On the "Easy" Task of Evaluating Chinese Irony Detection

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

Emmanuele Chersoni
Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University
emmanuelechersoni@gmail.com

Rong Xiang
Department of computing, The Hong Kong Polytechnic University
csrxiang@comp.polyu.edu.hk

Chu-Ren Huang
Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University
churen.huang@polyu.edu.hk

Qin Lu
Department of computing, The Hong Kong Polytechnic University
qin.lu@polyu.edu.hk

Abstract

In this paper, we present a discussion on the problem in the evaluation of irony detection in Mandarin Chinese, especially due to the difficulties of finding an exhaustive definition and to the current lack of a gold standard for computational models. We describe some preliminary results of our experiments on an irony detection system for Chinese, and analyze examples of irony or other related phenomena that turned out to be challenging for NLP classifiers.

1 Introduction

In recent years, irony became a hot topic which draws the attention of both cognitive linguists and computational linguists. As a special kind of rhetorical device, its most striking feature is the incongruity between its literal meaning and contextual meaning. This feature means that the processing procedure of irony should be more complex than other expressions for both humans and machines. And since ironic expressions are highly context-dependent while analyzing contextual information is not easy for computational systems, the automatic detection of irony is a hard task. However, if we cannot effectively detect it, entire sentences will be understood in a totally different way, affecting the performance in many NLP tasks.

Generally speaking, ironies are often defined as the expressions whose literal meanings are incongruous with their contextual meaning. However, according to our observation, the use of the word "incongruous" is inadequate. Some other kinds of expressions can also show incongruities between their literal and contextual meanings:

- Exaggeration or Hyperbole:

(1) I was fired and caught a cold in the same day. I must be the most luckless people in the world!!!

The speaker should know that there must be some more luckless people than him/her in the world. He/she does not mean to state the fact that he/she is the most luckless people, but just

Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 452-460, Hakodate, Japan, September 13-15, 2019

want to express his/her strong emotion by the exaggeration device.

- Metaphor:

(2) Mark Twain's work is a mirror of America.

The speaker does not use simile by using the words "as" or "like" in this sentence. Literally speaking, the meaning of the sentence is that "Mark Twain's work is a mirror". However, it is obvious that literature works cannot be a real mirror. The contextual meaning of the sentence is that Mark Twain's work shows the actualities of America.

Another rhetorical device which needs to be emphasized in this discussion is the pun. Ironies and puns at least have the following similarities:

- They both have some incongruities in several linguistic levels;
- In communication, not all the listeners can understand their meaning;
- Speakers may make pun / ironies unintentionally;
- There are bad ironies/puns like "icy jokes". That is, the expression is so unfunny / non-ironic / non-punny that it is kind of funny / ironic / punny.

However, they are not the same concept, neither. Puns just ask for double entendre. If an expression can express more than one meaning, it can be a pun.

For example:

(3) 人类失去联想，世界将会怎样？
ren2 lei4 shi1 qu4 lian2 xiang3, shi4 jie4 jiang1 hui4 zen3 yang4?
What will happen to the world if human lost their imagination?

This is a famous advertising slogan of Lenovo (a technology company) in China. Their Chinese name "联想 (lian2 xiang3)" means "imagination" in Chinese, so here the word "联想 (lian2 xiang3)" is a pun. It can not only refer to imagination, but also to Lenovo. However, since "imagination" and "Lenovo" are not contrast with each other, it just a pun instead of an irony.

(4) The dear leader played the "trump" card and played it very well.

In this sentence, the word "trump" refers to the Joker card in poker game, its basic meaning. It can also refer to the advantage that makes people more likely to succeed. It can even be a pun since the word "trump" can also refer to President Trump. However, this sentence is not necessarily an irony. Only when we can get further context, which can prove that the speaker is an opponent of President Trump (or at least, not the supporter of him), the sentence can be interpreted as ironic.

