A Survey in Automatic Irony Processing: Linguistic, Cognitive, and Multi-X Perspectives

Qingcheng Zeng
Department of Linguistics
Northwestern University
qingchengzeng@outlook.com

An-Ran Li
Department of Chinese and Bilingual Studies
The Hong Kong Polytechnic University
an-ran.li@connect.polyu.hk

Abstract

Irony is a ubiquitous figurative language in daily communication. Previously, many researchers have approached irony from linguistic, cognitive science, and computational aspects. Recently, some progress have been witnessed in automatic irony processing due to the rapid development in deep neural models in natural language processing (NLP). In this paper, we will provide a comprehensive overview of computational irony, insights from linguistic theory and cognitive science, as well as its interactions with downstream NLP tasks and newly proposed multi-X irony processing perspectives.

1 Introduction

Irony, which generally refers to the expressions that have opposite literal meanings to real meanings, is a representative rhetorical device in human languages (Li and Huang, 2020). For example, in sentences I just love when you test my patience!!, and Had no sleep and have got school now, compared to their literal meanings, both sentences are conveying reverse meanings and emotions, which mean not love and not happy. People are using ironies to communicate their affective states more implicitly or explicitly, to strengthen the claims depending on the needs, which contribute to neologism across human languages. However, the inner incongruity makes ironic expressions harder for machines to understand. Consequently, accurate irony processing systems are essential and challenging for downstream tasks and natural language understanding (NLU) research.

Previously, many researchers approached irony processing from various perspectives. The last survey in computational irony was published seven years ago (Wallace, 2015) and mainly focused on irony detection and pragmatic context model. Therefore, a systematic and comprehensive survey on automatic irony processing remains absent, encouraging us to focus on a review of advances in irony processing, from traditional machine learning, recurrent neural networks (RNNs) methods, to deeper pretrained language models (PLMs) throughout. Besides, del Pilar Salas-Zárate et al. (2020) revealed the great imbalance in figurative language research, in which sarcasm was dominant and twice as much as irony research. With this review, We aim to encourage a balanced and equal research environment in figurative languages.

Moreover, the design and further improvement in present irony processing systems should be concatenated with both the theoretical accounts in irony theories and its role in communication, followed by downstream applications in NLP tasks. Generally, most researchers approached irony processing as a single irony detection branch and three main shared tasks (Ghanem et al., 2019; Van Hee et al., 2018a; Ortega-Bueno et al., 2019) all focused on irony detection, whereas other aspects are constantly under-explored. And theoretical-informed or cognitive-informed discussions are rarely seen in irony research compared to tiny amendments in neural networks’ architectures.

In this paper, we aim to offer a comprehensive review in irony processing from machine learning, linguistic theory, cognitive science, and newly proposed multi-X perspectives and to evaluate irony processing in NLP applications. The remaining sections are organized as follows. In the second part, we will approach irony theories in linguistics and cognitive science research, and discuss the concrete differences between irony and sarcasm. Then, we will review irony datasets in world languages and discuss potential problems in annotation schemes throughout. In the fourth part we will retrospect research progress in automatic irony processing, including traditional feature engineering, neural network architectures, to PLMs and relatively under-explored research fields in irony processing. Finally, we will discuss irony’s in-
teractions with downstream NLP tasks like sentiment analysis and opinion mining, and multi-X perspectives for further development in computational irony research.

2 Theoretical Research in Irony

2.1 Irony Theories

Various definitions have been given to irony. Early studies suggested that irony is the expression whose real meaning is contradictory to its literal meaning (Grice, 1975). The Merriam-Webster Dictionary, The Oxford English Dictionary, and The Collins English Dictionary all adopted this definition and used the words "opposite" or "contrary" to explain the relationship between the literal and contextual meanings of irony.

However, more research into various types of ironic examples revealed that the contextual meaning of irony does not have to be "opposite" or "contrary" to the literal one. According to Sperber and Wilson (1986); Wilson and Sperber (2012), some expressions have no "literal meaning" to be challenged because no "literal meaning" is mentioned in the context, based on which they raised relevance theory and the "echoic" concept. They considered irony as "an echoic use of language in which the speaker tacitly dissociates herself from an attributed utterance or thought" (Wilson, 2006). That is, if the "echoic use" is incongruous in some ways, the expression can be ironic. Based on this theory, Seto (1998) put forward that there are some "echo-markers" like definitely, really, and indeed.

