Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/109/20241303

A Review of Humor Recognition Research

Xiang Zhou

School of Computer and Science and Technology, Dalian University of Technology, Dalian, China

zhouxiang@mail.dlut.edu.cn

Abstract. Humor recognition is a popular research area in Natural Language Processing (NLP), with the fundamental goal of detecting whether humor is present. With the rapid development of artificial intelligence and the growing demand for human-computer interaction, humor recognition has broad applications in areas such as social media analysis, human-computer interaction systems, and intelligent question-answering assistants. However, due to the highly subjective and complex nature of humor, teaching computers to understand and recognize humor in a manner similar to humans is an exceptionally challenging task. This paper reviews the major research progress in the field of humor recognition and the construction of humor corpora. It discusses humor recognition methods based on traditional machine learning, deep learning, and multimodal approaches, and then highlights the advantages and disadvantages of each. Based on these analyses, the paper identifies the main challenges and difficulties in current research and provides suggestions and prospects for the future development of humor recognition systems.

Keywords: humor recognition, deep learning, multimodality, humor dataset.

1. Introduction

Humor is widely used in daily communication as a way for speakers to create a relaxed and pleasant conversational atmosphere. It is a complex emotional phenomenon, often conveyed through wordplay, sarcasm, absurdity, or paradox at the linguistic level. Humor can also rely on context and implications, using innuendo, metaphor, or unexpected endings to provoke laughter. These complexities make computational humor recognition difficult to achieve. Additionally, humor is closely tied to the speaker's sociocultural background, with significant differences in how humor is expressed across different regions, cultures, and languages. This cultural dependency not only makes humor a complex form of emotional expression but also a contextualized linguistic phenomenon. Globally, humor may be interpreted differently across cultural boundaries. For instance, British humor is often characterized by dry wit, frequently using sarcasm and irony, while American humor tends to be more direct and vivid, and Chinese humor often combines metaphors or puns. These cross-cultural differences present a significant challenge to computational humor recognition. As such, humor recognition is considered one of the most difficult tasks in Natural Language Processing (NLP).

Early work on humor detection employed manually designed humor features, using models such as Naive Bayes and Support Vector Machines (SVM) to detect single-line jokes [1]. Their model achieved an accuracy rate of 97% in distinguishing single-line jokes from news articles. Chen and Soo proposed a humor recognition model that combines Convolutional Neural Networks (CNN) with Highway

DOI: 10.54254/2755-2721/109/20241303

Network [2]. To enhance efficiency, they reduced the network scale, addressing the challenges posed by increased network depth during model training. Hasan et al. [3] performed punchline detection on the multimodal humor dataset UR-FUNNY, extracting features from different modalities and proposing a Contextual Memory Fusion Network (C-MFN) model, which extracts multimodal contextual humor information. Experimental results demonstrated that C-MFN achieved higher accuracy on UR-FUNNY than human performance, indicating that there is still significant room for improvement in multimodal humor recognition.

Humor computation is a relatively emerging field in NLP. Teaching computers to understand, recognize, and even generate humor has both academic significance in NLP research and many practical applications. As one of the more complex tasks in NLP, humor recognition involves multi-layered linguistic phenomena and contextual understanding, advancing the field in areas such as implicit sentiment analysis, ambiguity resolution, and context analysis. Computer-based humor understanding can make human-computer interaction more natural, allowing intelligent assistants and social robots to interact with users in a more friendly manner, avoiding mechanical or cold responses. Humor computation can also be applied in social media analysis and sentiment detection. By analyzing users' comments and uncovering hidden humorous emotions, it can help assess emotional tendencies and public opinion trends. Additionally, humor, due to its cultural dependency, poses a barrier to cross-cultural communication. Computational humor could offer more precise humor translations, promoting cross-cultural exchanges.

However, few studies have conducted a comprehensive review of humor recognition. This paper reviews the major research progress in text-based humor recognition, discussing methods based on traditional machine learning and deep learning. By analyzing the strengths and weaknesses of different approaches, this paper highlights the major challenges in current research and provides an outlook on future research directions.

