Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5034–5039
Marseille, 11–16 May 2020

The Design and Construction of a Chinese Sarcasm Dataset

Xiaochang Gong[1], Qin Zhao[1], Jipeng Wu[1], Jun Zhang[2], Ruibin Mao[2], Ruifeng Xu[1,3]
[1]Harbin Institute of Technology (Shenzhen), China, [2]Shenzhen Securities Information Co.,Ltd, China
[3]Joint Lab of Harbin Institute of Technology and RICOH
newbiejailer@gmail.com, zhaoqin@hit.edu.cn, wujipeng@stu.hit.edu.cn
{zhangjun, maoruibin}@cninfo.com.cn, xuruifeng@hit.edu.cn

Abstract

As a typical multi-layered semi-conscious language phenomenon, sarcasm is widely existed in social media text for enhancing the emotion expression. Thus, the detection and processing of sarcasm is important to social media analysis.However, most existing sarcasm dataset are in English and there is still a lack of authoritative Chinese sarcasm dataset. In this paper, we presents the design and construction of a largest high-quality Chinese sarcasm dataset, which contains 2,486 manual annotated sarcastic texts and 89,296 non-sarcastic texts. Furthermore, a balanced dataset through elaborately sampling the same amount non-sarcastic texts for training sarcasm classifier. Using the dataset as the benchmark, some sarcasm classification methods are evaluated.

Keywords: sarcasm dataset, corpus design, Chinese

1. Introduction

Sarcasm is a typical multi-layered semi-conscious language phenomenon. Essentially, there are a dual purpose expression. That is, the meaning of what a speaker wants to express is very different from the superficial meaning of what he/she says, and even in most cases the two meanings are completely opposite. Because of sarcasm's unique language effect, it is widely used by users in internet applications such as social media and forums (Maynard and Greenwood, 2014). When users express their emotions through sarcasm, they tend to express the opposite of the emotional tendency that they want to express which always puzzle the sentiment analysis algorithms (Pang et al., 2008). Thus, the study on sarcasm detection and processing is important to improve the performance of text emotion analysis, question answering system and conversation robot. Currently, most existing sarcasm annotation corpus in on English text but few on Chinese, which is a barrier to sracasm detection research on Chinese (Walker et al., 2012; Joshi et al., 2015; Oraby et al., 2016; Khodak et al., 2018).

In this paper, we present the work on designing and constructing a large high-quality Chinese sarcasm dataset. The raw text are collected from the user comments text from a news sites. We construct a balanced annotated dataset, which contains 2,486 sarcastic texts and 2,486 non-sarcastic texts, and an unbalanced dataset which contains 2,486 sarcastic texts and 89,296 non-sarcastic texts. Based on the constructed dataset, the performance of some existing sarcasm classification methods are evaluated.

The rest of this paper is summarized as follows. Section 2. briefly reviews the existing Chinese sarcasm dataset and related work. The definition of sarcasm and the design issues of the sarcasm corpus are presented in Section 3.. Section 4. presents the workflow and detailed our annotation process. The statistics of constructed corpus is presented in Section 5.. In Section 6., we simply evaluate some sarcasm classification models by using the constructed dataset as the benchmark data. Finally, Section 7. concludes.

2. Related Work

Tang and Chen (2014) collected sarcastic texts from Weibo by using the emoji as clue and using linguistic features and sentiment determination for identifying sarcastic text. The the language structure and sarcastic elements were analyzed and annotated. Liu et al. (2014) constructed three unbalanced dataset based on sarcastic data from Sina Weibo, Tencent Weibo and Netease Forum, respectively. They also proposed a multi-strategy integrated learning method to solve the data imbalance problem in sarcasm detection. Lin and Hsieh (2016) constructed a dataset based on the assumption that positive sentimental comments on negative issues is highly likely to has sarcasm. Through the Gossiping section of PTT, they semi-automatically constructed a dataset consists of 17,256 sarcastic comments and 9,373 non-sarcastic comments. Based on the corpus data, the performances of of the sarcasm detection methods based on Support Vector Machine with naïve features and Convolutional Neural Network models were evaluated. Sun et al. (2016) constructed a sarcastic corpora with 1,030 documents from Sina Weibo through manually annotation, plus 1,000 sarcasm documents from Tang and Chen (2014) and random sampled 1,000 non-sarcasm documents from Weibo. Finally, a dataset containing 3,030 documents are constructed. The effectiveness of the fusion of convolutional neural network and LSTM sequence neural network model on the sarcasm detection task were evaluated. Lu et al. (2019) collected 200,000 Weibos data and constructed a dataset containing 2,398 sarcastic documents and 2,398 non-sarcastic documents through manual annotation. Based on the dataset, an sarcasm detection method based on Convolutional Neural Networks with linguistic features was proposed.