Compare with puns, although most of the ironies also have double meanings, there are some restrictions. If the two meanings of an expression are just incongruous instead of contradict with each other, it can't be an irony. Besides that, having double meanings even isn't an essential condition to ironies. For expressions like counterfactuals (such as "太阳从西边出来 (tai4 yang cong2 xi1 bian1 chu1 lai, The sun rise in the west.)") and satiations (such as "你相信吗？你相信吗？ (ni3 xiang1 xin4 ma? ni3 xiang1 xin4 ma?, Do you believe? Do you believe?)"), all the words are in their original meanings. However, the listeners still feel they are ironic.

From examples above, we know that a suitable definition of irony should not only show its linguistic features but also can effectively differentiate them from other linguistic phenomena. Both "incongruity" and "opposite" are not sufficient or necessary features, although it can involve both of them. Huang (2019) proposed that "reversal" is the critical nature of irony. All the features including "incongruity" and "opposite" can be seen as tools to achieve this "reversal".

The concept of "reversal" comes from "reversal theory" (Apter, 1982) in social psychology field. It is an important theory for personality changes as well as for change of belief / knowledge, in the context of studies on persuasion. This concept can include not only the reversed meanings, but also the situations in which the literal meaning is left unchanged and the reversal concerns only its semantic or pragmatic effect.

In this paper, we accept this definition and view irony as the expression which makes people experience a reversal during the understanding process.

In our experiments, we tried to use classifiers and word embedding methods to classify different types of ironic expressions, on the basis of the available resources. Starting from the definition basing on the concept of "reversal", we hope to find an efficient method to build reliable dataset which can be used to train and test irony detection models. Then we also apply our dataset to some machine learning models and report our preliminary results.

2 Related Work

A lot of efforts have recently been devoted to irony detection in natural language processing. The model by Buschmeier et al. (2014) give more than ten features including Imbalance (the overall polarity of words is different from the polarity of the whole sentence), Hyperbole (a sequence of three positive or negative words in a row), Quotes (up to two consecutive adjectives or nouns in quotation marks have a positive or negative polarity), Pos/Neg Punctuation (a span of up to four words contains at least one positive (negative) but no negative (positive) word and ends with at least two exclamation marks or a sequence of a question mark and an exclamation mark (Carvalho et al., 2009), Punctuation, Interjection (such as "wow" and "huh"), Laughter and Emoticon. Basing on these features, they use five different classifiers to analyze an Amazon review corpus which has 437 ironic reviews as well as 817 non-ironic reviews. The best F1-score they reported (74.4) is reached by combining star-rating with bag-of-words and specific features, and then using Logistic Regression to classify the corpus. A lot of later researches use features that are similar to their set.

Comparing with them, Reyes and Rosso (2014) used more lexical features. They divided the features they used into three layers. The signatures layer includes the features like Pointedness (refers to the contents that reflect a sharp distinction in the information. E.g. punctuation, emoticons and capitalized words), Counter-factuality (words hint an opposition. E.g. nevertheless, nonetheless and yet) and Temporal compression (words represent an abrupt change. E.g. *suddenly*, *now* and *abruptly*). The emotional scenarios layer includes the features like Activation (means the degree of emotion), Imagery (whether the words are easy to form a mental picture) and Pleasantness (the degree of pleasure of the words). The unexpectedness layer includes the features like Temporal Imbalance and Contextual Imbalance (whether there is an opposition of polarity or attitude in the time line / context). Their basic idea is much closer to our definition of irony since the three triggers of reversal (contingent events, frustration and satiation, see Apter, 1982) are all included in their features. They claimed that negation should be a useful grammatical category to detect ironies and also report the difficulties in the automatic detection task.

Sarcasm is a rhetorical device which share many important features with irony. The main difference between irony and sarcasm is whether the speakers intend to hurt someone by their words. Similar to irony, sarcasm experience a reversal in both meaning and sentiment, so sarcasm detection models can also give us some inspiration on irony detection. Ghosh et al. (2015) used a word embedding method to detect sarcasm and introduce a useful platform Amazon Mechanical Turk (MTurk). This platform can rephrase the sarcastic utterances to their intended meanings (e.g. "I love going to the dentist" can be rephrased as "I hate going to the dentist" or "I don't like going to the dentist"). They use this platform to rephrase 1,000 sarcastic Tweets and get 5,000 sarcastic – intended message pairs (each sarcastic message has five intended candidates). Meanwhile, they use co-training algorithm and statistical machine translation alignment method to extract 80 semantically opposite paraphrases. By extracting the context vectors with word embedding, they got the contextual features of each sarcastic utterance as well as its intended pairs. By using the distributional approach *w2vsg* with the Kernel classifier, they achieved the highest F1-score of 97.5% in their study.

