Embedding Lexical Features via Tensor Decomposition for Small Sample Humor Recognition

Zhenjie Zhao[1*] **Andrew Cattle**[1*] **Evangelos E. Papalexakis**[2†] **Xiaojuan Ma**[1]
[1] Department of Computer Science & Engineering, HKUST
[2] Department of Computer Science & Engineering, University of California, Riverside
{zzhaoao, acattle}@connect.ust.hk, epapalex@cs.ucr.edu, mxj@cse.ust.hk

Abstract

We propose a novel tensor embedding method that can effectively extract lexical features for humor recognition. Specifically, we use word-word co-occurrence to encode the contextual content of documents, and then decompose the tensor to get corresponding vector representations. We show that this simple method can capture features of lexical humor effectively for continuous humor recognition. In particular, we achieve a distance of 0.887 on a global humor ranking task, comparable to the top performing systems from SemEval 2017 Task 6B (Potash et al., 2017) but without the need for any external training corpus. In addition, we further show that this approach is also beneficial for small sample humor recognition tasks through a semi-supervised label propagation procedure, which achieves about 0.7 accuracy on the *16000 One-Liners* (Mihalcea and Strapparava, 2005) and *Pun of the Day* (Yang et al., 2015) humour classification datasets using only 10% of known labels.

1 Introduction

Recognizing humor automatically is an important step for natural human-computer interaction (Shahaf et al., 2015). While early works tend to frame humor recognition as a binary classification task (Mihalcea and Strapparava, 2005; Yang et al., 2015), the last few years have seen the emergence of humor recognition as a pairwise relative ranking task (Cattle and Ma, 2016; Shahaf et al., 2015). In addition to pairwise ranking, SemEval 2017 Task 6 also includes a global ranking subtask. However, the majority of submissions build global rankings using a series of pairwise comparisons (Potash et al., 2017). Only Yan and Pedersen (2017) attempt to predict global rankings directly, ranking documents inversely to their probability according to an n-gram language model.

State-of-the-art humor recognition algorithms usually require a considerable amount of training data with labels to learn effective features (Yang et al., 2015). However, such data are difficult to obtain – especially fine-grained humor annotations. First, the humor judgments differ from individual to individual. Thus, collecting perceptually consistent human labels is expensive and time-consuming. Second, fine-grained degrees of humor add a further challenge. Therefore, methods on small sample learning or even unsupervised rule-based methods merit investigation.

In this paper, considering the importance of lexical information for humor recognition (Radev et al., 2015), we propose a tensor decomposition method to capture the contextual nuances of a corpus. This allows us to model the lexical similarity of sentences regardless of the size of the corpus. In this way, we can rank the degree of humor effectively via lexical centrality (Radev et al., 2015), namely, regarding the distance to the lexical center as an indicator of the degree of humor. Experimental results on the SemEval 2017 Task 6 dataset (Potash et al., 2017) show that *without external training data*, the tensor embedding method can achieve performance equivalent to the second place on SemEval 2017 Task 6B without the need for any external training corpus. In addition, by applying a semi-supervised label propagation procedure (Zhou et al., 2003), we can also use the tensor embedding method for small sample humor recognition, achieving about 0.7 accuracy with only 10% of known labels on the *16000 One-Liners* (Mihalcea and Strapparava, 2005) and *Pun of the Day* (Yang et al., 2015) datasets.

[*] Zhenjie Zhao and Andrew Cattle contributed equally to this work.

[†] E. Papalexakis was supported by a UCR-China collaboration grant by the Bourns College of Engineering at UCR, and by the National Science Foundation CDSE Grant no. OAC-1808591

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 6376–6381, Hong Kong, China, November 3–7, 2019.

The contributions of this paper are: 1) we propose a tensor embedding method to model the lexical features of documents, which can capture lexical similarity effectively regardless of the size of the corpus, 2) we show that the lexical features can be used effectively for fine-grained humor ranking and small sample humor recognition. Our implementation is open-sourced, and can be found at https://github.com/zhaozj89/TensorEmbeddingNLP.

