Telling the Whole Story: A Manually Annotated Chinese Dataset for the Analysis of Humor in Jokes

Dongyu Zhang[1], **Heting Zhang**[1], **Xikai Liu**[2], **Hongfei Lin**[2], **Feng Xia**[1*]
[1]School of Software, Dalian University of Technology, 116620, China
[2]School of Computer Science and Technology, Dalian University of Technology,
116024, China
f.xia@ieee.org
hflin@dlut.edu.cn

Abstract

Humor plays important role in human communication, which makes it important problem for natural language processing. Prior work on the analysis of humor focuses on whether text is humorous or not, or the degree of funniness, but this is insufficient to explain why it is funny. We therefore create a dataset on humor with 9,123 manually annotated jokes in Chinese. We propose a novel annotation scheme to give scenarios of how humor arises in text. Specifically, our annotations of linguistic humor not only contain the degree of funniness, like previous work, but they also contain key words that trigger humor as well as character relationship, scene, and humor categories. We report reasonable agreement between annotators. We also conduct an analysis and exploration of the dataset. To the best of our knowledge, we are the first to approach humor annotation for exploring the underlying mechanism of the use of humor, which may contribute to a significantly deeper analysis of humor. We also contribute with a scarce and valuable dataset, which we will release publicly.

1 Introduction

Humor plays important role in human communication, which not only serves to exchange ideas or convey messages, but also involves emotion regulation such as provoking laughter, generating amusement, and reducing stress (Wooten, 1996; Morse, 2007). In particular, with the rapid growth of social media applications such as Facebook and Twitter, a significantly increasing number of individuals are using these social media public platforms to release humorous texts. Humor often arises when two incongruous concepts are applied and examined through one semantic frame (Lefcourt, 2001; Paulos, 2008). The two concepts often involve semantic disconnection in forms such as contradiction and contrast/comparison. Humor sometimes occurs due to ambiguity (Yang et al., 2015), such as unexpected homophones/homographs.

The importance and complexity of humor has thus gained attention in natural language processing (NLP), and many computational approaches to it have been proposed (Binsted et al., 2006; Yang et al., 2015; Baziotis et al., 2017; Ortega-Bueno et al., 2018; Liu et al., 2018). Corpora are fundamental in NLP for sound analysis of humor and for high-quality automatic humor detection. Scholars have been devoted to the study of the humor resources in both English and other languages. Mihalcea and Strapparava (2005) constructed a humorous witticism dataset with 16,000 text data for humor identification in English sentences. The dataset comes from one-liners, reuters titles, BNC sentences, and proverbs, and was annotated with humor and non-humor. Reyes et al. (2013) established an English irony dataset of 40,000 tweets for conducting the study of irony on tweets. The dataset contains the label of irony and other specific hashtags of non-irony (education, humor, and politics). Zhang and Liu (2014) established an English humor corpus with 3,000 tweets to recognize humor on Twitter. The dataset contains the annotation of humorous tweets, non-humorous tweets and humorous non-tweets. Potash et al. (2017) built a 12,734 tweets dataset of English for studying the comparative ranking of humor. The dataset comes from the midnight TV program called Hashtag Wars which published on Twitter. Castro et al. (2017) established a humorous text corpus containing 33,531 tweets for detecting humor in Spanish Tweets. The dataset involves humorous annotation and humor level annotation. The humor level annotation is based on a 5-point scale, 1 signifying the lowest level and 5 signifying the highest level. Castro et al. (2018) revised

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing
and the 9th International Joint Conference on Natural Language Processing, pages 6402–6407,
Hong Kong, China, November 3–7, 2019. ©2019 Association for Computational Linguistics

the Spanish Twitter corpus with crowd notes and presented a 27,000 tweets dataset in total. Specially, the authors used 5 different emojis to represent the 5 degrees of humor instead of using the 5-point annotation.