Joshi et al. (2017) summarized the main approaches on sarcasm detection tasks. Rule-based approaches focus on the rules which rely on indicators of sarcasm. Feature Sets approaches usually use bag-of-words as features. Learning algorithms mainly rely on different kinds of SVM models. And deep learning-based approaches can give us further insights when the datasets are big enough. They claimed that pattern discovery was the early trend in sarcasm detection, while the use of context will be the new trend of the task.

Author-specific context, conventional context and topical context will play more and more important roles in future research, which can be also be true for the irony detection task.

However, researches on Chinese irony detection are still limited. Only one Chinese irony corpus has been built by Tang and Chen (Tang and Chen, 2014). Basing on the NTU Sentiment Dictionary, they extracted 304,754 messages from Plurk. These messages are the texts with negative emoticons and positive words. They believed they are potential candidates of ironic expressions. Then they retrieved all the messages which contain the pattern "degree adverb + positive adjective" from the candidates then manually reviewed whether they are ironic expressions.

During the review task, if annotators found some new irony patterns, they would also retrieve all the messages which contain this new pattern and then manually check them. Finally, they found 1,005 ironic messages from all the candidates and divided them into five groups according to the patterns they have. These patterns are: degree adverbs + positive adjective, positive adjective with high intensity, positive noun with high intensity, "很好" (hen3 hao3, very good) and "可以再…一点" (ke3 yi3 zai4…yi4 dian3, It's okay to be worse).

They used these patterns to extract ironic expressions from Yahoo corpus and obtained 36 ironic texts from it. Their work is, without a doubt, a meaningful try but the patterns that they found are too short. A lot of dynamic and relatively abstract ironic expressions are not included in their corpus and whether just use one pattern (degree adverbs + positive adjective) at the very beginning of the bootstrapping procedure is adequate for this task is worth discussing.

Besides that, Deng, Jia and Chen (2015) construct a feature system for Weibo irony identification task. The system contains six features:

- the basic emotion feature of the words in the sentences: be recorded by unigram
- homophonic words: such as "河蟹 (he2 xie4, river crab)" and "和谐 (he2 xie2, harmony)"
- sequential punctuations: more than three
- length of the text: Weibo texts are divided into short, middle and long. They believed that the length of the text will affect the quantity of sentiment information.
- verb passivization: abnormal collocation of the structure "被 + verb" like "我被就业了 (wo3 bei4 jiu4 ye4 le, I am gotten a job)
- incongruities between emotions in and out of the quotation marks: whether the emotion words in the quotation marks is positive while the emotion words out of the quotation is negative or vice versa.

Basing on this system, they reported the highest precision rate and F-score from the Logistic Regression Model (Precision rate: 78.31%, F1-score: 71.13%) and the highest recall rate from Decision Tree Model (71.86%).

From current studies we can see that now we lack of Chinese irony resources. It is no doubt a big problem. On the one hand, we do not have a suitable corpus for both machines and researchers to extract features and find patterns. Only hundreds of examples cannot effectively help us to summarize the rules. Moreover, they usually do not cover enough types of ironic expressions. On the other hand, since both the theoretical and applied researches on Chinese ironies are limited, we do not have an adequate corpus as well as a standard to evaluate the quality of Chinese irony detection. However, constructing such a corpus completely by annotators is a hugely difficult task since ironic expressions account for a very small percentage in most corpora (usually less than 1%). In other words, ironic expressions are just like needle in the haystack. Therefore, it is meaningful to find a method which can filter ironic expressions automatically and precisely.

3 Classification Experiments

3.1 Data Collection

For our study, first we need to build a provisional dataset for the classifiers. The dataset need to include enough ironic expressions as well as non-ironic expressions that share some features with them, as representatives of the negative class. The Taiwanese irony corpus built by Tang and Chen (2014) is a suitable resource to form the ironic part. According to what they reported in their paper, this corpus has 1,005 ironic messages

collected from Plurk and five ironic patterns can be extracted from them. Therefore, its scale is big enough for the classifiers to detect features. And since it is manually-checked, the expressions are reliable and typical.