Li and Huang (2020) provided instances to show that "incongruity" does not have to be between the literal and contextual meanings of irony in certain circumstances. They believed that irony’s true nature is a psychological activity as much as a verbal representation. The speaker or listeners must finish the "reversal" process on a psychological level for it to be completed. When compared to the concepts of "echoic" and "incongruity," "reversal" is concerned not only with the results but also with the psychological processes that the speakers/listeners go through.

2.1.1 Types of Irony

Booth (1974) divided irony into tragic and comic irony by literary genre, as well as stable and unstable irony by determinacy. He also categorized irony into dramatic irony, situational irony, verbal irony, and rhetorical irony by the range of context it has to refer to. Researchers in computational irony paid the most attention to situational irony and verbal irony. Situational irony is a circumstance in which the outcome differs from what was expected, or a situation that contains striking contrasts (Imagine a situation that a lifeguard is saved from drowning). Verbal irony is the expression in which the speaker’s intended meaning significantly different from the literal one. Abrams and Harpham (2014) considered that verbal irony usually involves the explicit presentation of one attitude or evaluation, but with signals that the speaker wants to express a totally different attitude or opinion. It differs from situational irony in that it is purposely manufactured by speakers. For example, Very well, keep insulting me!

Verbal irony is the kind to which computational linguistics researchers pay the most attention. Li (2021) made a further classification of them. She put forward eight kinds of reversals which are rhetorical reversal, expectation reversal, evaluation reversal, reversal of sentiment, relationship reversal, reversal from opposite pair, and reversal from satiation. The paper considered that verbal ironies can be classified by the kind of reversal they generate.

2.2 Linguistic Features

2.2.1 Irony Markers and Constructions

Although most of the studies saw irony as a pragmatics phenomenon, people also considered that it can be reflected on the verbal, grammatical, or semantic level. For example, on the verbal level, people often use words like thank, congratulate, welcome, happy, and interesting to express ironic meanings. Laszlo (2017) found 15 core evaluative words which often show in ironic expressions. She generated patterns from these core evaluative words to extract ironic sentences from the corpus. For example, when the word love is in the pattern “NP + would/ ’d/ wouldn’t + love”, it is highly possible to be an ironic expression. On a grammatical level, people often use the subjunctive when they intend to be ironic. Besides that, semantic conflict is the most direct way to express ironic meaning. The incompatibility between the main words of the proposition leads to the ridiculousness of the proposition (e.g. It’s very considerate of you to make such a loud noise while I was asleep). Besides, (Ghosh and Muresan, 2018) also categorized irony markers according to trope, morphosyntactic, and
Li (2021) considered that ironies are often expressed by specific “constructions”, especially in short discourses. Larger than “core evaluative words” in Laszlo (2017), the “constructions” mentioned in Li (2021) are mostly in the form of idioms or phrases. The crucial feature of them is the lack of predictability. Most of them do not have to rely on too much contextual information, they themselves can provoke the process of reversal for readers or listeners (e.g. 贵人多忘事 (honorable people frequently forget things)).

2.2.2 Irony in Communication
Researchers claim that by using ironies, people have several kinds of intentions.

Be polite: According to Brown et al. (1987), when unfavorable attitudes such as resistance, criticism, and complaints are stated with irony, the threat to the listener’s reputation is reduced. The irony, as stated in Giora (1995), is an indirect negation. People prefer to utilize indirect negation to be polite to their listeners because direct negation can generate great unhappiness;

Ease criticisms: As reported by Dews and Winner (1995), irony helps to ease the expression’s evaluative function. They believe that the incompatibility between literal meaning and contextual meaning can make it difficult to articulate negative feelings. However, Toplak and Katz (2000) argued that, while irony literally avoids conflict, it is more aggressive from the perspective of the speaker’s goal;

Self-protection: Sperber and Wilson (1986) proposed the “echoic” idea, which stated that irony is a detached utterance that is simply an echo of another people’s thought. It’s a self-protection tactic, especially when the speakers are members of marginalized groups. According to Gibbs (2000), the irony is an “off-record” statement that allows speakers to deny their true intentions and avoid being challenged;

Be amusing: Gibbs (2000) reported that when young people intend to be humorous, 50% of their communication is ironic. It can assist people in creating a dialogue platform on which speakers and listeners can agree and communicate more easily.