However, while previous work focuses on textual humor annotation of humorous/non-humorous and degree of funniness, such annotations do not provide adequate knowledge and scenarios to explain how humor arises, so they may not provide a deep analysis of the underlying mechanism of humor. In addition, as the majority of data came from Twitter, the data source lacks variety. To this end, we create a dataset on humor with 9,123 manually annotated jokes in Chinese. We propose a novel scheme with annotations of key words that make text humorous as well as character relationship, scene, humor category, and degree of funniness. The annotation agreement analyses for multiple annotators are described. We also conduct analysis and exploration on the dataset. Our contributions are as follows.

- We propose a novel annotation scheme to explain how humor arises in text. Unlike previous work, we annotate not only what is humorous, but also what causes humor.
- We contribute to a new, sizeable, and scarce joke dataset, which is being released publicly and particularly valuable in languages other than English.

2 Data Collection

To make the dataset objective and comprehensive, we collected joke data involving both diachronic and synchronic relationships simultaneously from a variety of fields. Also, We selected jokes based on a four-dimensional model. On the time axis, our dataset includes jokes from books, literary journals, etc. published over the past decade, which satisfies the diachronic requirement. It also includes jokes posted on websites and microblogs, many of which are novel, which conforms to the synchronic requirement. On the spatial axis, the dataset contains both domestic and translated foreign jokes. On the subject axis, the perception of intensity of jokes also varies from person to person, due to their varying backgrounds and senses of humor. On the style axis, jokes from books have various themes, and they are relatively canonical, while online jokes seem more oral and informal. The source information is in Table 1.

Sources	Words	Sentences	Jokes
Websites	2,397,816	23,508	5,463
Micro-blogs	1,207,856	11,614	2,581
Books,Journals	504,920	4,855	1,079
Total	4,110,592	39,977	9,123

Table 1: Information on data sources.

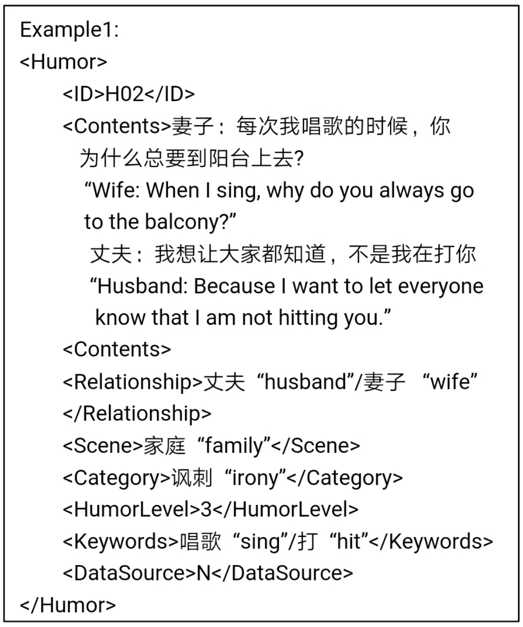
3 Annotation Scheme

3.1 Annotation model

The annotation model is as follows:

JokeModel = (Relationship, Scene, Category, HumorLevel, Keyword, DataSource)

- Relationship: We annotated the mutual relationship between the main characters such as teacher-student, doctor-patient, lovers, superior and subordinate, etc. because accessing the relationship between people in jokes is helpful for a clearer understanding of the contextual coordinates on the joke (Popa, 2005).
- Scene: The scene refers to the place where the joke occurs. Previous studies indicate humor plays an important role in the places including campus (Morrison et al., 2012), workplace (Blumenfeld and Alpern, 1994), family (Lovorn, 2008) and public space (Thornton, 2007). We therefore selected the campus, workplace, family and public space for the annotation of scene in humor.
- Category: There is no consensus on the category of humor in the literature. Based on our investigation of a wide range of literature, we focused on eight main types of the most frequently appearing humor including homophonic, harmonic, antiphrasis, analogy, euphemism, irony, exaggeration, and reversal.
- Level: We weighed the fine-grained annotation of humor (Bressler and Balshine, 2006; Castro et al., 2018; Deckers and Devine, 1981) and presented a 5 point-scale for humor rating, 1 signifying the non-humor, 2-5 signifying the gradually increasing degree of humor. (Gan, 2015; Stein, 1998; Tsakona, 2009)
- DataSource: Table 1 presents the source of jokes.
- Keyword: We define the key words as words that trigger humor and that may have conflicting, incongruous and ambiguous meanings in jokes (Van Hee et al., 2016). Van Hee et al. (2016) proposed the annotation item based on the text spans of contrasting which contains the type of explicit



```
Example1:
<Humor>
      <ID>H02</ID>
      <Contents>妻子：每次我唱歌的时候，你
        为什么总要到阳台上去?
         "Wife: When I sing, why do you always go
         to the balcony?"
          丈夫：我想让大家都知道，不是我在打你
          "Husband: Because I want to let everyone
           know that I am not hitting you."
      <Contents>
      <Relationship>丈夫 "husband"/妻子 "wife"
      </Relationship>
      <Scene>家庭 "family"</Scene>
      <Category>讽刺 "irony"</Category>
      <HumorLevel>3</HumorLevel>
      <Keywords>唱歌 "sing"/打 "hit"</Keywords>
      <DataSource>N</DataSource>
</Humor>
```