3. Corpus Design

3.1. Problem Definition

Sarcasm is a type of complicated language phenomenon. We summarize the characterises of Chinese sarcasm into the following two points: first, there is an opposite relation-

Corresponding Author: xuruifeng@hit.edu.cn ship between the literal meaning and implicit meaning of the text, such as "五个国家一起围剿一个公司，真有出息啊。" (*It is really promising for five countries to against a company together.*) which literally praise the five countries' behavior, but whose implicit meaning is that the five countries is bullying the company. If we change above text to "一个公司对抗五个国家，真有出息啊。" (*"It is really promising for a company to against five countries."*, then there is absolutely no sarcastic meaning in it. Second, sarcasm is directional and aggressive. The target of sarcasm should be person, organization, country, etc. In the previous example, the target is "五个国家" (*"the five countries"*. Besides, we need to take sentence as a whole when judge it is or not. For example, "这么大项目才投资5000万美元，够买一个杯子不？" (*Only $50 million is invested in such a large project. Is it enough to buy a cup?*) Although the last part is ambiguous, the first part is a straight negative expression. So it is a non-sarcastic sentence.

Accurate detection of sarcasm requires a wealth of information, including context and background knowledge (Hazarika et al., 2018). For instance, "这个业务水平！牛逼！" (*Very professional! Awsome!*) This sentence is a plain positive expression if we do not take its context into consideration. However, when given the context, which is "保胎药开成打胎药，妈妈胎儿不保，"医生回复'笔误'" (*The fetal-protection medicine was prescribed as an abortion medicine, which leads to a mother losing her fetus. The doctor responded with a "clerical error"*), the above sentence is definitely sarcastic now. Most existing Chinese sarcasm dataset is based on Weibo data, which is a relatively free medium without contextual information and related background knowledge. To solve this problem, in this study, we collect both the target text and its context/background text at the same time during dataset construction.

3.2. Data Collection

To build a Chinese sarcasm data set, the first step is to choosing the proper raw text. Considering the openness of the data and the cost of labeling, etc., the candidate raw text need to meet following requirements:

1. The data must be open and easy to collect. Since the collected data is only used for academic research which is open and shared, the data of the collected objects must be open and can be legally reused.

2. The data should be mainly short texts. The study of Chinese sarcasm detection is in its infancy, and the sarcasm of this linguistic phenomenon usually only appears in a clause in a sentence or even a semantic segment composed of several words. In order to start from the most basic and core issues, we consider the study object as a form of data dominated by short texts. Long texts always introduce unnecessary noise into the problem study, and weaken the effect of the true sarcastic clauses or semantic fragments in the sentence on the final recognition.

3. The proportion of sarcastic texts in whole data must be relatively high. As we know, sarcasm, as a special linguistic phenomenon, usually with some aggressiveness and occurs in a specific context, those factors make it not appear in high proportions in general text (Wallace et al., 2014). The anonymity of the network and the enthusiasm of discussion on social hot issues leads to frequent conflicts of different opinions, which makes the probability of sarcasm increases on social medias. But in general, the proportion of sarcastic data is normally small which brings much difficulty to sarcasm corpus construction. To save the cost for corpus construction, different collection of text objects are observed and compared. The data sources with relatively high proportion of sarcastic texts are selected for further annotation.

To meet the above requirements, the user comments text from the news website is regarded as good candidate. Firstly, the news texts of the major news websites and their corresponding user comment texts are all open. Since the relevant texts in the news websites do not involve sensitive information or commercial information, the difficulty for raw text collection is relatively low. In addition, news are very time-sensitive and always focus on current hot topics, which increase readers' enthusiasm for expressing their opinions. Furthermore, the possibility of collision between different opinions from different readers is high, which increases the probability of the occurrence of sarcastic comment.

Guanchazhe[1] is a news and commentary integration website, which integrates news communication, humanities and social science research, reflecting the current confrontation between various trends of thought in China and the world. It focusing on various comments to international and domestic issues from inside and outside China. The website has the characteristics of fast update of news content, plenty of active users, abundant user comments on news events, and active discussion between users. These characteristics met the above requirements for building the sarcasm corpus. So we finally choose the user comment texts of Guanchazhe as the annotation target while the corresponding original news text is regarded as background text.