2.3 Irony and Sarcasm
Most of the studies saw sarcasm as a subset of irony (Tannen et al., 2005; Barbe, 1995; Kumon-Nakamura et al., 1995; Leggitt and Gibbs, 2000; Bowes and Katz, 2011). Sarcasm is often recognized as “a nasty, mean-spirited, or just relatively negative form of verbal irony, used on occasion to enhance the negativity expressed relative to direct, non-figurative criticism” (Colston, 2017).

One of the peculiarities of sarcasm is whether or not the speakers intend to offend the listeners. Kumon-Nakamura et al. (1995), for example, believe that sarcastic irony always conveys a negative attitude and is intended to harm the object being discussed. The non-sarcastic irony, on the other hand, can communicate either a good or negative attitude, and it is rarely meant to be hurtful. Barbe (1995) concurred that the core difference was “hurtful”.

She claimed that irony is a face-saving strategy while sarcasm is a face-threatening action. Ridicule is another feature of sarcasm. According to Lee and Katz (1998), sarcasm is closer to ridicule than irony. Their experiment revealed that sarcasm is directed at a single person, but irony is directed toward a large group of people. Haiman et al. (1998) claimed that one of the most distinguishing characteristics of sarcasm is that the literal meaning of its words is always positive. However, he did not convey his thoughts on irony. Whereas Litman and Mey (1991) viewed this topic from another angle. While there are many various forms of ironies, they believe that there is only one type of sarcasm because “sarcasm cannot exist independently of the communication setting.”

Cognitive scientists approached the difference in experimental studies. Previous research in child language acquisition (Glenwright and Pexman, 2010) reported that children understood sarcastic criticism later than they could understand the non-literal meanings of irony and sarcasm, implying different pragmatic purposes of irony and sarcasm. Filik et al. (2019) utilized fMRI and found out sarcasm is associated with wider activation of semantic network in human brains compared to irony.

However, most computational linguistics researchers used irony and sarcasm interchangeably, since the boundary between these two concepts is too vague for even human beings, let alone for machines. Joshi et al. (2016) and Sulis et al. (2016) verified this claim from both human annotators and computational perspectives. Although in this paper we will mainly focus on the literal irony processing and discuss the inspirations from recent research output in sarcasm processing, we aim to encourage unify sarcasm under the framework
of irony via fine-grained annotation schemes.

3 Irony Datasets Perspectives

3.1 Irony Textual Datasets

Some main target databases for irony processing research include social media platforms like Twitter and online shopping websites like Amazon. For example, Reyes and Rosso (2012) collected a 11,861-document irony dataset based on customer reviews from several websites. Some other preliminary attempts to build irony benchmark datasets is Reyes et al. (2012) and Reyes et al. (2013), in which they used self-generated hashtag #irony as the gold standard and constructed 40,000-tweet and 50000-tweet datasets from Twitter respectively, each including 10,000 ironic tweets and remaining non-ironic ones. The irony benchmark dataset that is now widely used is from Van Hee et al. (2018a), consisting of 4,792 tweets and half of them were ironic. This dataset was also constructed via searching hashtags including #irony, #sarcasm, and #not.

There were also tremendous attempts to construct benchmark datasets in other languages. For example, Tang and Chen (2014) firstly built the NTU Irony Corpus including 1,005 ironic messages from Plurk, a Twitter-like social media platform, by mining specific ironic patterns and manually checking extracted messages. A recent Chinese benchmark dataset for irony detection was constructed by Xiang et al. (2020), which includes 8,766 Weibo posts, labelled from not ironic to strongly ironic in a five-scale system. Besides, irony datasets in Spanish, Greek, and Italian are also widely available. A comprehensive overview of the irony datasets is listed in Table 1.