2. Humor-Related Theories and Corpora

2.1. Humor-Related Theories

Early humor research primarily approached the subject from psychological and philosophical perspectives. The most mainstream humor theories include superiority theory, relief theory, and incongruity theory [4]. Superiority theory posits that the feeling of superiority over others is the source of humor. Although this theory explains the psychological reasons for laughter, it is not applicable to all scenarios. Relief theory suggests that humor serves as a form of escape from "prohibited thoughts," allowing the release of repressed tensions and thereby relieving stress. Incongruity theory, on the other hand, holds that humor arises from the contrast between expectation and reality when a sentence contains two different interpretations, with the greater the incongruity, the stronger the humor effect.

Building on incongruity theory, Raskin developed the first linguistic humor theory, the Semantic Script Theory of Humor (SSTH). SSTH defines the structure of a joke as composed of a "setup" and a "punchline," where the setup allows for both explicit and implicit interpretations. The punchline triggers the implicit meaning in the setup, creating an unexpected effect and thus generating humor. Attardo and Raskin further expanded on SSTH by proposing the General Theory of Verbal Humor (GTVH), which adds six essential elements for analyzing humorous texts: script opposition, logical mechanism, situation, target, narrative strategy, and language. In addition to these theories, humor feature extraction can also be based on linguistic characteristics such as rhyme, puns, ambiguity, and semantic distance.

2.2. Humor Corpora

In NLP research, corpora are crucial resources for model training and evaluation. For humor recognition, which is a complex task, a suitable corpus must not only cover various forms of humor but also provide appropriate annotations to accurately identify humorous content. However, the subjective and diverse nature of humor makes collecting high-quality corpora particularly challenging. Nevertheless, with the

DOI: 10.54254/2755-2721/109/20241303

rise of humor recognition research, several humor corpora have emerged in recent years, providing valuable resources for research in this field.

Yang et al.[5]constructed the "Pun of the Day" dataset, collecting over 2,000 positive samples from the pun website Pun of the Day, while negative samples were sourced from AP News, The New York Times, Yahoo! Answers, and proverbs. Due to domain differences between the positive and negative datasets, all selected negative samples were restricted to vocabulary present in the positive samples, with sentence lengths limited to 10-30 words.

To address the limitations in both the type and scale of humor datasets, Orion Weller and Kevin Seppi [6] collected over 550,000 jokes from the r/Jokes section on Reddit. The humor level of each joke was quantified based on user feedback from the r/Jokes community.

Badri N. Patro et al. [7] developed a multimodal humor dataset (MHD) based on the American sitcom The Big Bang Theory, which contains humorous dialogues from multiple characters. They labeled dialogue fragments as humorous or non-humorous by using the laugh track from live audience recordings as indirect manual annotations. Since humor is often expressed through contextual dialogue, Patro and colleagues grouped several lines of dialogue into sets and labeled them for humor. Each dialogue segment was also annotated with various attributes that could influence the humor model, such as scene, speaker, audience, participants, and start/end times.

3. Research Methods

3.1. Traditional Machine Learning-Based Approaches

In traditional machine learning, humor recognition is often based on linguistic features. After preprocessing the text, features and text representations are extracted. A classifier is then built and trained based on the test data to predict the classification results.

Yang et al.[5] focused on the underlying semantic structure of humor, constructing humor features from four aspects: inconsistency, ambiguity, interpersonal impact, and speech style. Inconsistency features are derived by measuring semantic disjunctions in a sentence, with Word2Vec used to extract two types of features: maximum semantic distance and minimum semantic distance, assessing the distance between content words in a sentence. Ambiguity is often a key component of humor [8], and to capture ambiguity in a sentence, a part-of-speech tagger [9] is used to identify nouns, verbs, adjectives, and adverbs. WordNet is then applied to obtain possible meanings of the words $\{\omega 1, \omega 2 \ldots \omega k\}$, and the combinations of meanings are calculated as a feature. To extract interpersonal impact features, the lexical resource from Wilson et al. is used [9], which provides annotations and clues to measure subjectivity and sentiment associated with words. This results in positive/negative polarity (the number of positive/negative words) and strong/weak subjectivity (the number of strong/weak subjective words). For speech style, alliteration and rhyme features are designed using the CMU Pronouncing Dictionary. After designing the humor features, Yang et al. used a random forest classifier for training, treating humor recognition as a traditional text classification task. Experimental results showed that combining handcrafted humor features with Word2Vec outperformed using Word2Vec or language models alone, as this approach considers both underlying structure and semantic meaning.