The raw data is divided into two parts: news report related data and user's comment related data.

- News related data is collected with following fourteen attributes: news subject content, news category, news ID, news keyword, news title, news source, news tag, author avatar, author name, author's personal homepage link, author title, name of the editor, the time of the press release, and the last update time. Table 1 shows an example of news related data. It is worth noting that if some old news' content is mentioned by current news, then the related old news will be linked to current news through internal links of the website. Obviously, such relationship is directional, that is, it only points to the old news from the latest news. This relationship constitutes a tree-like structure, in which the root node is the latest news and the leaf nodes are those related old news. During data collection, such tree-like

[1] https://www.guancha.cn/

Field	Value
doc_id	397130
title	台军役男"享福" 只需服4个月的兵役还能分2年服役(*Male Taiwanese enjoy military benefits: only needs to complete 4 months of military service within 2 years*)
category	military affairs
source	观察者网(*Guanchazhe*)
tags	台军，台湾(*Taiwan army, Taiwan*)
keywords	nan
author_name	于文凯(*Kaiwen Yu*)
body	台军实行募兵制后，其他适龄役男只需要参加4个月军事训练就算服役。今日台"国防部"宣布了一个"好消息"，今年的役男可以延续"暑期分阶段"便民政策，就是说4个月的服役可以拆分成2半，每个暑假培训8周，两个暑假后就算完成服役...(*After the Taiwan military implemented a recruiting system: male of the right age only needs to complete four months of military service. Today Taiwan's "Ministry of Defense" announced a "good news". In this year, individuals can continue the "summer stage" convenience policy. That is, divide four months of militate service into two parts. Train in each time for 8 weeks and finish the service in two summer vocations...*)
last_update_time	2019-02-21 22:41:55
release_time	2019-02-21 20:03:32

Table 1: Example of news related data

relationship is retained by news ID, news related keywords and corresponding link to facilitate subsequent research.

- For the user's comment related data, we collect information such as comment text, comment ID, number of replies, number of likes, number of points, user ID, and user's website name. Table 2 shows an example of this type of data. Since there are cases in which users reply to each other in the comment, the relationship between the comment and the reply is retained by the current comment ID and the parent comment ID.

We collected 2,197 news published on Guanchazhe from April to May 2019. The news cover themes such as international, military, financial, economics, technology, automotive. In total 178,237 related user's comments are also collected.

Field	Value
comment_id	12749170
parent_comment_id	0
root_comment_id	0
doc_id	397130
user_id	224423
user_name	渔排守望者(*Fish raft watcher*)
content	很好，很自由，很人性化，很娘娘化，很...，大伙儿再想想很什么谢谢。(*Very good, very free, very humanize, very motherly, very ..., help me to figure out more features. Thanks.*)
reply_num	0
praise_num	5
tread_num	0

Table 2: Example of comment related data

4. Corpus Annotation

In this study, sarcasm detection is regarded as a special binary text classification problem. Therefore, label sarcastic text as 1 and label non-sarcastic text as 0. In order to ensure the quality of annotated data and reduce the deviation of manual judgment, the following guidelines are developed:

1. To determine whether a sentiment expression is sarcastic, the annotator is suggested to pay more attention to the contradiction between the literal meaning of a sentence and its implied actual meaning. For example, "没有纸尿裤可以用，好怕怕哦" (*There is no urinary use, so scare*). The text of this commentary literally expresses a feeling that 好怕怕哦" (*scare*). The implicit meaning is to express "we are not afraid at all". There is a clear contradiction between these two meanings, so this case is annotated as sarcastic.

2. The labeling process only focuses on whether there is sarcasm in a sentence rather than the sentiment polarity of the sentence,because that the sentiment polarity of sarcastic sentence can be either positive or negative, and there is no direct connection between sentiment polarity and sarcasm.

3. To ensure the annotation quality, we prepared several rules for ambiguous data annotation. First, we synthesize the opinions of at least 5 people for the ambiguous cases. Then, we adopt the majority if more than 80 percent people vote for it, and drop the data otherwise. Actually, sarcasm is sensitive to many factors such as context, background knowledge, and thus the ambiguous cases are widely exist.

