Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 2, No. 1, 2022

Open Chinese Internet Sarcasm Corpus Construction: An Approach

Yizhang Zhu

College of Computer Science, Chongqing University, Chongqing 401331, China
20194113@cqu.edu.cn

Abstract: Sarcasm is a commonly-used language phenomenon particularly on the Internet, which is often to convey criticism or negative emotions. A proper sarcasm corpus to help sarcasm study and detection can contribute to linguistic research and assist sentiment analysis, but an open Chinese corpus is found extremely lacking. In this paper, we referenced existing methods and data and constructed a balanced open Chinese Internet sarcasm corpus in a new approach to improve efficiency and data quality. The balanced open corpus contains multi-source and labeled 2,000 texts selected from bigger corresponding origin datasets. In our corpus, sarcasm and non-sarcasm, longer and shorter texts are both in 1:1 ratio.

Keywords: Sarcasm; Open corpus; Chinese Internet.

1. Introduction

Sarcasm can show opposite meanings or attitudes instead of their literal ones when being expressed. In the Cambridge Dictionary, the explanation of the word "sarcasm" is "the use of remarks that clearly mean the opposite of what they say, made to hurt someone's feelings or to criticize something in a humorous way". In Contemporary Chinese Dictionary, the corresponding word "Fanfeng" is explained as "satire from the opposite side to expose or criticize the bad or stupid, often use metaphors and exaggerations". However, this can also interfere with sentiment recognition and analysis for hiding their real means.

Due to the need for expression on some specific topics or emotions, or to circumvent relatively more stringent censorship mechanisms, sarcasm widely exists on the Internet. Therefore, sarcasm detection can be conducive to tasks requiring capturing people's real sentiments, including sentiment analysis, hot events monitoring, and harassment detection[1]. There have been authoritative and proven datasets for sarcasm detection so far, but most of them are English-based[2][3]. Open Chinese corpus in this field, particularly simplified Chinese corpus, is extremely lacking. Furthermore, although sarcastic messages widely exist in online social media or comment areas, they do not account for a significant proportion[4], which makes conventional data collection methods labor-intensive and time-consuming.

During corpus construction, we proposed an ingenious and more efficient method to assist data collection. We also combined the existing corpus constructed by Tang and Chen[5], and referenced some suggestions on data collection from Li and Huang[4]. In this paper, an open balanced Chinese Internet sarcasm corpus is constructed, containing 500 longer and 500 shorter sarcastic texts, 500 longer and 500 shorter non-sarcastic texts respectively, 2,000 labeled texts in total.

2. Related Work

As a linguistic phenomenon, sarcasm is studied by literary and linguistic or Natural Language Processing scholars from many aspects. To assist and help further study, sarcasm corpora are constructed, most of which are English-based. News Headlines dataset for Sarcasm Detection introduced by Misra et al. is collected from The Onion (an American website particularly producing sarcastic news) for sarcastic news headlines and HuffPost for non-sarcastic ones[6]. Khodak et al. constructed the Self-Annotated Reddit Corpus (SARS), which is a much larger corpus containing 1.3 million labeled comments scraped from Reddit through the use of the tag "\s" which implies sarcastic meaning in Reddit[2].

As for Chinese corpus, they are much smaller in number and scale compared with their English counterparts. Tang and Chen collected 1,005 microblogs from the Plurk website and formed the earliest open Chinese corpus (unsimplified Chinese)[5]. Based on it, Sun and He expanded the dataset through a semi-automatic method: using existing sarcasm detection algorithms roughly screen out "quasi-sarcastic" texts then selecting manually to ensure correctness[7]. Gong et al. constructed a corpus containing 2,486 manually selected and annotated sarcastic texts collected from Guanchazhe which means observers, a Chinese integration website of news and commentary. However, these corpora are non-open sources[3]. Lu et al. constructed a balanced corpus containing 2,348 manually selected sarcastic and non-sarcastic texts from Sina Weibo respectively[8]. Partial data was mentioned to be opened, but the link has been no longer accessible.

Li and Huang tried to analyze cues of sarcasm at grammatical and semantical levels[4]. According to their statistics, the average proportion of sarcastic messages existing on Sina Weibo is less than 1% which makes it difficult and inefficient to collect data if aimless. They also summarized formalized features of sarcastic expressions based on linguistic analysis and designed an Irony Identification Procedure (IIP) to help detection.

3. Proposed Corpus Construction Methods

3.1. Principles of Collection

Li and Huang's work illustrated the features of sarcastic expressions' existence on the Internet: high volume and low proportion[4]. Therefore, aimless collection can make the corpus construction procedure time-consuming and labor-

intensive. However, to ensure the accuracy and correctness of corpus materials and their labels, corpus construction should base on manual selection. To improve the efficiency of data collection and corpus construction, there are 2 principles to be cleared and followed in this paper.