2 Related Work

2.1 Humor Feature Extraction

Modeling and learning humor features are critical for automatic humor recognition. Previous works tend to use a combination of phonological, stylistic, semantic, and content-based features. Phonological features include acoustic features extracted from sitcom audio tracks (Bertero and Fung, 2016) and "phonetic embeddings" generated using a character-to-phoneme LSTM encoder-decoder (Donahue et al., 2017). Stylistic features include alliteration, rhyming, negative sentiment, and adult slang (Mihalcea and Strapparava, 2005) as well as emotional scenarios (Reyes et al., 2012). Semantic features range from attempts to measure incongruity (Cattle and Ma, 2018; Shahaf et al., 2015; Yang et al., 2015) to the use of word embeddings as inputs to neural models (Bertero and Fung, 2016; Donahue et al., 2017). Content-based approaches include word frequency (Mihalcea and Strapparava, 2005), n-gram probability (Yan and Pedersen, 2017), and lexical centrality (Radev et al., 2015).

Centrality is based on the observation that humorous responses to common stimuli tend to cluster around a small number of core jokes (Radev et al., 2015; Shahaf et al., 2015), with more central documents benefiting from "wisdom of the crowd". While most humor features involve making population-level inferences based on document-level features, centrality is instead population-level feature directly. Radev et al. (2015) calculate their centrality feature using LexRank, a graph-based text summarization method (Erkan and Radev, 2004). Compared with more traditional lexical similarity measures like tf-idf, this method is better suited to short humor texts due to their short lengths leading to sparse vector representations (Erkan and Radev, 2004).

2.2 Small Sample Humor Recognition

Once the humor features have been extracted, the next step is training a machine learning model to make predictions. Although learning-based methods have shown significant performance improvement recently (Yang et al., 2015), one of their main bottlenecks is the lack of appropriate training corpora. While previous works have employed data crawled from websites (Mihalcea and Strapparava, 2005; Yang et al., 2015), Twitter (Cattle and Ma, 2016; Reyes et al., 2012), sitcom subtitles (Bertero and Fung, 2016; Purandare and Litman, 2006), or the New Yorker Cartoon Caption Contest (Radev et al., 2015; Shahaf et al., 2015), these datasets are generally not released publicly. Owing to the difficulty in obtaining fine-grained labeled humor data, it is critical to study how to recognize humor by a small training sample or even without labeled data.

3 Method

3.1 Tensor Embedding

Contextual patterns of words can be used to measure lexical similarity for humor recognition. State-of-the-art learning-based approaches like doc2vec (Le and Mikolov, 2014) or sent2vec (Pagliardini et al., 2018) usually require a large amount of data. This is difficult to obtain for humor recognition. We propose to use a novel tensor decomposition method to obtain lexical features of short humor texts. To capture lexical similarity for humor recognition, we propose to represent the tensor through a novel word-word co-occurrence method, which has only been explored in the context of fake news detection (Hosseinimotlagh and Papalexakis, 2018). Considering a corpus $\mathcal{D} = \{s_1, s_2, \ldots, s_D\}$ with D sentences, we first build a vocabulary for it, namely, $w_1, w_2, \ldots, w_V$, where V is the number of words. For each sentence s in $\mathcal{D}$, we count the word-word co-occurrence in a small window H, and build a frequency matrix $\mathbf{W}_s \in \mathbb{Z}^{V \times V}$, where $\mathbb{Z}$ denotes the set of integers. In particular, $\mathbf{W}_s(i, j)$ indicates the frequency that word w_i and w_j co-occur in s within the window H. In this way, we can capture the lexical patterns of s in $\mathbf{W}_s$. We then stack all $\mathbf{W}_s$ as a three-dimensional tensor $\boldsymbol{\mathcal{W}} \in \mathbb{Z}^{V \times V \times D}$. The objective of tensor decomposition is to find an approximation $\hat{\boldsymbol{\mathcal{W}}}$ of $\boldsymbol{\mathcal{W}}$ so