The collected user comment data comes from the Internet, so the comment text contains some invalid strings such as web links, identifiers, extra spaces. Thus, data cleaning and pre-processing is performed on the comment text. It should be noted that, some of the parts that need to be removed in the conventional text cleaning work are reserved in this topic. Table 3 shows the different usage frequencies of the exclamation mark in sarcastic text and non-sarcastic text. It is shown that the frequency of exclamation marks in sarcastic text is much higher than non-sarcastic text. It indicates that users often express their strong sarcasm emotion through the use of exclamation mark. So, exclamation marks in texts are reserved during text cleaning.

punctuation	sarcasm	non-sarcasm
!	1395	728
!!	153	68
!!!	89	39

Table 3: Exclamation marks in sarcastic/non-sarcastic text

In addition, the sarcastic text comes from many different news, and has a certain tendency to the topic of the news, such as the possibility of sarcastic comments appearing in news related to international topics is relatively high, as shown in Table 5. In order to improve the persuasiveness of this sarcasm dataset, we collect non-sarcastic texts from those news which already contain sarcastic comments, and the number is consistent to the number of sarcastic text. For example, there are 3 sarcastic user comments in news A, then 3 non-sarcastic user comment texts are randomly sampled from news A. In this way, the topic distribution of sarcastic user comment and non-sarcastic user comment is consistent.

5. Corpus Statistics

In order to have a more intuitive understanding of our developed sarcasm dataset, we do some related data statistics work, and also validate the classification effect of some commonly used text classification models on our balanced dataset.

Table 4 shows the comparison of our dataset and exsiting sarcasm corpora including data source, data scale and obtaining method. As one can see, the number of sarcastic data contained in our dataset is the largest among these manually annotated dataset, as well as the size of our unbalanced dataset. In addition, the source of the previous dataset is mostly social media such as Weibo, while our developed dataset comes from the news network. In comparison, our dataset contains more information than plain text, such as news information, news structure information, comments, and so on. Those additional information increases the scalability of future research.

The topic distribution among sarcastic text is shown in Table 5 , one may observe that news related to international affairs, military affairs and politics occupy the most part. Those topics often have hot issues which can cause widespread concern which leads to intense discussion between different users. So, sarcasm is widely used by users to refute other's point of view due to its aggression. Figure 1 shows the length distribution of sarcastic text and non-sarcastic text. It is observed that the number of non-sarcastic text is more than that of satirical text in the length range from 0 to 15, which means that short text normally can not express sarcasm very well. Such differences in distribution shows that sarcasm is a complex form of expression which requires sufficient context information to understand.

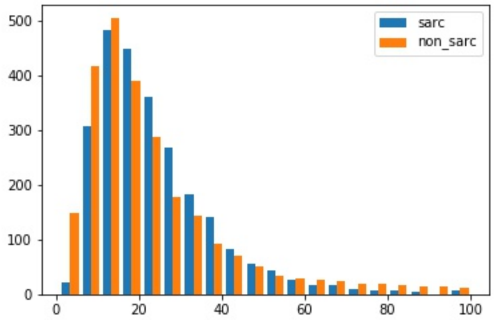

Figure 1: Length distribution of sarcastic text and non-sarcastic text.

6. Evaluation

In this section, several sarcasm detection method based on typical text classification models are evaluated, in order to provide comparable baseline results for future research. The evaluation metrics used to measure the performance of models are accuracy and F1 score.

As for the baseline models, they are chosen as follows:

- **textCNN:** We use textCNN (Kim, 2014) as the baseline model to learn the feature representations from comment text, a softmax layer is used to generate final classification result.
- **LSTM:** Compared to textCNN, long-short term memory (LSTM) (Hochreiter and Schmidhuber, 1997) is expected to learn the longer dependency in sequence for sarcasm detection.
- **textCNN+attention:** Considering that different words within a sentence should have different contribution to the representation of the whole sentence, attention mechanism (Vaswani et al., 2017) is adopted to improve the textCNN model.
- **LSTM+attention:** Regular LSTM also treat every token in sequence equally, so attention mechanism is added to strengthen its performance.
- **BERT:** The pre-trained language model BERT (Devlin et al., 2019) is very popular in recent NLP field due to its powerful semantic encoding, and yield good result in many other tasks.

The achieved performances are listed in Table 6. Due to limitation of the dataset, all results come from average of 10-fold cross validation. It is observed that BERT achieved the highest performance on small dataset which is similar to the previous research by (Sun et al., 2019).