Traditional machine learning methods for humor recognition typically require less computational power and are suitable for small-scale data. The research focus is on designing appropriate humor features to achieve optimal performance. However, as the corpus size increases, efficiency and quality may decline to varying degrees. Because traditional methods rely on handcrafted features, they may exhibit poor generalization in different contexts. Moreover, humor recognition involves numerous factors, and simply relying on designed humor features cannot cover all scenarios. Traditional methods are limited to surface-level linguistic features of humor, raising the question of whether the models are truly capturing and understanding humorous expressions or merely confusing the writing style and vocabulary of jokes with actual humorous traits [10].

DOI: 10.54254/2755-2721/109/20241303

3.2. Deep Learning-Based Approaches

While traditional machine learning models are relatively intuitive and efficient, their limitations in humor recognition have been discussed above. Additionally, they face issues such as the high-dimensional sparsity of text representation and weak feature expression. Deep learning models (e.g., GloVe, BERT) can convert high-dimensional sparse bag-of-words models into low-dimensional dense word vector representations. These representations not only reduce dimensionality but also capture the semantic relationships between words. Unlike traditional methods that require manually designed humor features—a challenging task for handling complex emotional language—deep learning models, through end-to-end learning, can automatically extract high-level semantic features from large-scale data, reducing the need for manual feature extraction and enabling more accurate capture of the deeper semantics of humor.

Orion Weller and Kevin Seppi chose the pre-trained BERT model as the basis for their humor recognition system. BERT [11], based on the Transformer architecture, can effectively identify key words in a sentence and focus on critical points in the linguistic structure. Another key feature of BERT is its ability to fine-tune: by adding an additional output layer to the pre-trained model, it can adapt to other tasks. Weller attributes the success of this model largely to the self-attention layers in Transformers. Their experiment used two baseline models: a CNN with Highway Layers as described by Chen and Soo [2], and developed by Srivastava et al. [12], and human performance data from Amazon Mechanical Turk. They used the metrics including Accuracy, Precision, Recall and F1 score as a comparison. Accuracy measures overall correctness, Precision evaluates positive prediction quality, Recall assesses sensitivity to positive instances, and F1 Score balances Precision and Recall.

Table 1. Comparison of Methods on Pun of the Day Dataset.

Previous Work	Accuracy	Precision	Recall	F1
Word2Vec+HCF	0.797	0.776	0.836	0.705
CNN	0.867	0.880	0.859	0.869
CNN+F	0.892	0.886	0.907	0.896
CNN+HN	0.892	0.889	0.903	0.896
CNN+F+HN	0.894	0.866	**0.940**	0.901
Orion Weller et al.[11]	**0.930**	**0.930**	0.931	**0.931**

In Table 1, HCF represents Human Centric Features, F for increasing the number of filters, and HN for the use of highway layers in the model. It shows that on the "Pun of the Day" dataset, Transformer achieved an accuracy of 93%. Although the baseline CNN model employed various techniques to extract features from the dataset, its accuracy was still slightly lower than the Transformer.

Table 2. Results on Short Jokes Identification

Method	Accuracy	Precision	Recall	F1
CNN+F+HN	0.906	0.902	0.946	0.924
Transformer	**0.986**	0.986	0.986	0.986

As shown in Table 2, the accuracy and F1 score of the Transformer on the "Short Jokes" dataset were 0.986, an 8% improvement over the CNN.