	Prompt						Average
	Christmas	**Shakespeare**	**Bad Job**	**Break Up**	**Broadway**	**Cereal**	
Our system	0.909	1.000	0.909	0.909	0.727	0.909	0.887 (±0.113)
Duluth	0.818	0.909	1.000	0.636	1.000	0.909	0.872 (±0.137)
TakeLab	0.909	0.909	1.000	0.818	1.000	0.818	0.908 (±0.081)
QUB	0.818	0.909	0.818	1.000	1.000	0.909	0.924 (±0.081)
SVNIT	0.818	1.000	0.909	1.000	1.000	0.818	0.938 (±0.089)

Table 1: The results of our lexical centrality system using tensor embeddings along with the top four SemEval 2017 Task 6B systems reproduced from Potash et al. (2017). Prompt names refer to specific prompts from the SemEval 2017 Task 6 evaluation set.

	Pun of the Day				**16000 One-Liners**			
	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
5%	0.670	0.707	0.578	0.633	0.665	0.690	0.602	0.642
10%	0.700	0.735	0.611	0.667	0.681	0.704	0.624	0.661
30%	0.729	0.752	0.680	0.714	0.700	0.722	0.641	0.679
90%	0.745	0.752	0.723	0.737	0.705	0.721	0.667	0.693
Yang et al. (2015)	0.7958	0.761	0.862	0.808	0.798	0.801	0.793	0.797

Table 2: The results of our label propagation system. XX% refers to XX% of the data used as training and (100-XX)% as test. Baseline results reproduced from Yang et al. (2015)

that:

$$\hat{\mathcal{W}} = \sum_{r=1}^{R} \mathbf{v}_r \otimes \mathbf{v}_r \otimes \mathbf{d}_r, \quad (1)$$

where $\mathbf{v}_r \in \mathbb{R}^V$, $\mathbf{d}_r \in \mathbb{R}^D$, R is the pre-defined rank parameter, and $\otimes$ is the outer product, namely, $\mathbf{v}_r \otimes \mathbf{v}_r \otimes \mathbf{d}_r$ being a three-dimensional tensor, and

$$\mathbf{v}_r \otimes \mathbf{v}_r \otimes \mathbf{d}_r(i,j,k) = \mathbf{v}_r(i) \cdot \mathbf{v}_r(j) \cdot \mathbf{d}_r(k). \quad (2)$$

With the tensor decomposition, we can find low-rank embeddings of sentences that capture the similarity of contextual patterns (Hosseinimotlagh and Papalexakis, 2018). In particular, $\mathbf{C} = [\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_R] \in \mathbb{R}^{D \times R}$, where the s-row of $\mathbf{C}$ is the embedding vector of sentence s. The Euclidean distance of embeddings is used to measure the similarity of two sentences.

3.2 Lexical Centrality

We use lexical centrality to rank the degree of humor (Radev et al., 2015). While Radev et al. (2015) utilize a graph-based definition of centrality, we instead take a vector-space approach. Given the decomposed $\mathbf{C} = [\mathbf{c}_1^T, \mathbf{c}_2^T, \ldots, \mathbf{c}_D^T]^T$, we compute a centroid as the average $\mathbf{m}$ of all sentence vectors of a corpus:

$$\mathbf{m} = \frac{1}{D} \sum_{k=1}^{D} \mathbf{c}_k. \quad (3)$$

The Euclidean distance to the center is then taken as an indicator of the degree of humor. In other words, given two sentences s_1 and s_2 and their embeddings $\mathbf{x}_1$ and $\mathbf{x}_2$, $d(\mathbf{m}, \mathbf{x1}) < d(\mathbf{m}, \mathbf{x2})$ implies s_1 is funnier than s_2.

3.3 Label Propagation

With the lexical similarity captured by tensor embeddings, we can build a similarity graph, and use a label propagation algorithm (Zhou et al., 2003) for semi-supervised humor recognition. In this way, we can use only a small portion of labeled data to predict the remaining unlabeled data effectively (Zhou et al., 2003). In particular, with the tensor embeddings, we first find the K nearest neighbors of each data point, and build a similarity graph $\mathcal{G}$. We then form an affinity matrix W, where $W_{ij} = 1$ if i and j are connected, otherwise, $W_{ij} = 0$. Afterwards, we iterate:

$$F(t+1) = \alpha W F(t) + (1-\alpha) Y, \quad (4)$$

and can get the results F^* as the limit of this sequence. Equation (4) means we propagate the labels of each data point to its neighbors in a weighted average way, where α is the ratio of propagating labels each iteration. For each point x_i, its label is $y_i = \arg\max_{j \le c} F^*_{ij}$.