The non-ironic expressions, for the moment, are of two different types. The first type is sentences extracted from microblogs. Among those 1,005 ironic messages in Taiwanese corpus, 993 of them have both "positive sentiment" tag and "ironic" tag. It means that they are not only ironic but also have positive lexicons. We use them as patterns. Meanwhile, we screen out all the positive sentences from two sentiment-labeled microblog corpora as candidates. After that, we calculate all the cosine similarities between each patterns and candidates by using bag-of-word vectors. We filter out all the candidates whose cosine similarities are less than 0.5 and choose the best five candidate matches of each pattern. We finally extract 2,241 candidates from microblogs corpus. All of these sentences share high similarities of words with the ironic expressions in Taiwanese corpus. However, they are non-ironic since the sentiment tag of each whole sentence is still positive.

The second type is puns. As we mentioned, puns share a lot of similarities with irony. It can even confuse humans in some cases so we wonder whether they are also confounding factors to computers. Introducing puns in our dataset can broaden our range from a theoretical point of view, no matter what the classification results show us.

If the classifiers can correctly classify most of (or even all of) the ironic puns as ironies and filter out the non-ironic ones, it is no doubt an exciting result to show that the detection method is strong enough to filter out irony-like expressions from real ironies. If the classifiers classify all the puns as ironies, at least it shows that the filter can identify double entendre from other expressions. Finding out the features that different kinds of double entendre have in common and work out why these features are effective enough to differentiate double entendre from other expressions should be a new and meaningful topic to do research on. Even if the result shows that ironic and non-ironic puns are classified randomly, it can also be a treasurable resource for error analysis and future researches. There should be some rules inside the wrong results. Why some of the non-ironic puns can confound the classifiers while others cannot? It is also a worthwhile topic.

For this part of our dataset, first we extract 906 candidates online. Then we manually checked these candidates to find typical puns. The second meaning of these puns can be easily recognized and the two meanings are different enough to differentiate from each other. We finally selected 176 puns from 906 candidates. Besides that, we also add 30 *xiehouyu* to the dataset. *Xiehouyu* is a kind of Chinese idiom that usually has two parts. The first part of it is descriptive while the second part carries its double meanings. Since *xiehouyu* are conventional expressions, no matter whether the second parts are shown in the discourse, they are typical and popular template for puns.

Finally, the three parts we mentioned above construct our classification dataset. They are:

Positive examples:

- 993 ironic expressions with positive lexicons (from Taiwanese corpus, Tang and Chen (2014))

Negative examples:

- 2,241 non-ironic expressions which have high similarities with those ironic expressions (from sentiment-labeled microblog corpora, Zhou et al. (2018) and Wang et al. (2016))
- 206 puns: 176 complete sentences with typical puns (randomly extract from different websites) and 30 popular xiehouyu.

Therefore, now we have 3440 sentences in the database. We automatically mix them and divide them into training set (3,097 sentences) and test set (343 sentences).

3.2 Classification Task and Results

In this section we use two widely-used classifiers (Support Vector Machine (SVM) and Logistic Regression (LR)) to train and classify our data. The Support Vector Machine takes as input a sentence feature vector, a representation that is

computed as the mean vector of the embedding of the words in the sentence. The results of SVM are as follows:

Kind of Sentences	Number of Sentences	Precision	Recall	F1-score
Negative Examples	244	0.98	0.97	0.97
Positive Examples	99	0.92	0.95	0.94
Average / Total	343	0.96	0.96	0.96

Table 1: Results of Support Vector Machine (SVM)

Similar to SVM, the Logistic Regression (LR) also uses a form of binary model to analyze the input sentences. Here the results of LR are as follows:

Kind of Sentences	Number of Sentences	Precision	Recall	F1-score
Negative Examples	244	0.98	0.98	0.98
Positive Examples	99	0.96	0.96	0.96
Average / Total	343	0.98	0.98	0.98

Table 2: Results of Logistic Regression (LR)

As we can see, both of the classifiers show good results. It is no doubt excited but it seems that the classification of irony sentences is too easy for computational models. Therefore, we still wonder that whether the limitation of the data affect the result. The reason why we have this consideration is that the ironic sentences in the Taiwanese corpus just have five typical ironic constructions but in actual discourses there should be more. We want to confirm whether the diversity of our data lower the difficulty of the classification task.