Figure 1: An example of annotation

and implicit. The special word pair in the contrasting text spans triggers the production of humor. Specially, we annotated the keyword based on the thought of contrasting text spans in humor in the format of prototype. An annotation example shows in figure 1.

3.2 Keyword annotation

The keyword is the most challenging of the six annotating items. Following Van Hee et al. (2016), our annotation of key words is at the relation level, which involves the identification of incongruous or ambiguous vocabulary, resulting in a comic effect. To discriminate key words, the annotators followed the below guidelines:

• Read the entire text-discourse to establish a general understanding of the meaning.

• For each word in the text, establish its meaning in context.

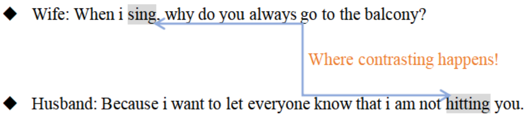
• Determine which words have the meaning of incongruous/conflicting/ambiguous/unexpected or strong emotions that make text humorous in the given context.

• Decide whether the contextual meaning can be understood.

• If yes, mark the word as a key word.

Figure 2 shows the example of keyword annotation in humor.

The words "sing" and "hit" act as keywords, because the two phrases are the main indicators of humor: "she sings badly, and it sounds like she is being beaten and screaming." The comparison of "sing" and "hit" invokes the humor, so they are keywords.



◆ Wife: When i sing, why do you always go to the balcony?

◆ Husband: Because i want to let everyone know that i am not hitting you.

Figure 2: An example of keyword annotation

3.3 Annotation process

Eight postgraduate students and one PhD student worked together to complete the annotation of the joke dataset. The participants were divided into four groups of two. Each group annotated the jokes using cross-validation. The PhD student arbitrated. During the annotation process, when two people reached agreement on the annotation result, then the marking was complete; when there was disagreement, the arbitrator attempted to resolve it. When the arbitration was inconsistent with the views of the two persons' judgment: Case 1. If the inconsistency was in the degree of humor, we used the average value of the three people. Case 2. If there was any disagreement about the generation mechanism, it was discussed by the whole group of nine people, and the mechanism receiving the largest number of votes was the final result.

3.4 Annotation agreement and challenges

To evaluate inter-annotator agreement, we let three annotators annotate the same 600 sentences to assess inter-annotator agreement. We used Fleiss' s kappa (Fleiss, 1971). The agreement on the relationship annotation was κ = 0.85; the agreement on the scene annotation was κ = 0.79; the agreement on the category annotation was κ = 0.71, the

