finally, we plan to further extend and use the humour dataset to investigate open questions on the linguistics of humour, such as what relationships hold between a pun's phonology and its "successfulness" or humorousness (Lagerquist, 1980; Hempelmann and Miller, 2017).

Acknowledgments

This work has been supported by the German Federal Ministry of Education and Research (BMBF) under the promotional reference 01UG1816B (CEDIFOR), by the German Research Foundation (DFG) as part of the QA-EduInf project (grants GU 798/18-1 and RI 803/12-1), by the DFG-funded research training group "Adaptive Preparation of Information from Heterogeneous Sources" (AIPHES; GRK 1994/1), and by the Austrian Science Fund (FWF) under project M 2625-N31. The Austrian Research Institute for Artificial Intelligence is supported by the Austrian Federal Ministry for Science, Research and Economy.

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Salvatore Attardo. 1994. *Linguistic Theories of Humor*. Mouton de Gruyter, Berlin.

Daniel Beck, Trevor Cohn, and Lucia Specia. 2014. Joint emotion analysis via multi-task Gaussian processes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1798–1803. 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. Association for Computational Linguistics.

Nancy D. Bell. 2017. Failed humor. In Salvatore Attardo, editor, *The Routledge Handbook of Language and Humor*, Routledge Handbooks in Linguistics, pages 356–370. Routledge, New York.

Max Black. 1955. Metaphor. *Proceedings of the Aristotelian Society*, 55(1):273–294.

Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Yitao Cai, Yin Li, and Xiaojun Wan. 2018. Sense-aware neural models for pun location in texts. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 546–551. Association for Computational Linguistics.

Hugo Carretero-Dios, Cristino Pérez, and Gualberto Buela-Casal. 2010. Assessing the appreciation of the content and structure of humor: Construction of a new scale. *Humor: International Journal of Humor Research*, 23(3):307–325.

Santiago Castro, Luis Chiruzzo, and Aiala Rosá. 2018a. Overview of the HAHA task: Humor analysis based on human annotation at IberEval 2018. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages*, volume 2150 of *CEUR Workshop Proceedings*, pages 187–194. Spanish Society for Natural Language Processing.

Santiago Castro, Luis Chiruzzo, Aiala Rosá, Diego Garat, and Guillermo Moncecchi. 2018b. A crowd-annotated Spanish corpus for humor analysis. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 7–11. Association for Computational Linguistics.

Peng-Yu Chen and Von-Wun Soo. 2018. Humor recognition using deep learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 2, pages 113–117. Association for Computational Linguistics.

Wei Chu and Zoubin Ghahramani. 2005. Preference learning with Gaussian processes. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 137–144. ACM.

Trevor Cohn and Lucia Specia. 2013. Modelling annotator bias with multi-task Gaussian processes: An application to machine translation quality estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 32–42. Association for Computational Linguistics.

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

Jonathan Dunn. 2010. Gradient semantic intuitions of metaphoric expressions. *Metaphor and Symbol*, 26(1):53–67.

Jonathan Dunn. 2014. Measuring metaphoricity. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 745–751. Association for Computational Linguistics.

Tom Fawcett. 2006. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874.

Paul Felt, Eric K. Ringger, and Kevin D. Seppi. 2016. Semantic annotation aggregation with conditional crowdsourcing models and word embeddings. In *Proceedings of the 26th International Conference on Computational Linguistics*, pages 1787–1796.

Jenny Rose Finkel, Christopher D. Manning, and Andrew Y. Ng. 2006. Solving the problem of cascading errors: Approximate Bayesian inference for linguistic annotation pipelines. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 618–626. Association for Computational Linguistics.

Adam Finn and Jordan J. Louviere. 1992. Determining the appropriate response to evidence of public concern: The case of food safety. *Journal of Public Policy & Marketing*, 11(2):12–25.

Terry N. Flynn and A. A. J. Marley. 2014. Best–worst scaling: Theory and methods. In Stephane Hess and Andrew Daly, editors, *Handbook of Choice Modelling*, pages 178–201. Edward Elgar Publishing, Cheltenham, UK.

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. Association for Computational Linguistics.

