Improving Humor Detection by Converting Word Embeddings to Feature Probabilities

Ren Tu
University of California, Berkeley
Berkeley, CA
rt97@berkeley.edu

Abstract

In this paper, we reinterpret an alternative representation of word embeddings and apply it through new feature engineering to improve the humor classification task on a new dataset. The new representation of word embeddings relies on the idea that each value of a word embedding vector is a function of the probability of possessing that dimension's feature. When these embeddings are transformed into feature probabilities, we explore additional tools at our disposal such as K-L divergence and cross-entropy to represent incongruity and ambiguity features among words that are useful in the humor detection task. We demonstrate that the feature probability representations do contain some incremental humor signals and provide some insight on where linguistic-feature-based approaches excel and where they are still lacking in humor recognition.

1: Introduction

Humor is an advanced level of communication that provokes laughter, builds social bonds, and benefits the mental and physical health of humans. Having computers better understand humor could lead to many societal benefits as humans have ever-increasing levels of human-machine interactions.

However, detecting humor is a challenging task for computers. Humor can be interpreted differently in varying contexts. There are many different types of humor such as irony, sarcasm, wordplay and metaphors. Oftentimes humor requires knowledge of contextual subtleties or external cultural factors that break away from typical language rules. These factors make it difficult to detect humor using a systematic rules-based approach.

In this project, our goal is to treat humor detection as a classification task. We explore features that are potential semantic indicators of humor including incongruity, ambiguity, polarity and subjectivity, with a particular focus on whether there are additional signals to be extracted from word embeddings by converting them to feature probabilities. We want to examine whether linguistic features using feature probabilities can improve classification performance compared to features used in existing literature.

2: Background

Many existing studies on humor recognition have focused on using linguistic features to build classifiers. For example, Yang et al. (2015) explored several semantic structures behind humor: incongruity, ambiguity, interpersonal effect, phonetic style, and humor anchors. Incongruity refers to the opposition or contradiction of words that violates expectations, thus creating humor. Ambiguity theory states that the usage of words with multiple definitions can be an important part of humorous sentences. Interpersonal effect is related to how strong sentiment or opinion in words can elicit laughter. Phonetic style pertains to the clever usage of word sounds such as alliteration or rhymes to create humor. Humor anchors are defined as the minimal combinations of word spans in a sentence that create humor. Liu et al. (2018) built on the work of Yang et al. (2015) by adding sentiment association features for classification. By parsing sentences into discourse units, humor may be detected if there are types of sentiment conflict or sentiment transition between discourse units.

Other studies have used deep learning to tackle the humor detection task. Xie et al. (2021) focused on the incongruity theory of humor by exploring a set-up versus punchline framework for humor and using GPT-2 for predictions. Chen et al. (2018) focused on using convolutional neural networks on GloVe word embeddings to detect humor without directly creating linguistic features.

Besides the previous literature directly tackling humor recognition, Bhat et al. (2020) explored a new way to interpret word embeddings by transforming them into feature probabilities of each individual embedding vector dimension. This transformation converts each word vector into a fuzzy set to allow for set theoretic operations such as union, difference and intersection and similarity calculations such as cross-entropy and K-L divergence that the authors argue are more asymmetric notions of similarity that are closer to human interpretations of language similarity.

This project seeks to improve humor classification performance by taking some of the linguistic features commonly used in existing humor detection literature as baselines and creating new linguistic features using the feature probability concepts introduced by Bhat et al. (2020).

3: Methods

Given that humor recognition is a classification task, many researchers create features that represent qualities of humor. In this section, we explore features associated with three theories of humor: interpersonal effect, incongruity and ambiguity. For features that use Word2Vec vectors, we chose to create skip-gram negative sample Word2Vec vectors like those used by Bhat et al. (2020) since we want to reflect the ideas of feature probability tuples expressed by those authors. Due to computational resource constraints, we chose to use the Gensim Text8 Wikipedia corpus as the training text for the Word2Vec vectors, a smaller corpus relative to other full-size Wikipedia datasets.

3.1 Interpersonal Effect

Interpersonal effect refers to how strong sentiment or opinion can produce humor in a text. According to Yang et al. (2015), a text is more likely to be humorous with high-sentiment words such as ‘idiot’ in the following humor example:

Your village called. They want their idiot back.