3.3. Multimodal Approaches

With the rise of multimodal techniques in NLP, recent studies have begun exploring humor recognition through multimodal approaches. Humor often relies not just on textual information but also on the speaker's tone, actions, or facial expressions, which can serve as sources of humor. This approach, which integrates multiple modalities, more closely resembles how humans process humor. Multimodal information typically includes text, audio, and video, and the challenge lies in how to interact and fuse information across these different modalities to obtain multimodal representations.

DOI: 10.54254/2755-2721/109/20241303

Hasan et al. [3], after establishing the multimodal dataset UR-FUNNY, proposed the Contextual Memory Fusion Network (C-MFN), which consists of a contextual network (Figure 1) and a memory fusion network (Figure 2). The contextual network models each modality (text, audio, video) separately using three LSTMs, with each LSTM corresponding to one modality. The multimodal contextual network then learns the multimodal representation of the context based on the output of the unimodal contextual networks, using a self-attention model to discover asynchronous spatial-temporal relationships within the context. After learning the unimodal and multimodal representations of the context, the Memory Fusion Network (MFN) is used to model the punchline. MFN consists of two types of memory: unimodal and multimodal. The unimodal and multimodal outputs of the contextual network initialize the memory in the MFN. MFN operates at the word level, where each punchline word and its accompanying visual and acoustic descriptors are fed into an LSTM system. Multi-view gated memory is updated during each LSTM iteration through a Delta-memory attention network. The final prediction of humor is based on the last state of the LSTM system and multi-view gated memory, calculated through affine mapping and a Sigmoid activation function.

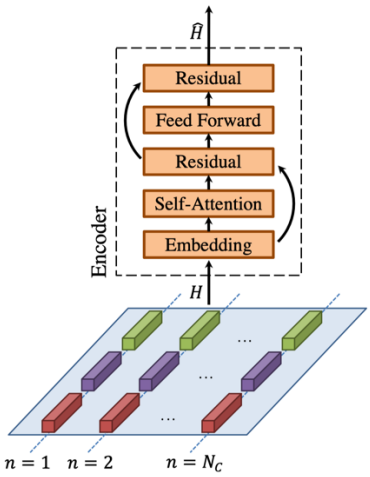

Figure 1. The structure of Multimodal Context Network[3].

Experimental results showed that each neural component of the C-MFN model contributed to improving humor prediction. The findings also demonstrated that modeling humor from a multimodal perspective can lead to successful results, with C-MFN achieving the highest accuracy of 65.23% across text, audio, and video modalities. However, this is still lower than human performance on the UR-FUNNY dataset (82.5%).

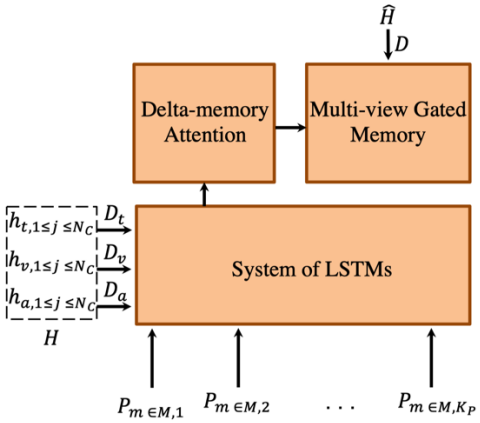

Figure 2. The initialization and recurrence process of Memory Fusion Network (MFN)[3].

DOI: 10.54254/2755-2721/109/20241303

4. Challenges and Future Prospects

Humor recognition is considered one of the more challenging tasks in NLP. In recent years, significant progress has been made thanks to advancements in deep learning and NLP research. Current large language models are capable of engaging in humorous conversations with users to some extent. However, humor recognition still faces several key challenges:

Diversity and Cultural Dependency: Humor's diversity and reliance on cultural contexts remain difficult obstacles for humor recognition. While deep learning can extract high-level semantic features, it lacks transparency compared to traditional machine learning methods. These deep models often operate as "black boxes," making it hard to explain how they arrive at their decisions.

Lack of Sufficient and Diverse Humor Corpora: Existing humor datasets, such as r/Jokes and "Pun of the Day," provide training resources for models. However, these datasets are often too narrow in scope, failing to encompass the wide variety of humor across different cultures and forms. Additionally, beyond English and Chinese, humor corpora in other languages remain scarce.