Christian F. Hempelmann. 2003. *Paronomasic Puns: Target Recoverability Towards Automatic Generation*. Ph.D. thesis, Purdue University, West Lafayette, IN, USA.

Christian F. Hempelmann. 2008. Computational humor: Beyond the pun? In Victor Raskin, editor, *The Primer of Humor Research*, number 8 in Humor Research, pages 333–360. Mouton de Gruyter, Berlin.

Christian F. Hempelmann and Tristan Miller. 2017. Puns: Taxonomy and phonology. In Salvatore Attardo, editor, *The Routledge Handbook of Language and Humor*, Routledge Handbooks in Linguistics, pages 95–108. Routledge, New York.

Matthew D. Hoffman, David M. Blei, Chong Wang, and John William Paisley. 2013. Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303–1347.

Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 133–142. ACM.

Maurice George Kendall. 1948. *Rank Correlation Methods*. Griffin, Oxford, UK.

Hiroyuki Kido and Keishi Okamoto. 2017. A Bayesian approach to argument-based reasoning for attack estimation. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 249–255. International Joint Conferences on Artificial Intelligence.

David C. Kingsley and Thomas C. Brown. 2010. Preference uncertainty, preference refinement and paired comparison experiments. *Land Economics*, 86(3):530–544.

Svetlana Kiritchenko and Saif M. Mohammad. 2016. Capturing reliable fine-grained sentiment associations by crowdsourcing and best–worst scaling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 811–817. Association for Computational Linguistics.

Svetlana Kiritchenko and Saif M. Mohammad. 2017. Best–worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 465–470. Association for Computational Linguistics.

Linnea M. Lagerquist. 1980. Linguistic evidence from paronomasia. In *Papers from the Sixteenth Regional Meeting Chicago Linguistic Society*, pages 185–191. University of Chicago.

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

Vasileios Lampos, Nikolaos Aletras, Daniel Preoţiuc-Pietro, and Trevor Cohn. 2014. Predicting and characterising user impact on Twitter. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 405–413. Association for Computational Linguistics.

Louis G. Lippman and Mara L. Dunn. 2000. Contextual connections within puns: Effects on perceived humor and memory. *Journal of General Psychology*, 127(2):185–197.

R. Duncan Luce. 1959. On the possible psychophysical laws. *Psychological Review*, 66(2):81–95.

Anthony A. J. Marley and Jordan J. Louviere. 2005. Some probabilistic models of best, worst, and best–worst choices. *Journal of Mathematical Psychology*, 49(6):464–480.

Rada Mihalcea and Carlo Strapparava. 2006. Learning to laugh (automatically): Computational models for humor recognition. *Computational Intelligence*, 22(2):126–142.

Rada Mihalcea, Carlo Strapparava, and Stephen Pulman. 2010. Computational models for incongruity detection in humour. In *Computational Linguistics and Intelligent Text Processing: 11th International Conference, Cicling 2010*, number 6008 in Theoretical Computer Science and General Issues, pages 364–374, Berlin/Heidelberg. Springer.

Elena Mikhalkova and Yuri Karyakin. 2017. Detecting intentional lexical ambiguity in English puns. In *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue" (2017)*, volume 1, pages 167–178. HSE – Higher School of Economics National Research University.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, volume 2, pages 3111–3119.

Tristan Miller, Christian F. Hempelmann, and Iryna Gurevych. 2017. SemEval-2017 Task 7: Detection and interpretation of English puns. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, pages 58–68. Association for Computational Linguistics.

Frederick Mosteller. 1951. Remarks on the method of paired comparisons: I. The least squares solution assuming equal standard deviations and equal correlations. *Psychometrika*, 16(1):3–9.

Hannes Nickisch and Carl Edward Rasmussen. 2008. Approximations for binary Gaussian process classification. *Journal of Machine Learning Research*, 9:2035–2078.

Seth Ovadia. 2004. Ratings and rankings: Reconsidering the structure of values and their measurement. *International Journal of Social Research Methodology*, 7(5):403–414.

Nathalie Parde and Rodney D. Nielsen. 2018. A corpus of metaphor novelty scores for syntactically-related word pairs. In *Proceedings of the 11th International*

Conference on Language Resources and Evaluation, pages 1535–1540. European Language Resources Association.