We will use the methodologies employed by Yang et al. (2015) and Liu et al. (2018) to create the following interpersonal effect features:

- High and Low Polarity: Count of occurrences of highly positive and negative words respectively in a text. This captures the among of high-sentiment words in either direction.
- High and Low Subjectivity: Count of occurrences of high and low subjectivity words respectively in a text. This captures strongly and weakly opinionated words.

3.2 Incongruity – Word2Vec

Incongruity theory states that laughter can come from inconsistency in words that lead to communication that fall outside of expectations. We will use the feature engineering techniques employed by Yang et al. (2015) and Liu et al. (2018) to create the following incongruity features:

- Maximum Word Pair Similarity: Highest cosine similarity between possible pairs of Word2Vec word vectors in a text. This is a proxy for the lack of incongruity in a text.

- Minimum Word Pair Similarity: Lowest cosine similarity between possible pairs of Word2Vec word vectors in a text. This is a proxy for how much incongruity exists.

3.3 Ambiguity – WordNet

Ambiguity can produce humor as a result of multiple meanings of words producing unexpected interpretations. We plan to capture ambiguity through avenues previously explored by Yang et al. (2015), Liu et al. (2018) and Xie et al. (2021):

- Word Sense Total Combination: Calculated as the log of the product of all possible senses of each word in a text according to WordNet. This provides us the magnitude of possible meaning combinations between words.
- Maximum Sense Similarity: The highest path similarity of any word sense pair in a text according to WordNet.
- Minimum Sense Similarity: The lowest path similarity of any word sense pair in a text according to WordNet.

3.4 Transformation into Feature Probabilities

The underlying idea of feature probabilities introduced by Bhat et al. (2020) is that each value of a word embedding vector is a function of the probability of possessing that dimension's feature. To convert word embeddings into feature probabilities, the dimensional values of each word vector are first exponentiated and then normalized with the individual word vectors. Then the transformed values are normalized relative to the same dimensional values across the entire vocabulary, thereby producing a tuple of dimensional feature probabilities of each word relative to the vocabulary.

3.5 Incongruity – Feature Probabilities

With words represented by feature probabilities, K-L divergence can be a sensible metric to explore as it is a measure of similarity between two probability distributions. Bhat et al. (2020) argued that since K-L divergence is an asymmetric measure of similarity, it can more closely approximate human language understanding of similarity that also tends to be asymmetric. We plan to capture incongruity through word feature probabilities using the following features:

- Maximum K-L Divergence: The highest K-L divergence between any pairs of word feature probabilities in a text.
- Minimum K-L Divergence: The lowest K-L divergence between any pairs of word feature probabilities in a text.

3.6 Ambiguity – Feature Probabilities

With feature probabilities available, another avenue to explore ambiguity is through cross-entropy, which is a measure of uncertainty between probability distributions. Bhat et al. (2020) argues that cross-entropy can represent polysemy as words with high cross-entropy have a more distributed set of feature probabilities, which capture the potential multiple meanings and senses among words. We believe the following features may contain additional signals on ambiguity:

- Maximum Cross-Entropy: The highest cross-entropy between any pairs of word feature probabilities in a text
- Minimum Cross-Entropy: The lowest cross-entropy between any pairs of word feature probabilities in a text
- Top 5 Cross-Entropy: The sum of the five highest cross-entropy values between any pairs of word feature probabilities in a text

3.7: Dataset

This project will use the ColBERT dataset used in Annamoradnejad et al. (2021) that contains 200,000 labeled short texts with balanced classes of 100,000 humor and 100,000 non-humor. The short texts range between 30-100 characters and 10-22 words. Due to computational resource constraints, this project will randomly choose half of the ColBERT dataset for analysis (50,000 humor and 50,000 non-humor short texts). We believe this is a sufficiently large size to produce robust results.

4: Results and Discussion

In this section, we illustrate our experimental design using features discussed in the previous section and interpret the classification results.