3.2 Data Source and Construction Methodology

Twitter, as one of the most trending social platforms, is the major source of irony benchmark datasets. Given Plurk and Weibo’s similarity to Twitter, online short-text social medias are almost the only origin for present datasets, which might lead to several potential problems. For example, the limitation of 140-word introduced specific bias towards short-text classification and long texts remain a problem. Besides, the judgment of irony might be highly dependent on contextual information like previous comments or retweets and one single tweet could be meaningless itself. At last, topics on social media platforms might also be highly biased towards political or sports topics. Interestingly, previous research (Ghosh and Muresan, 2018) reported that for Twitter and Reddit, different ironic markers played the most important roles, further emphasizing the needs of multiple sources for robust NLU.

As for the construction methodology, most datasets adopted "keywords and hashtags" filtering strategy, and some of them followed by human annotations. With annotation schemes put aside, Sykora et al. (2020) explored how self-generated tags could correspond to real labels correctly via a manual semantic analysis. At the worst case, only 16% of the tagged #sarcasm tweets are unambiguous sarcastic tweets according to well-trained linguists, which further emphasized the necessity of human annotation.

3.3 Annotation Schemes

As shown in Table 1, various datasets have been labelled differently. Most datasets were annotated as binary classes, ironic versus non-ironic. Both Chinese datasets adopted different strategies by annotating ironic elements and intensity respectively. Van Hee et al. (2018a) is the only one doing fine-grained labelling, into verbal irony by polarity contrast, other verbal irony, situational irony, and non-ironic.

Another annotation scheme was proposed by Cignarella et al. (2020b). They annotated irony activators at the morphosyntactic level and distinguished different types of irony activation. This annotation scheme dived into syntactic information and offered information for analyzing ironical constructions.

3.4 Future work in datasets

A high-quality benchmark dataset is crucial to measure NLU capability and advance future research. We are calling improvements for a future benchmark dataset from following perspectives.

1) Diverse data sources like literature, daily conversations, and news articles should be involved, and the distribution of text length should be relatively balanced.

2) A uniform annotation scheme and strategy is awaiting for construction. For example, situational irony is needed; ironic intensity should be labelled as reference rather than exclusive labels in Xiang et al. (2020); a scheme to unify sarcasm and irony could be expected.
Multi-X perspectives should be incorporated into the construction of datasets. For example, a multimodal and multilingual dataset could enhance irony identification in a grounded environment.

4 Irony Processing Systems

In this section, we will discuss the progress of irony processing systems comprehensively, organized along the development of machine learning and deep learning.

4.1 Irony Detection

4.1.1 Rule-based Detection Methods

Tang and Chen (2014) extracted ironic expressions based on five patterns summarized from mandarin Chinese. Li and Huang (2020) further expanded ironic constructions to more than twenty and proposed a systematic irony identification procedure (IIP). Besides Chinese, Frenda (2016) utilized sentiment lexicons, verb morphology, quotation marks, and etc. to design an Italian irony detection model and got competitive performance with machine learning models. However, rule-based models are too complex and hard to generalize for wider applications.

4.1.2 Supervised non-neural network era

Most research took irony detection as a simple classification problem. Before the popularity of deep learning, feature engineering is crucial for accurate irony detection. Generally, features could be divided into several levels.

Lexical features Lexical features are at the foundational level of NLP features, basically divided into bags of words (BOW) sets, word form sets, and conditional n-gram probabilities (Van Hee et al., 2018b). Representative BOW sets mainly include n-grams and character n-grams. Word form sets focus on number and frequency, such as punctuation numbers, emoticon frequencies, character repetitions, etc. Despite their easiness to get, lexical features were proved effective in much research.

Syntactic features Syntax is mainly quantified via parts-of-speech and named entities. After tagging, the number and frequency of both characteristics could act as features in classification models. Besides, hand-crafted syntactic features also included clash before verb tenses (Reyes et al., 2013) and dependency parsing (Cignarella et al., 2020a).

Semantic features Van Hee et al. (2018b) approached semantic features based on the presence or not in semantic clusters, which were trained on a irony Twitter corpus with the Word2Vec algorithm (Mikolov et al., 2013).