Combine with existing high-quality open corpus. Because of the low proportion, large-scale sarcastic texts are difficult to collect but important for research. To enlarge the corpus as much as possible without compromising quality, we processed irony corpus data collected by Tang and Chen[5], and add it to the corpus we constructed to avoid the reinvention of the wheels.

Select materials in “high-density” places. Another way to improve efficiency is to select sarcastic or non-sarcastic materials where they are more likely to appear and gather together, which are so-called “high-density places”, so that time of collection can be reduced and data labeling can be saved. In Li and Huang’s work, they chose materials already known or marked as ironic[4], which is similar to Misra et al. who collected English sarcastic materials from The Onion[6]. It is an inspiring method, and we referenced their idea and took a step further to apply it not only in the sarcasm area but non-sarcasm area.

3.2. Sarcasm Corpus Collection

The sarcastic corpus has two sources: the corpus constructed by Tang and Chen[5] and sarcastic messages from Sina Weibo.

The corpus constructed by Tang and Chen contains 1,005 labeled ironic texts of unsimplified Chinese collected from the Plurk website which is a social networking site[5]. In this corpus, Tang and Chen added corresponding sarcastic tags for each sentence element based on their linguistic summarization. Here are two examples:

<message><rhetoric> 很 </rhetoric><ironic sentiment="pos">好</ironic><context sentiment="neg">又失眠了</context><rhetoric>!!</rhetoric> :-(</message>

·Awesome! I’m having insomnia again!! :-(

<message><context sentiment="neg"> 能 夠 天 天 出 錯 </context> 也 是 一 件 <ironic sentiment="pos"> 了 不 起 </ironic>的事<rhetoric>!!!</rhetoric> (annoyed)</message>

To be able to make mistakes every day is also quite impressive!!!

In the corpus constructed by Tang and Chen, the texts are unsimplified Chinese and these sentence element labels are not necessary for our study, therefore, the data needs to be further processed to suit our needs. Furthermore, in this corpus, Plurk microblogs which are relatively shorter. To construct a balanced corpus, we introduced another source to collect longer sarcastic materials.

Similar to The Onion in America, on the Chinese Internet, there is a user account on Sina Weibo called "Yangcong Gushihui (Onion The Storyteller)", providing sarcastic news and commentary in simplified Chinese and most of them are relatively longer texts. Not all of the materials provided by this account are suitable and usable for sarcasm corpus construction, there are also advertisements and raffle information, but the proportion of sarcastic text content is far higher than the platform-wide average, which is particularly what we called "high-density" places. To construct our corpus, we manually select 500 sarcastic texts provided by Onion the Storyteller. Here is an example from Onion the Storyteller:

Because there are always students who do not take their studies seriously and dream of becoming a star one day, the star dissuasion business came into being. A local celebrity education institution specializes in recruiting such students to practice singing, dancing, and acting according to the standards of celebrities, and they generally respond that it is not difficult. The institution will even regularly invite first-line stars to practice with the students, trying to make them feel the gap between professional and non-professional. The results of the students found that professional is indeed far worse than non-professional, so the desire to become a star is stronger, and the business ended in failure.

Compared with the corpus constructed by Tang and Chen, most sarcastic materials collected from Onion the Storyteller are relatively longer, containing more contextual information which will allow for sarcasm detection that is not limited to sentence components and corpus balance.

3.3. Non-sarcasm Corpus Collection

Similar to the methods we collected sarcasm corpus materials, we selected "high-density places" to collect non-sarcasm corpus, and the balance of short and long texts is also considered.

As for longer non-sarcastic materials, formal official news can be practical. We crawled messages from the official Weibo account of Renmin Daily and Xinhua News Agency as the original dataset to be processed. Renmin Daily is an official and authoritative newspaper in China, and Xinhua News Agency is China’s state and worldwide news agency. Their reports and commentaries are serious no matter pro or con standpoints, and hardly sarcastic or ironic, which are suitable as non-sarcastic materials. Here is an example from Renmin Daily:

Hello, tomorrow. The survey shows that more than 80% of young people believe that they can be the master of their own emotions. On the road of life, there are slow currents and dangerous beaches, and there are suns and storms. To harness emotions instead of being harnessed, and to defeat difficulties instead of being defeated, is the only way to have a bright and cheerful life. Praise for the robust self-confidence of young people, but also to listen to their voices, help them face setbacks, resolve their annoyance, and embrace a better life.

To construct a balanced corpus, we need shorter non-sarcastic as well. However, if aimlessly crawl online messages, widely existing sarcastic messages are not neglectable and can affect data quality even though they account for a small percentage.

Therefore, we turned to the comment section of the official agencies’ WeChat public account, which are equipped with two main features: 1) most comments are short; 2) because of the given function of the WeChat App itself, every comment for an article of any public account will be filtered and vetted by author or admin before being displayed to the public. We selected two official public accounts where the comments have been stricter filtered in the past: Renmin Daily, Central Committee of the Communist Youth League of China (Chinese Youth League). It is reasonable to assume that comments from the sections mentioned above qualify as short non-sarcastic materials as the authors had manually screened out inappropriate comments for us in advance. Here are several examples:

Healthy diet and exercise routine should not be neglected!!!! ∑(Д)

The Internet is not a lawless area. Disinformation must be punished!