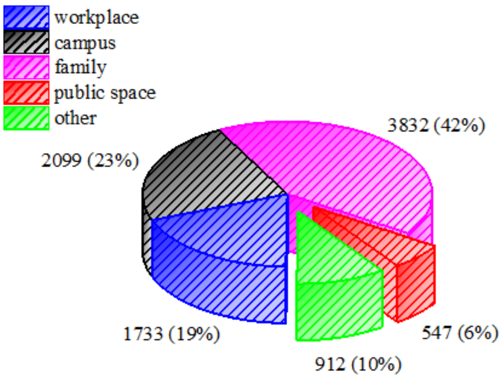


Figure 3: Quantity of joke scenes

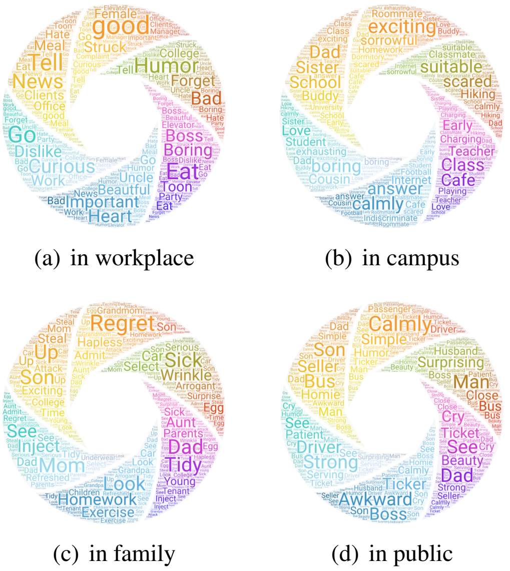
agreement on the humor level annotation was $\kappa = 0.65$, and the agreement on the keywords annotation was $\kappa = 0.59$.

The keywords and humor level annotation was the most challenging part of the annotation, due to the subjective nature of cognition and the different background knowledge of people. To minimize the problems annotators faced, we held a seminar once a week to discuss the ambiguities. Then, the guide gave an authoritative explanation. Finally, ambiguity points and measures were added into the annotating guide manual to help annotators to make judgments quickly and correctly when they encountered the same problem.

4 Dataset Analysis

The dataset contains 9,123 jokes, 39,977 sentences and 4,110,592 words in total, with an average of 4.38 sentences per joke. In the dataset, there are five scenes of joke (where the joke occurs): workplace, campus, family, public space and others. Family accounts for 42% of the dataset. This is perhaps because family life accounts for the largest proportion of life as a whole. These statistical data fully confirm that the jokes originate from life, and that they are known to the general public, which shows that the audience for jokes is very wide. The specific distribution is in Figure 3.

Figure 4 shows the vocabulary that appears most frequently in each scene of joke. As Figure 4 shows, the high-frequency words for each scene of jokes are various. For instance, the top five high-frequency words in campus jokes are ”school, class, exciting, suitable, answer”; in family jokes are ”son, mom, regret, homework, see”; in workplace jokes are ”boring, boss, curious, forget, tell”; in public space jokes are ”man, bus, surprising, cry, calmly”. It is intriguing that some high-frequency words in certain jokes are related to certain scenes.



(a) in workplace

(b) in campus

(c) in family

(d) in public

Figure 4: Word cloud in each scene (translated from Chinese)

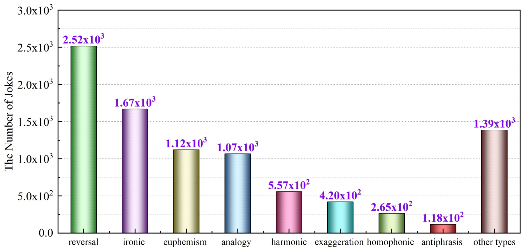


Figure 5: Humor categories

For instance, the high-frequency words in workplace jokes are ”boss, boring, forget, which not only is in line with the bias of the work place, but also proves the validity of this classification to some extent.

We also analyzed the humor categories because they may associate with underlying mechanism of the use of humor. The quantitative statistics for humor categories are shown in Figure 5.

Our annotation not only contains the degree of funniness, but also key words that trigger humor, as well as character relationship, scene, and humor categories. Specially, we have improved on the study of Castro et al. (2018) by providing evidence of what causes humor and explaining how humor arises in text. Furthermore, our data have come from a range of sources in numerous domains rather than only from Twitter.