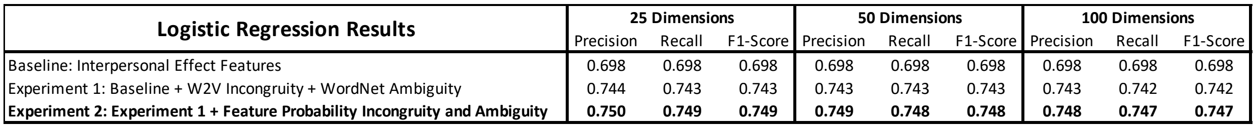
4.1: Baseline and Experiments

Since our project is focused on improvements using feature probabilities, our baseline will be the interpersonal effect features that have no associated feature probability analogue. The first level experiment will add the existing literature incongruity and ambiguity features to the baseline to see the performance difference of adding those features. Then the second level experiment will have all existing features plus the new incongruity and ambiguity variables utilizing feature probabilities to see if there is additional room for humor recognition improvement. Also, given that Bhat et al. (2020) discovered that feature probabilities tend to decrease in effectiveness with increasing original vector dimensions, each experiment will have three iterations using 25-dimension, 50-dimension, and 100-dimension Word2Vec vectors to test for dimensional effects.

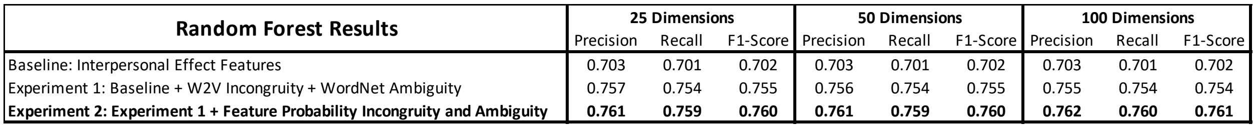
4.2: Modeling Methods and Metrics

For each experiment iteration, we will implement humor classification through logistic regression and random forest. These two methods can effectively capture linear and non-linear relationships respectively between the features and the humor target variable. Every model will be run through a 5-fold cross validation to increase the robustness of the results. Our key performance metrics will be the weighted precision, recall and F1-scores as we want to correctly classify as much of both humor and non-humor cases as possible.

4.3: Results



Logistic Regression Results	**25 Dimensions**			**50 Dimensions**			**100 Dimensions**		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Baseline: Interpersonal Effect Features	0.698	0.698	0.698	0.698	0.698	0.698	0.698	0.698	0.698
Experiment 1: Baseline + W2V Incongruity + WordNet Ambiguity	0.744	0.743	0.743	0.743	0.743	0.743	0.743	0.742	0.742
Experiment 2: Experiment 1 + Feature Probability Incongruity and Ambiguity	**0.750**	**0.749**	**0.749**	**0.749**	**0.748**	**0.748**	**0.748**	**0.747**	**0.747**



Random Forest Results	**25 Dimensions**			**50 Dimensions**			**100 Dimensions**		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Baseline: Interpersonal Effect Features	0.703	0.701	0.702	0.703	0.701	0.702	0.703	0.701	0.702
Experiment 1: Baseline + W2V Incongruity + WordNet Ambiguity	0.757	0.754	0.755	0.756	0.754	0.755	0.755	0.754	0.754
Experiment 2: Experiment 1 + Feature Probability Incongruity and Ambiguity	**0.761**	**0.759**	**0.760**	**0.761**	**0.759**	**0.760**	**0.762**	**0.760**	**0.761**

Sample of true positives across all models for Experiment 2:

- I wish I had emo hair so it would cut itself
- What’s the difference between Trump and Bush? Trump builds towers and Bush knocks them down
- My friends keep telling me to stop impersonating butter. But I can’t, I’m on a roll now

- I’m no gynecologist, but I can take a look
- What did the police do when they wanted to interrogate Mark? Question Mark.