Only when I went to college did I realize that what I missed the most was high school, the unforgettable part of my youth.

Remarkably, given that there is still quite a few unhelpful and invalid information existing in comment sections where we collected data and mentioned above like any other online platforms, such as too short like "ok/good" or meaningless like "Uh…….". Therefore, at this stage, manual selection is still a necessity.

3.4. Data Processing and Corpus Construction

According to the methods mentioned above, table 1 is the summary table of corpus data sources, and table 2 illustrates the scales and collecting methods of origin data.

To construct a properly balanced corpus, we processed origin data as follows:

Delete the sentence element labels in the corpus constructed by Tang and Chen, and transform them into simplified Chinese, in order to maintain uniformity of text type in the whole corpus.

Unify the format of the original data. Our data are multi-sourced, so in the corpus, the data format needs to be unified. In our corpus, each piece of data has two attributes: label and text. The label indicates sarcasm (label: 1) or non-sarcasm (label: 0), text records the content, as table 3.

Remove irrelevant text features. Materials from the different online platforms may have specific content-irrelevant features. For example, on Sina Weibo, official media accounts tend to use "[]" or "##" to mark out headlines and topics which is very rare in Onion the Storyteller. Irrelevant features can have negative effects on feature extraction and interfere with text classification tasks.

Randomly select to construct a balanced corpus. We want in our final corpus, sarcasm and non-sarcasm, longer and shorter texts are both in 1:1 ratio, so in this step, we randomly select texts as table 4 illustrates.

Shuffle and adjust. Randomly disrupted the dataset and make final adjustments.

Table 1. Summary of Data Sources

	Sarcasm	Non-sarcasm
Longer	Onion the Storyteller (Weibo account)	① Renmin Daily (Weibo account) ② Xinhua News Agency (Weibo account)
Shorter	Corpus constructed by Tang and Chen[5]	① Renmin Daily (WeChat comment section) ② Chinese Youth League (WeChat comment section)

Table 2. Summary of Origin Data Information

Data source	Scale (texts account)	Collect
Onion the Storyteller	500	Crawl and Manual
Renmin Daily (Weibo)	930	Crawl
Xinhua News Agency (Weibo)	1001	Crawl
Corpus constructed by Tang and Chen[5]	1005	--
Renmin Daily (WeChat)	250	Manual
Chinese Youth League (WeChat)	250	Manual

Table 3. Example of Data Format

Label	Text
0	(In translation) Embarrassing is as one feels embarrassed!
0	(In translation) I remember an aunt stole and sold a sack of textbooks that we had packed before the college entrance examination.
…	……

Table 4. Random Selection

Data source	Origin Scale	Random Selection
Onion the Storyteller	500	All
Renmin Daily (Weibo)	930	250
Xinhua News Agency (Weibo)	1001	250
Corpus constructed by Tang and Chen	1005	500
Renmin Daily (WeChat)	250	All
Chinese Youth League (WeChat)	250	All

4. Conclusion

We constructed a balanced Chinese Internet Sarcasm Corpus with 2,000 multi-sourced labeled text. Our corpus balanced sarcasm and non-sarcasm, longer and shorter materials, both in 1:1 ratio, and we opened the source on Github to make it available to anyone. The open source repository is https://github.com/Yizhang-Zhu/open-Chinese-Internet-Sarcasm-Corpus. Hopefully, our corpus can help and facilitate Chinese sarcasm detection in further study.

References

[1] Cai, Y., Cai, H., & Wan, X. (2019, July). Multi-modal sarcasm detection in twitter with hierarchical fusion model. In Proceedings of the 57th annual meeting of the association for computational linguistics (pp. 2506-2515).

[2] Khodak, M., Saunshi, N., & Vodrahalli, K. (2017). A large self-annotated corpus for sarcasm. arXiv preprint arXiv:1704.05579.

[3] Gong, X., Zhao, Q., Zhang, J., Mao, R., & Xu, R. (2020, May). The design and construction of a Chinese sarcasm dataset. In Proceedings of the 12th Language Resources and Evaluation Conference (pp. 5034-5039).

[4] Li, A. R., & Huang, C. R. (2020). A Method of Modern Chinese Irony Detection. In From Minimal Contrast to Meaning Construct (pp. 273-288). Springer, Singapore.

[5] Tang, Y. J., & Chen, H. H. (2014, August). Chinese irony corpus construction and ironic structure analysis. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers (pp. 1269-1278).

[6] Misra, R., & Arora, P. (2019). Sarcasm detection using hybrid neural network. arXiv preprint arXiv:1908.07414.

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

[8] Lu, X., Li. Y., & Wang S. (2019). Linguistic Features Enhanced Convolutional Neural Networks for Irony Recognition. Journal of Chinese Information Processing, 33(5), 31-38.