4 Experiment

To evaluate the effectiveness of the tensor embedding method, we conduct two experiments on global humor ranking and binary humor classification separately. The alternating least squares

method of CANDECOMP/PARAFAC tensor decomposition (Sidiropoulos et al., 2017) is used to calculate the low-rank sentence embeddings as implemented in the Matlab tensor toolbox [1].

4.1 Global Humor Ranking

To show the effectiveness of the lexical centrality of our tensor embedding method, we conduct an experiment on SemEval 2017 Task 6B (Potash et al., 2017) consisting of tweeted responses to specific thematic prompts generated as part of a TV show. For each prompt, the writing staff of the show pick a top 10 and an overall winner. These humor judgments are used as gold standard labels.

Tensor embeddings and centroids are computed on a per-prompt basis and responses are ranked according to their distance from the centroid. We run a grid search procedure to determine the optimal rank value as 100, the window size as 5. For evaluation, we adopt the same edit distance-based metric used in Potash et al. (2017).

The results of our lexical centrality system using tensor embeddings is shown in Table 1, where the official results of other state-of-the-art systems are taken from Potash et al. (2017). Our system outperforms all but the Duluth (Yan and Pedersen, 2017) system in the official results for SemEval 2017 Task 6B (Potash et al., 2017), making our performance equivalent to second place. It is notable that our system can perform well on the **Broadway** prompt, where other methods usually fail. Moreover, because our system does not have a learning procedure, the performance is more stable than others.

4.2 Binary Humor Classification

To show the effectiveness of label propagation of our tensor embedding method for small sample humor recognition, we conduct an experiment on two humor classification datasets *16000 One-Liners* (Mihalcea and Strapparava, 2005) and *Pun of the Day* (Yang et al., 2015). Similarly, we run a grid search procedure to find optimal parameters, and set the rank as 10, window size as 5, neighbor number as 50, α as 0.2. $F(0)$ is set as a zero matrix initially. For each dataset, we randomly select 5%, 10%, 30%, and 90% of the data for training. We run a 10-fold procedure, and report the average accuracy, precision, recall, and F1 score values.

The results of humor classification are shown in Table 2. Our own implementation of Yang et al. (2015) is included as a baseline. While Yang et al. (2015) uses a large portion of data for training and combine different features, we find that at similar portion of training data (90%), the results of our method are comparable to it. In addition, with only a small portion of training data, our method still achieves good results.

4.3 Discussion

4.3.1 Lexical Centrality

The most notable aspect of our tensor embedding/lexical centrality approach is how little training data our system requires. Our system's unsupervised nature means that we do not need to use the 106 training prompts included with the SemEval 2017 Task 6 dataset. Our results are obtained exclusively using the six evaluation prompts. By comparison, almost all the systems reported in Potash et al. (2017) take a supervised approach and make full use of the training set. Furthermore, since we consider prompts one-at-a-time and since each prompt only contains approximately 100 responses, we are able to achieve a state-of-the-art performance with 100 training documents. The only system reported in Potash et al. (2017) to take an unsupervised approach is Duluth (Yan and Pedersen, 2017). Like ours, their results are obtained without using the training set. However, their system uses an n-gram language model trained on a 6.2GB subset of the News Commentary Corpus and the News Crawl Corpus. Similarly, most other systems use some form of external training corpora for training word embeddings, phoneme models, semantic models, and so on.

Another advantage of our approach is the ease of interpretability, in contrast to neural-based state-of-the-art baselines. Because our lexical feature is in an Euclidean space, we can compare and rank humor level more easily. Tweets labeled as "overall winners" exhibited a smaller mean distance from their respective centroids (0.848) than those labeled as "merely in the top 10" (0.942). These tweets then in turn exhibited smaller distances than those labeled as "not in the top 10" (1.00). A one-way ANOVA test gives mild evidence that overall winners are drawn from a different distribution than tweets not in the top 10 ($p = 0.106$). This slight result is likely due to the fuzzy nature of humor and the relatively small

[1] www.tensortoolbox.org

dataset. Finally, ad hoc analysis of tweets with distances > 2 revealed these to be mostly "not in the top 10".