Linguistic-motivated features Irony processing is deeply associated with sentiments and emotions. Therefore, researchers have offered many characteristics to capture irony patterns. For example, Reyes et al. (2013) proposed the feature of contextual imbalance, which was quantified via measuring the semantic similarity pairwise. Generally, most features could be categorized into ambiguity (Reyes et al., 2012) and incongruity (Joshi et al., 2015). Take incongruity as an example, implicit incongruity was defined as a boolean feature checking containing implicit sentiment phrases or not; explicit incongruity was defined as number of times a polarity contrast appears. Theoretical research (Ghosh et al., 2020) is encouraging more semantic and pragmatic features to better capture ironies.

Features at various levels were concatenated with classifiers, including naive bayes, decision tress, support vector machines (SVMs), etc. (Van Hee et al., 2018b) to get final classification results.
4.1.3 Supervised Neural Network Era

There have been irony detection tasks in various languages after 2015, among which most participants used convolutional neural networks (CNNs) and RNNs methods to detect ironical expressions. For example, in SemEval-2018 Task 3 (Van Hee et al., 2018a), Wu et al. (2018) classified irony tweets and their ironical types in a densely connected LSTM network, together with multitask learning (MTL) objectives in the optimization. In IroSvA (Ortega-Bueno et al., 2019), González et al. (2019) utilized transformer encoders only for detecting Spanish ironical tweets.

Besides, Ilić et al. (2018) firstly utilized contextualized word representations ELMo (Peters et al., 2018) with Bi-LSTM to detect ironies. Zhang et al. (2019) enhanced irony detection with sentiment corpora based on attention Bi-LSTM RNNs, and achieved state-of-the-art results on Reyes et al. (2013) dataset.

4.1.4 Pretraining and Fine-tuning Paradigm

Since BERT (Devlin et al., 2019), the “pretraining and fine-tuning paradigm” has become the mainstream in NLP research due to its extraordinary capacity in dealing with contextualized information and learning general linguistic knowledge. Recent development in irony detection also witnessed the usage of PLMs. Xiang et al. (2020) released several baseline results along with the dataset, among which BERT had highest accuracy (5% higher than Bi-LSTM methods). Potamias et al. (2020) proposed a recurrent convolutional neural network (RCNN)-RoBERTa (Liu et al., 2019) strategy and improved the results on the SemEval-2018 dataset by a large degree. Besides, Cignarella et al. (2020a) explored syntax-augmented irony detection with multilingual BERT (mBERT) in multilingual settings. To sum up, some representative studies in irony detection are detailed in Table 2.

4.2 Irony Generation

Irony generation is mostly an underexplored research field besides Zhu et al. (2019), in which they defined irony generation as a style transfer problem, and utilized a Seq2Seq framework (Sutskever et al., 2014) with reinforcement learning to generate ironical counterparts from a non-ironic sentence. Concretely, they designed the overall reward as a harmonic mean of irony reward and sentiment reward, which was trying to capture the sentiment incongruity. In terms of the evaluation, besides traditional natural language generation metrics like BLEU, they also designed task-specific evaluation metrics, which should be further enhanced in irony and even figurative language research.

Future work in irony generation could be advanced in new PLMs and theoretical accounts. For example, no attempts were made to generate ironical expressions after generative PLMs like BART (Lewis et al., 2020). Controllable irony generation and its interaction with agents are interesting topics remaining for future exploration. Besides, irony theories could be further utilized. In recent research on unsupervised sarcasm generation (Mishra et al., 2019; Chakrabarty et al., 2020), context incongruity, valence reversal, and semantic incongruity were merged to enhance the generation.

5 Discussion

5.1 Irony for Downstream NLP Tasks

Irony is directly associated with downstream NLU tasks like sentiment analysis and opinion mining. For example, the sentence retrieved from Filatova (2012) I would recommend this book to friends who have insomnia or those who I absolutely despise. is classified as positive by fine-tuned sentiment analysis RoBERTa model (Heitmann et al., 2020), which is apparently opposite to human evaluation. Wrong sentiment judgments will potentially lead to contrary opinion mining. We suggest that irony could be further captured through introducing incongruity embedding or specific pattern matching. Joshi et al. (2015) designed linguistic-motivated features implicit and explicit incongruity, which are inspiring for enhancing irony understanding. Consider another example task, machine translation, in which wrong translation will potentially lead to totally opposite meanings. We encourage to model discourse features (Voigt and Jurafsky, 2012), such as ironic patterns and punctuation as embeddings for robust irony translation.