R. L. Plackett. 1975. The analysis of permutations. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 24(2):193–202.

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. 2018. Collecting diverse natural language inference problems for sentence representation evaluation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 337–340. Association for Computational Linguistics.

Peter Potash, Alexey Romanov, and Anna Rumshisky. 2017. SemEval-2017 Task 6: #HashtagWars: Learning a sense of humor. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, pages 49–57. Association for Computational Linguistics.

Amruta Purandare and Diane Litman. 2006. Humor: Prosody analysis and automatic recognition for *Friends*. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 208–215. Association for Computational Linguistics.

Dragomir Radev, Amanda Stent, Joel Tetreault, Aasish Pappu, Aikaterini Iliakopoulou, Agustin Chanfreau, Paloma de Juan, Jordi Vallmitjana, Alejandro Jaimes, Rahul Jha, and Robert Mankoff. 2016. Humor in collective discourse: Unsupervised funniness detection in the *New Yorker* cartoon caption contest. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*. European Language Resources Association.

Victor Raskin. 1985. *Semantic Mechanisms of Humor*, volume 24 of *Synthese Language Library: Texts and Studies in Linguistics and Philosophy*. D. Reidel Publishing, Dordrecht, Netherlands.

Carl E. Rasmussen and Christopher K. I. Williams. 2006. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, USA.

Dafna Shahaf, Eric Horvitz, and Robert Mankoff. 2015. Inside jokes: Identifying humorous cartoon captions. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1065–1074. ACM.

Ekaterina Shutova. 2010. Models of metaphor in NLP. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 688–697. Association for Computational Linguistics.

Ekaterina Shutova. 2015. Design and evaluation of metaphor processing systems. *Computational Linguistics*, 41(4):579–623.

Edwin Simpson and Iryna Gurevych. 2018. Finding convincing arguments using scalable Bayesian preference learning. *Transactions of the Association for Computational Linguistics*, 6:357–371.

Edwin D. Simpson, Matteo Venanzi, Steven Reece, Pushmeet Kohli, John Guiver, Stephen J. Roberts, and Nicholas R. Jennings. 2015. Language understanding in the wild: Combining crowdsourcing and machine learning. In *Proceedings of the 24th International Conference on World Wide Web*, pages 992–1002. International World Wide Web Conferences Steering Committee.

Jonas Sjöbergh and Kenji Araki. 2007. Recognizing humor without recognizing meaning. In *Applications of Fuzzy Sets Theory: 7th International Workshop on Fuzzy Logic and Applications, WILF 2007, Camogli, Italy, July 7–10, 2007. Proceedings*, number 4578 in Lecture Notes in Artificial Intelligence, pages 469–476, Berlin/Heidelberg. Springer.

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*, volume 14 of *Converging Evidence in Language and Communication Research*. John Benjamins Publishing, Amsterdam.

Louis L. Thurstone. 1927. A law of comparative judgment. *Psychological Review*, 34(4):273–286.

Ivan Titov and Alexandre Klementiev. 2012. A Bayesian approach to unsupervised semantic role induction. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 12–22. Association for Computational Linguistics.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 248–258. Association for Computational Linguistics.

Hui Yuan Xiong, Yoseph Barash, and Brendan J. Frey. 2011. Bayesian prediction of tissue-regulated splicing using RNA sequence and cellular context. *Bioinformatics*, 27(18):2554–2562.

Diyi Yang, Alon Lavie, Chris Dyer, and Eduard Hovy. 2015. Humor recognition and humor anchor extraction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2367–2376. Association for Computational Linguistics.

Yi-Hsuan Yang and Homer H. Chen. 2011. Ranking-based emotion recognition for music organization and retrieval. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):762–774.

Georgios N. Yannakakis and John Hallam. 2011. Ranking vs. preference: A comparative study of self-reporting. In *Affective Computing and Intelligent Interaction: 4th International Conference, ACII 2011,* *Memphis, TN, USA, October 9–12, 2011, Proceedings, Part I*, volume 6974 of *Lecture Notes in Computer Science*, pages 437–446, Berlin/Heidelberg. Springer.

Renxian Zhang and Naishi Liu. 2014. Recognizing humor on Twitter. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 889–898. ACM.