4.3.2 Label Propagation

Although the semi-supervised framework provides a good alternative for small sample humor recognition, our method still cannot achieve a state-of-the-art performance with the same portion of training data. There is still space to improve the method; for example, by modeling not only the lexical similarity, but also other features, such as word association (Cattle and Ma, 2016), and the like, that are important for humor recognition. In addition, label propagation cannot handle unbalanced data well. Adding prior knowledge of the ratio of labels, e.g., the unbalanced SemEval 2017 Task 6 dataset, also deserves further investigation.

5 Conclusion

In this paper, we show the importance of lexical features for small sample humor recognition. We propose a tensor embedding method to capture the lexical similarity effectively. Without training data, on SemEval 2017 Task 6B, we can achieve a relatively good result. Under a semi-supervised framework, the tensor embedding method can also achieve pretty good results for small sample humor classification. It is interesting to further investigate a unified tensor embedding model to combine not only lexical, but also other features that are important for the sense of humor.

References

Dario Bertero and Pascale Fung. 2016. A long short-term memory framework for predicting humor in dialogues. In *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016-Proceedings of the Conference*, pages 130–135.

Andrew Cattle and Xiaojuan Ma. 2016. Effects of semantic relatedness between setups and punchlines in twitter hashtag games. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 70–79, Osaka, Japan. The COLING 2016 Organizing Committee.

Andrew Cattle and Xiaojuan Ma. 2018. Recognizing humour using word associations and humour anchor extraction. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1849–1858, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

David Donahue, Alexey Romanov, and Anna Rumshisky. 2017. HumorHawk at SemEval-2017 task 6: Mixing meaning and sound for humor recognition. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 98–102, Vancouver, Canada. Association for Computational Linguistics.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

Seyedmehdi Hosseinimotlagh and Evangelos E Papalexakis. 2018. Unsupervised content-based identification of fake news articles with tensor decomposition ensembles. In *WSDM 2018 Workshop on Misinformation and Misbehavior Mining on the Web (MIS2)*.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages II–1188–II–1196. JMLR.org.

Rada Mihalcea and Carlo Strapparava. 2005. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 531–538. Association for Computational Linguistics.

Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2018. Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features. In *NAACL 2018 - Conference of the North American Chapter of the 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 (SemEval-2017)*, pages 49–57, Vancouver, Canada. Association for Computational Linguistics.

Amruta Purandare and Diane Litman. 2006. Humor: Prosody analysis and automatic recognition for F*R*I*E*N*D*S*. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 208–215. Association for Computational Linguistics.

Dragomir R. Radev, Amanda Stent, Joel R. Tetreault, Aasish Pappu, Aikaterini Iliakopoulou, Agustin Chanfreau, Paloma de Juan, Jordi Vallmitjana, Alejandro Jaimes, Rahul Jha, and Robert Mankoff. 2015. Humor in collective discourse: Unsupervised funniness detection in the new yorker cartoon caption contest. *CoRR*, abs/1506.08126.

Antonio Reyes, Paolo Rosso, and Davide Buscaldi. 2012. From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering*, 74:1–12.

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*, KDD '15, pages 1065–1074, New York, NY, USA. ACM.

Nicholas D. Sidiropoulos, Lieven De Lathauwer, Xiao Fu, Kejun Huang, Evangelos E. Papalexakis, and Christos Faloutsos. 2017. Tensor decomposition for signal processing and machine learning. *IEEE Transactions on Signal Processing*, 65(13):3551–3582.

Xinru Yan and Ted Pedersen. 2017. Duluth at semeval-2017 task 6: Language models in humor detection. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 384–388, Vancouver, Canada. Association for Computational Linguistics.

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, Lisbon, Portugal. Association for Computational Linguistics.

Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. 2003. Learning with local and global consistency. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, NIPS'03, pages 321–328, Cambridge, MA, USA. MIT Press.