In addition, we are looking forward to the research of irony and sarcasm processing in NLP for social good (NLP4SG), especially considering the strong sentiments hidden in ironies. A recent work (Chia et al., 2021) explored cyberbullying detection and this was a starting point to handle online harmful ironical contents.

5.2 Multi-X Perspectives

Recent developments in NLP and sarcasm processing encourage us to approach discourse processing...
from multiple angles. In this part, we will review and suggest several multi-X perspectives for irony and figurative language processing.

### 5.2.1 Multimodal Irony Processing

Linguistic interactions are not solely consisted of texts. Besides, facial expressions and speech communications are crucial to convey emotions and feelings. For example, Skrelin et al. (2020) reported people could classify ironies based on phonetic characteristics only. Consequently, it is conceivable that multimodal methods could help with irony detection. Schifanella et al. (2016) made the first attempt in multimodal sarcasm detection, in which they extracted posts from three multimodal social media platforms based on hashtags. Then they used SVMs and neural networks to prove the validity of information in enhancing sarcasm detection.

Castro et al. (2019) made a great improvement in multimodal sarcasm detection by introducing audio features into the dataset. The experiments also verified the importance of more modalities in sarcasm processing.

Future work in multimodal irony processing should include a comprehensive multimodal irony dataset based on MUSTARD dataset (Castro et al., 2019) with more fine-grained annotation schemes. Additionally, most methods (Pan et al., 2020; Liu et al., 2021) explored sarcasm by introducing inter-modality and intra-modality attention in single-stream setting. How double-stream multimodal pretrained models (MPMs) will encode and interact in complex discourse settings remains an interesting problem to solve.

#### 5.2.2 Multilingual Irony Processing

To understand irony in a multilingual context is even harder due to cultural gaps. Previously listed dataset includes a Hindi-English code-mixed irony dataset (Vijay et al., 2018), in which they offered an example:

- **Text:** The saying ‘old is gold’ is old. Today’s generation thinks ‘gold is old’ but only the old one makes sense. #MindF #Irony.

- **Translation:** The saying ‘old is gold’ is old. Today’s generation thinks ‘gold is old’ but only the old one makes sense. #MindF #Irony.

Cignarella et al. (2020a) explored how mBERT performed in multiple languages’ irony detection tasks separately. Given it has been proved code-switching patterns are beneficial for NLP tasks like humor, sarcasm, and hate speech detection in RNNs settings (Bansal et al., 2020), a future direction is to merge the irony detection datasets from multiple languages (consider Karoui et al. (2017)) or even code-mixed texts, and explore how multilingual datasets could enhance irony understanding in mBERT.

#### 5.2.3 Multitask Irony Processing

MTL is to make models learn several tasks simultaneously rather than independently once at a time. Recent work in figurative language processing proposed several MTL strategies to improve the performance interactively. Chauhan et al. (2020) proposed a MTL framework to do sentiment, sarcasm, and emotion analysis simultaneously and the framework yielded better performance with the help of MTL.
Generally, we will classify figurative language into several categories like metaphors, parodies, humors, ironies, and etc. However, noted that there are not clear differences between each other, a single task figurative language processing will only focus on one particular aspect and fail to capture the interactions. Recent work (Ao et al., 2022) also verified the combination of humor and sarcasm could improve political parody detection.

PLMs could understand figurative language better than random but apparently worse than human evaluation (Liu et al., 2022). We suggest that future work should consider domain adaptation towards figurative language as a whole via weak supervision. MTL strategy could utilize previous research in corpus linguistics, and design an appropriate proportion in summing the loss function. Besides, a unified framework to model figurative languages could be expected.

5.2.4 Multiagent Irony Processing

Human-like language generation is a central topic in multiagent interactive systems. Besides robots’ ironical understanding, we are also curious about how robots could generate ironical expressions. Unlike transferring non-ironic sentences to ironic, multiagent irony measures the performance during the interactions in dialogues. Ritschel et al. (2019) improved the robots by introducing ironic expressions, which showed better user experiences in human evaluation.

Further explorations in multiagent irony could aim at better dialogue state tracking and understand when irony should be introduced.