In future, we plan to collect more varieties of ironic expressions to extent our database. However, not matter whether the results will change significantly, our method should be a good standard to evaluate Chinese irony detection tasks.

3.3 Error Analysis

According to the results of LR model, these sentences are not correctly classified.

Ironic sentences which are classified as non-ironic:

(5) 以為訂的是晚上七點半回新竹的票，七點在自動售票機取到票才發現是晚上八點。特地再到網路訂位取票的窗口問可不可以換到七點半，結果票務人員發現我訂的是晚上八點從新竹到台北的票。我真的可以再天兵一點。。。:'-(

(6) 很好，團購了一個可以拿來澆花的、可能會摔破的、大陸製自行車水壺:-(

(7) 我真是太幸運啦！今天朋友抽籤，竟然抽到裡面最老的一位奶奶，８４歲。我真的是開心到不知道要說什麼．．．就是傻很大～而且他又喜歡話中代日語，所以我想．．這次專題真的很有挑戰性～:-(

(8) 今天真是太太太幸了,在台九上，大卡迎面而，不知哪一拳大的石，1秒2秒的，恰恰好直落在我的上方!碰的巨我－跳:-好人平安。

Non-ironic sentences which are classified as ironic:

(9) 回到纯爱的美好记忆。”””””欢迎爱光临”””””。”

(10) 飞蛾扑火-自取灭亡

(11) 上完课啦 终于告别早起的苦逼日子来全家吃个盒饭补充一下元气先下午接着上bec3 fighting!!!

(12) 为公司年会特意准备的谢谢！是不是很有：欧巴桑的feel“‘o(∩)o

We are not sure why (5) and (6) are not correctly classified since it seems that they have enough features for classifiers to make correct decisions (Example (5) have ironic features “ 真的 (zhen1 de, really)” and “ 可以再…一点 (ke2 yi3 zai4…yi4 dian3, it can be more…)” with negative emoticon “:'-(”. Example (6) has an ironic

feature as well as positive adjective " 很好 (hen2 hao3, very well)" with negative emoticon ":-(".). However, for example (7) and (8), even humans may also confuse whether the speakers are ironic. It is because that the event they described can either be considered as lucky or unlucky. The judgements just rely on the parties view the events from which angle. However, the speakers here use a lot of positive markers with relatively same quantity of negative markers. These markers will give too many vectors to the sentences and finally confuse the classifier. For non-ironic sentences, example (9) and (12) just have positive markers, we guess maybe the occurrences of sequential punctuations and seldom-used emoticon are marked as ironic features. Example (11) may be affected by the co-occurrence of "苦逼 (ku3 bi1, bitter)" and "fighting". According to this analysis, how to give different ironic constructions a reasonable weight in the classification task should be a meaningful topic.

Meanwhile, although we are not sure about the reason why example (10) are classified as irony, it is more than excited to see all the puns are correctly classified except it. It shows us although both irony and pun have double meanings, there must be some features which are strong enough to differentiate them from each other. It also supports our hypothesis that the critical nature of irony it reversal instead of incongruous or opposite in meaning since puns also have the latter. Finding out the features which can differentiate ironies from puns can be a valuable theoretical contribution.