Dataset	Source	Sarcastic	Non-sarcastic	Total	Method
Tang et al. 2014	Sina Weibo	950	0	950	semi-automatic
Liu et al. 2014	Sina Weibo	238	3,621	3,859	manual
	Tencent Weibo	359	5,128	5487	manual
	Netease Forum	546	9,810	10,356	manual
Liu et al. 2016	PTT	17,256	9,373	26,629	semi-automatic
Sun et al. 2016	Sina Weibo	2,000	1,000	3,000	manual
Lu et al. 2019	Sina Weibo	2,398	2,398	4,796	manual
our balanced dataset	Guanchazhe	2,486	2,486	4,972	manual
our whole dataset		2,486	89,296	**91,782**	manual

Table 4: Chinese sarcasm dataset comparison

topic	number	percentage
international affairs	1136	45.70%
military affairs	355	14.28%
politics	243	9.77%
science	172	6.92%
industry	169	6.80%
economy	93	3.74%
others	318	12.79%
total	2486	-

Table 5: Topic distribution of sarcastic texts

Method	Accuracy	F1 Score
textCNN	0.6522	0.6519
LSTM	0.6584	0.6549
textCNN+attention	0.6770	0.6733
LSTM+attention	0.6708	0.6646
BERT	**0.7611**	**0.7368**

Table 6: Performances of sarcasm classification model on our balanced dataset

7. Conclusion

In this study, using the user comments on news website as the candidate raw text and their corresponding news text as the background, we design and development a sarcasm annotated corpus. Up to now, it is the largest high-quality Chinese sarcasm dataset based on manual annotations in world, based on our knowledge. Using the corpus as the benchmark data, several existing sarcasm detection algorithms are evaluated. It is hoped that researchers in related fields can make good use of this dataset to promote the Chinese sarcasm detection research.

8. Acknowledgements

This work was supported by National Natural Science Foundation of China (61632011 and 61876053), Key Technologies Research and Development Program of Shenzhen JSGG20170817140856618, Shenzhen Foundational Research Funding JCYJ20180507183527919 and JCYJ20180507183608379, and Joint Research Program of Shenzhen Securities Information Co., Ltd. JRPSSIC2018001.

9. Bibliographical References

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Hazarika, D., Poria, S., Gorantla, S., Cambria, E., Zimmermann, R., and Mihalcea, R. (2018). Cascade: Contextual sarcasm detection in online discussion forums. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1837–1848.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Joshi, A., Sharma, V., and Bhattacharyya, P. (2015). Harnessing context incongruity for sarcasm detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 757–762.

Khodak, M., Saunshi, N., and Vodrahalli, K. (2018). A large self-annotated corpus for sarcasm. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018).*

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Lin, S.-K. and Hsieh, S.-K. (2016). Sarcasm detection in chinese using a crowdsourced corpus. In *Proceedings of the 28th Conference on Computational Linguistics and Speech Processing (ROCLING 2016)*, pages 299–310.

Liu, P., Chen, W., Ou, G., Wang, T., Yang, D., and Lei, K. (2014). Sarcasm detection in social media based on imbalanced classification. In *International Conference on Web-Age Information Management*, pages 459–471. Springer.

Lu, X., Li, Y., and Wang, S. (2019). Linguistic features enhanced convolutional neural networks for irony recognition. *Journal of Chinese Information Processing*, 33(5):31.

Maynard, D. and Greenwood, M. (2014). Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4238–4243, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Oraby, S., Harrison, V., Reed, L., Hernandez, E., Riloff, E., and Walker, M. (2016). Creating and characterizing a diverse corpus of sarcasm in dialogue. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 31–41.

Pang, B., Lee, L., et al. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.

Sun, X., He, J., and Ren, F. (2016). Pragmatic analysis of irony based on hybrid neural network model with multi-feature. *Journal of Chinese Information Processing*, 30(6):215.

Sun, C., Qiu, X., Xu, Y., and Huang, X. (2019). How to fine-tune bert for text classification? *arXiv preprint arXiv:1905.05583*.

Tang, Y.-j. and Chen, H.-H. (2014). Chinese irony corpus construction and ironic structure analysis. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1269–1278.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Walker, M., Tree, J. F., Anand, P., Abbott, R., and King, J. (2012). A corpus for research on deliberation and debate. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 812–817.

Wallace, B. C., Kertz, L., Charniak, E., et al. (2014). Humans require context to infer ironic intent (so computers probably do, too). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 512–516.