5.3 New Tasks: Inspiration from Sarcasm

Compared to sarcasm, irony is rarely seen as a term in NLP conferences. Recently we have witnessed great improvements in sarcasm processing and in this part we will discuss how new tasks in sarcasm could motivate irony research.

Data Collection As discussed, datasets are highly dependent on hashtags as a signal to extract ironical expressions. Shmueli et al. (2020) proposed an algorithm to detect sarcastic tweets from a thread based on exterior cue tweets. A distant supervision based method for extracting ironies from platforms is crucial, given ironies in conversational contexts are central topic in the future.

Intended and Perceived Irony Oprea and Magdy (2019) explored how author profiling affected the perceived sarcasm (manual labelling) versus the intended sarcasm (hashtags), and verified the difference between both. Further, Oprea and Magdy (2020) introduced iSarcasm dataset which divided intended sarcasms and perceived sarcasms. The state-of-the-art sarcasm detection models performed obviously worse than human evaluation on this dataset. Future work could focus on multimodal perceived and intended irony, especially across various cultures.

Target Identification Sarcasm target identification was firstly proposed in Joshi et al. (2018), in which sarcasm targets were classified as one target, several targets and outside. Patro et al. (2019) introduced sociolinguistic features and a deep learning framework, and improved target identification by a lot. For irony processing, most ironical expressions do not equip a specific target in itself as previously discussed. However, its ironical effects are likely in dialogue or visually grounded environment, which encourages us to enhance irony datasets in aforementioned ways.

Irony Explanation Irony, according to the definition, have opposite real meanings to literal meanings. However, this does not mean adding a single negation could interpret ironies well. Kumar et al. (2022) proposed a new task, sarcasm explanation in dialogue. Irony explanation might encounter more complex problems due to relatively low proportion of targets. Still, we should include irony explanation as a branch of multimodal irony processing like Desai et al. (2021).

5.4 Explainable Irony Processing

Explainable machine learning is of interest for most researchers to uncover the blackbox. In irony processing, we are also curious about why specific expressions are recognized as ironies. Buyukbas et al. (2021) explored explainability in irony detection using Shapley Additive Explanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) methods. Results showed that puncuations and strong words play important roles in irony detection.

For future work, we suggest using explainable methods in multimodal settings and check how different modalities act various roles in making a class label.

6 Conclusion

In this paper, we reviewed the development in automatic irony processing from underexplored theo-
retorical and cognitive science to computational perspectives, and offered a comprehensive analysis in future directions. We hope that our work and thinking will encourage further interdisciplinary research between linguistics and human language technology, motivate the research interests in irony and even, figurative languages.

Acknowledgement

This review is based on the first author’s previous research proposal. We would like to express our thanks to Professor Chu-Ren Huang and Dr. Yat Mei Lee for their suggestions.

References


Alessandra Teresa Cignarella, Manuela Sanguinetti, Cristina Bosco, and Paolo Rosso. 2020b. *Marking


José-Ángel González, Lluís F. Hurtado, and Ferran Plà. 2019. Elirf-upv at irosva: Transformer encoders for spanish irony detection. In IberLEF@SEPLN.


Aditya Joshi, Vaibhav Tripathi, Pushpak Bhattacharyya, Mark Carman, Meghna Singh, Jaya Saraswati, and


Shivani Kumar, Atharva Kulkarni, Md Shad Akhtar, and Tamnoy Chakraborty. 2022. When did you become so smart, oh wise one?! sarcasm explanation in multimodal multi-party dialogues.


Emmy Liu, Chen Cui, Kenneth Zheng, and Graham Neubig. 2022. Testing the ability of language models to interpret figurative language.


Debora Nozza, Elisabetta Fersini, and Enza Messina. 2016. Unsupervised irony detection: A probabilistic model with word embeddings. In KDIR.


Jasabanta Patro, Srijan Bansal, and Animesh Mukherjee. 2019. A deep-learning framework to detect sarcasm


Hannes Ritschel, Ilhan Aslan, David Sedlbauer, and Elisabeth André. 2019. Irony man: Augmenting a social robot with the ability to use irony in multimodal communication with humans. In AAMAS.


Andrea Santilli, Danilo Croce, and Roberto Basili. 2018. A kernel-based approach for irony and sarcasm detection in italian. In EVALITA@CLiC-it.


Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In NIPS.