4 Towards New Datasets for Chinese Irony Detection

As what we mentioned in 3.2, we wonder whether the scale of the database will affect our results. In order to richer the database, we need more ironic sentences as positive examples. These sentences must be as various as possible so that we can include enough actual instances for the classifiers to extract features. In the first step, we use the strategy which is similar to the Taiwanese corpus. We'll use some ironic constructions as key words to find candidates then manually check them. During the checking task, we may find some new ironic constructions. We'll use these new constructions to find more candidates as well as more ironic sentences. It just like makes a snowball. Using more constructions as key words in this step should be good for us to get more candidates. According to our definition, ironies should either express or facilitate reversals at different linguistic levels, so ironic detection should be more effective and accurate if we model it as a reversal detection instead of an incongruity detection task. For now, we've found at least seven kinds of ironic reversal:

1. Rhetorical Reversal: In Chinese, rhetoric questions can be formed in different ways such as:

   a) adding tag question with the verb 是 (shi4, to be) followed by a question particle 吗 (ma).

   b) adding emphasis with wh-words on manner/degree. For example:

(13) noun phrase+有这么+ verb phrase +的吗

Noun phrase + you3 zhe4 me + verb phrase + de ma

Is there anything can be done like this?

(14) 你以为你是谁？

ni3 yi3 wei2 ni3 shi4 shui2?

Who do you think you are?

   c) repetition of a normal question: It is a kind of satiation in reversal theory. Repeating an expression (no matter it is a question or not) again and again will makes listeners to question whether it is in its original meaning. It indicates stronger ironic intention.

2. Imperative sentences as dares: It is a kind of threaten to stop listeners from doing what they dare to do. Speakers use imperative but actually it is a prohibition. For example:

(15) noun phrase + 再 + verb phrase + (一+ quantifier) + (试试)

noun phrase + zai4 + verb phrase + (yi2 + quantifier) + (shi4 shi)

(Somebody) can (try to) do it (once more)

3. Evaluative reversal: This kind of reversal usually include some special lexical markers such as "亏 (kui1, fortunately)", which is marked in 现代汉语词典 (Xian Dai Han Yu Ci

Dian, 2016) to express irony/sarcasm.

4. Opposite pairs: This kind of expressions show ironic meaning by directly using contrastive linguistic pairs. (Ding, 2018)

5. Counterfactual constructions: These constructions reverse the factuality of a statement. It can be marked with adverbs such as "要不是(yao4 bu2 shi4, but for)", or formulaic counterfactual expressions such as "太阳从西边出来(tai4 yang cong2 xi1 bian1 chu1 lai, The sun rise in the west.)". (Jiang, 2019)

6. Reversal of sentiment: This happens when positive emotion words are used to express negative emotion, and vice versa.

7. Satiation: As what we mentioned, if speakers repeat an expression several times, listeners will question whether it is in its original meaning. Similarly, if speakers overuse certain polarity words (such as hyperbole), the listeners will also experience a reversal. If there are more than one assertive words or high degree adverbs in one sentence, it is highly possible to be an ironic expression.

Each kind of reversal can separate out more than one ironic constructions. Only using constructions from first four kinds we can easily extract 2,363 candidates from a single microblog corpus. Since most of the constructions are highly formalized and easy to retrieve, we are confident of finding more ironic constructions as well as positive examples by this method.

Meanwhile, in order to manually check the candidates in a standard way, similar to what Pragglejaz Group (2007) and Gerard J. Steen et al. (2010) did on metaphors, we construct an Irony Identification Procedure (IIP) to help annotators to make judgements. In short, the procedure should be as follows:

1. Read the entire sentence as well as the context (if available) to sketch an overall understanding of the meaning.

2. Determine the contextual meaning of core constructions of the text. These core constructions include idioms, adjective phrases, rhetorical devices, clauses which are linked by conjunctions and some other constructions which can express the attitudes of the speakers. Annotators should pay special attention to sentiments, evaluations and logic relations which are shown by these constructions in the given context.

3. Determine the literal meaning of each core construction. When finding literal meanings, researchers should neither consider about the construction meanings emerge after the combination of the components nor refer to any context. Literal meanings have to be: direct (can be understood without any context), formal (can be found in dictionaries) and common (frequently-used but do not use any rhetorical devices).

4. Compare the contextual meaning and the literal meaning of the construction to see whether the contextual one is the reversal of the literal one. Researchers should notice that the evaluation criterion is whether there is a reversal in the expression instead of just "incongruous". For example, if the literal sentiment of the construction is joy while the contextual sentiment of the construction is grossness or even wrath, it can be a reversal. If sentiment just changes from joy to excitement or from grossness to wrath, it is an "incongruity".