Early Stage of Multimodal Humor Recognition: Multimodal humor recognition research is still in its infancy. On one hand, it involves complex perception and reasoning processes, and the information density and expression forms across different modalities are not equivalent, potentially leading to inconsistent interpretations when integrating them. On the other hand, humor can arise from many modalities, and current multimodal datasets cannot cover all these cases. The robustness of multimodal humor recognition is also an area that needs further exploration.

In the future, researchers could explore the following directions: 1) Developing more universal, cross-cultural humor recognition models. 2) Using multimodal techniques to integrate various information, such as speech, facial expressions, and text, to improve the accuracy of humor recognition. 3) Enhancing models' understanding of the essence of humor. 4)Building larger, more diverse textual humor corpora and multimodal humor datasets to provide more comprehensive training data for models.

5. Conclusion

Humor recognition, as a complex task in natural language processing, is challenging due to its subjectivity, diversity, and cultural dependency. This paper reviewed the major research progress in text-based humor recognition, including traditional machine learning, deep learning, and multimodal approaches. Traditional machine learning methods can achieve some success on small-scale datasets through handcrafted features. However, as data volume increases and tasks become more complex, deep learning models—especially those based on pre-trained language models—show greater humor recognition capabilities. Moreover, using multimodal techniques to detect humor enables machines to gain a deeper understanding of humor.

However, challenges such as the lack of humor corpora, understanding the underlying principles of humor, and cross-modal integration still need to be addressed in current research. Future work should dive deeper into these challenges, developing more general and intelligent humor recognition systems capable of handling humor across different contexts.

References

[1] Mihalcea, R., & Strapparava, C. (2005). Making computers laugh: Investigations in automatic humor recognition. In R. J. Mooney (Ed.), Proceedings of the conference on human language technology and empirical methods in natural language processing (pp. 531–538). Association for Computational Linguistics. https://doi.org/10.3115/1220575.1220642

[2] Chen, L.-C., & Soo, V.-W. (2018). Humor recognition using deep learning. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2, 113–117. https://aclanthology.org/N18-2018

[3] Hasan, M. K., Ling, Y., Jones, C., & Sah, S. (2019). UR-FUNNY: A multimodal language dataset for understanding humor. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2046–2056. https://aclanthology.org/D19-1211

DOI: 10.54254/2755-2721/109/20241303

[4] Strapparava, C., Stock, O., & Mihalcea, R. (2011). Computational humour. Springer Berlin Heidelberg.
[5] Yang, D., Lavie, A., Dyer, C., & Hovy, E. H. (2015). Humor recognition and humor anchor extraction. Proceedings of the Conference on Empirical Methods in Natural Language Processing.
[6] Weller, O., & Seppi, K. (2020). The rJokes dataset: A large scale humor collection. In Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020) (pp. 6093–6098). European Language Resources Association (ELRA). https://aclanthology.org/2020.lrec-1.753
[7] Patro, B. N., Lunayach, M., Srivastava, D., Singh, H. S., & Namboodiri, V. P. (2021). Multimodal humor dataset: Predicting laughter tracks for sitcoms. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (pp. 576–585). IEEE.
[8] Miller, T., & Gurevych, I. (2015). Automatic disambiguation of English puns. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 719–729). Association for Computational Linguistics.
[9] Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing.
[10] West, R., & Horvitz, E. (2019). Reverse-engineering satire, or "paper on computational humor accepted despite making serious advances." Proceedings of the AAAI Conference on Artificial Intelligence, 33(1), 7265–7272. https://doi.org/10.1609/aaai.v33i01.33017265
[11] Weller, O., & Seppi, K. (2019). Humor detection: A transformer gets the last laugh. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 3621–3625. https://aclanthology.org/D19-1372
[12] Rupesh Kumar Srivastava, Klaus Greff, and J¨urgen Schmidhuber. 2015. Highway Networks. arXiv e-prints, page arXiv:1505.00387.