Sample of false positives across all models for Experiment 2:

- Exxon supports a carbon tax except when there is a vote on a carbon tax
- About a 50/50 chance a computer will be doing your job in 20 years
- North Korea is a bad actor, but not a state sponsor of terrorism
- I hope this week’s graduates acquired what I did at college: lifelong friendships
- Alabama governor: if we can put a man on the moon, we can build more prisons

Sample of true negatives across all models for Experiment 2:

- Is vitamin D a good idea for diabetics?
- Trump attempts another campaign reset after declaring he won’t ‘pivot’
- New York City mayor declares November 19 ‘Angie Martinez Day’
- Claim: American Apparel’s ousted CEO allegedly misused funds
- Digging below the surface of membership growth at credit unions

Sample of false negatives across all models for Experiment 2:

- Who built King Arthur’s round table? Sir Cumference
- Latest election news: Donald Trump narrowly leads Hillary Clinton by 4 lies
- Canadian territory puns? Yukon be serious! I’m having Nunavut
- Support bacteria, it’s the only culture some people have
- Did you hear about the constipated Chancellor of the Exchequer? He couldn’t budge-it!

4.4: Discussion

From the model results, the feature probability variables in Experiment 2 did improve performance to a small extent across metrics, dimensions and model types over the existing literature features in Experiment 1. This shows that the tuples of feature probabilities converted from Word2Vec embeddings do contain some incremental signal to help detect humor. We hypothesize that the incremental improvements are only small due to the overlap in signals with the non-feature-probability incongruity and ambiguity features. The similar incremental performance across dimensions also suggests that increasing word embedding dimensionality does not necessarily decrease the effectiveness of feature probabilities. The better performance across metrics for random forest models for Experiments 1 and 2 suggest that the ambiguity and incongruity features have some non-linear properties to help identify humor.

When we examine the sample of classification results from our best models in Experiment 2, we can better understand where the models excelled and where they were lacking. From the list of true positives, the features capture incongruity and ambiguity in humor well with multiple-meaning and opposite-meaning word combinations such as emo/cut, butter/roll, build/knock and interrogate/question/mark. From the false positive examples, the models have a difficult time with texts that have humor-like setups but are actually just factual statements. The false negatives are also telling in that our models cannot fully capture clever word-play and word combinations requiring deep contextual understanding such as round/Sir Cumference, Trump/Clinton/lies, Yukon/Nunavut and constipated/budge-it.

5: Conclusion

In this paper, we reinterpreted a novel approach to word embeddings using feature probabilities to create new features that contained incremental signals in the task of humor recognition and classification. We quantitatively demonstrated that incongruity and ambiguity representations of word combinations using the K-L divergence and cross-entropy of word feature probabilities improved the precision, recall and f1-scores of classification models relative to ones only using existing literature features on interpersonal effect, incongruity and ambiguity. Through this work, we confirmed the validity of the work of Bhat et al. (2020) on word feature probabilities and demonstrated its potential to further improve language tasks including humor detection. We also recognize the limitations of our approach using solely linguistic features as more comprehensive and sophisticated representations of context can likely further improve the humor recognition task.

References

I. Annamoradnejad and G. Zoghi, “ColBERT: Using BERT Sentence Embedding for Humor Detection”, *arXiv preprint arXiv*: 2004.12765, 2021
https://arxiv.org/pdf/2004.12765.pdf

S. Bhat, A. Debnath, S. Bannerjee and M. Shrivastava, “Word Embeddings as Tuples of Feature Probabilities”, in *Proceedings of the 5th Workshop on Representation Learning for NLP (RepL4NLP-2020)*, 2020, pages 24–33
https://aclanthology.org/2020.repl4nlp-1.4.pdf

P.-Y. Chen and V.-W. Soo, “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 (Short Papers)*, 2018, pages 113–117
https://aclanthology.org/N18-2018.pdf

L. Liu, D. Zhang and W. Song, “Modeling Sentiment Association in Discourse for Humor Recognition”, in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers)*, 2018, pages 586–591
https://aclanthology.org/P18-2093.pdf

Y. Xie, J. Li and P. Pu, “Uncertainty and Surprisal Jointly Deliver the Punchline: Exploiting Incongruity-Based Features for Humor Recognition,” in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Short Papers)*, 2021, pages 33–39
https://aclanthology.org/2021.acl-short.6.pdf

D. Yang, A. Lavie, C. Dyer and E. Hovy, “Humor Recognition and Humor Anchor Extraction,” in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pages 2367–2376
https://www.cc.gatech.edu/~dyang888/docs/emnlp_yang_16.pdf