5. If the contextual meaning of a construction experiences a reversal from its literal meaning, mark it as an "ironic construction". If it hasn't been included in current ironic construction set, add it to the set and further use it to retrieve new candidates.

6. Basing on core constructions, judge whether the whole text experience a reversal. If so, chose it as a positive example.

As what Joshi et al., 2017 claimed in their paper, pattern discovery was the early trend of sarcasm detection and researchers will rely more on context information in the future. Therefore, we will also try to take context features into consideration. Features like logic confusion and topical context will be new topic we concern about.

References

Micheal J Apter. 1982. The Experience of Motivation: The Theory of Psychological Reversals. Academic Press.

Micheal J Apter. 1984. Chinese Irony Corpus Construction and Ironic Structure Analysis. In *Journal of Research in Personality*, vol. 18(3): 265-288.

Konstantin Buschmeier, Philipp Cimiano, and Roman Klinger. 2014. An Impact Analysis of Features in a Classification Approach to Irony Detection in Product Reviews. In *Proceedings of the EMNLP Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*.

Paula Carvalho, Luis Sarmento, Mario Silva, and Eugenio De Oliveira. 2009. Clues for Detecting Irony in User-Generated Contents: oh...!! It's So Easy ;-). In *Proceedings of the International CIKM workshop on Topic-Sentiment Analysis for Mass Opinion*.

Zhao Deng, Xiu-Yi Jia, Jia-Jun Chen 邓钊, 贾修一, 陈家骏. 2015. Research on Chinese irony detection in microblog 面向微博的中文反语识别研究. In *Computer Engineering Science* 计算机工程与科学, vol. 37(12):2312-2317.

Jing Ding. 2018. A lexical semantic study of Chinese opposites. Springer Singapore, Singapore.

Debanjan Ghosh, Weiwei Guo, and Smaranda Muresan. 2015. Sarcastic or Not: Word Embeddings to Predict the Literal or Sarcastic Meaning of Words. In *Proceedings of EMNLP*.

Pragglejaz Group. 2007. Chinese and Counterfactual Reasoning. In *MIP: A Method for Identifying Metaphorically Used Words in Discourse. Metaphor and Symbol*, Vol. 22 (1): 1-39.

Chu-Ren Huang. 2019. Double Meaning and Reversal: Toward an empirical linguistic account of irony. In 2*019 Joint Conference of Linguistic Societies in Korea The 26th Joint Workshop on Linguistics and Language Processing* (JWLLP-26), Seoul, Korea.

Yan Jiang. 2019. Chinese and Counterfactual Reasoning. In Chu-Ren, H., Barbara, M., and Zhuo, J.-S. (eds.):*The Routledge Handbook of Chinese Applied Linguistics*. Routledge, Abingdon.

Aditya Joshi, Pushpak Bhattacharyya, and Mark J Carman. 2017. Automatic Sarcasm Detection: A Survey. In *ACM Computing Surveys*, vol. 50(5).

Shu-Xiang Lv, Sheng-Shu Ding (eds.) 吕叔湘, 丁声树 编撰. 2016. Modern Chinese Dictionary 现代汉语词典. The Commercial Press 商务印书馆, Beijing, China.

Antonio Reyes and Paolo Rosso. 2014. On the Difficulty of Automatically Detecting Irony: Beyond a Simple Case of Negation. In *Knowledge and Information Systems*, vol. 40(3): 595–614. Springer.

Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna A. Kaal, Tina Krennmayr and Trijntje Pasma. 2010. A Method for Linguistic Metaphor Identification: From MIP to MIPVU. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Yi-Jie Tang and Hsin-Hsi Chen. 2014. Chinese Irony Corpus Construction and Ironic Structure Analysis. In *Proceedings of COLING*.

Zhongqing Wang, Yue Zhang, Sophia Yat Mei Lee, Shoushan Li and Guodong Zhou. 2016. A bilingual attention network for code-switched emotion prediction. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 1624-1634.

Frank Z Xing and Yang Xu. 2015. A Logistic Regression Model of Irony Detection in Chinese Internet Texts. In *Research in Computing Science*, no. 90: 239–249.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of Thirty-Second AAAI Conference on Artificial Intelligence*.