5 Conclusion

We propose a novel annotation scheme to explain how humor arises in text. Unlike previous work, we annotate not only what is humorous, but also what causes humor. Our dataset creation involved nine volunteer students for 8 months. We will release the dataset publicly. With 9,123 Chinese jokes and 39,977 sentences in total, and with fine-grained annotation of humor, the dataset provides a new, sizeable, and scarce joke dataset, which is particularly valuable in languages other than English for scholars in many disciplines, such as computational, linguistic, and cognitive studies.

Acknowledgments

This research was supported by NSFC Programs (No. 61602079; No.61632011); Ministry of Education (No.16YJCZH141).

References

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. Datastories at semeval-2017 task 6: Siamese lstm with attention for humorous text comparison. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 390–395.

Kim Binsted, Anton Nijholt, Oliviero Stock, Carlo S-trapparava, G Ritchie, R Manurung, H Pain, Annalu Waller, and D O'Mara. 2006. Computational humor. *IEEE Intelligent Systems*, 21(2):59–69.

Esther Blumenfeld and Lynne Alpern. 1994. *Humor at work: The guaranteed, bottom-line, low-cost, high-efficiency guide to success through humor*. Peachtree Publishers.

Eric R Bressler and Sigal Balshine. 2006. The influence of humor on desirability. *Evolution and Human Behavior*, 27(1):29–39.

Santiago Castro, Luis Chiruzzo, Aiala Rosá, Diego Garat, and Guillermo Moncecchi. 2018. A crowd-annotated spanish corpus for humor analysis. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 7–11.

Santiago Castro, Matłas Cubero, Diego Garat, and Guillermo Moncecchi. 2017. Is this a joke? detecting humor in spanish tweets. *Inteligencia Artificial*.

Lambert Deckers and John Devine. 1981. Humor by violating an existing expectancy. *The Journal of Psychology*, 108(1):107–110.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Xiaoli Gan. 2015. A study of the humor aspect of english puns: Views from the relevance theory. *Theory and Practice in Language Studies*, 5(6):1211–1215.

Herbert M Lefcourt. 2001. *Humor: The psychology of living buoyantly*. Springer Science &amp; Business Media.

Lizhen Liu, Donghai Zhang, and Wei Song. 2018. Modeling sentiment association in discourse for humor recognition. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 586–591.

M G Lovorn. 2008. Humor in the home and in the classroom: The benefits of laughing while we learn. *Journal of Education and Human Development*, 2(1):1–12.

Rada Mihalcea and Carlo Strapparava. 2005. Making computers laugh: Investigations in automatic humor recognition. In *Conference on Human Language Technology & Empirical Methods in Natural Language Processing*, pages 531–538.

Mary Kay Morrison et al. 2012. *Using humor to maximize living: Connecting with humor*. R&L Education.

D. R. Morse. 2007. Use of humor to reduce stress and pain and enhance healing in the dental setting. *J N J Dent Assoc*, 78(4):32–36.

Reynier Ortega-Bueno, Carlos E Muniz-Cuza, José E Medina Pagola, and Paolo Rosso. 2018. Uo up-v: Deep linguistic humor detection in spanish social media. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018)*, pages 204–213.

John Allen Paulos. 2008. *Mathematics and humor: A study of the logic of humor*. University of Chicago Press.

Diana-Elena Popa. 2005. Jokes and translation. *Perspectives: Studies in Translatology*, 13(1):48–57.

Peter Potash, Alexey Romanov, and Anna Rumshisky. 2017. Semeval-2017 task 6:# hashtagwars: Learning a sense of humor. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 49–57.

Antonio Reyes, Paolo Rosso, and Tony Veale. 2013. A multidimensional approach for detecting irony in twitter. *Language Resources & Evaluation*, 47(1):239–268.

Howard F Stein. 1998. Organizational euphemism and the cultural mystification of evil. *Administrative Theory & Praxis*, pages 346–357.

Kendell C Thornton. 2007. Relationship closeness and embarrassment. *Individual Differences Research*, 5(1).

Villy Tsakona. 2009. Language and image interaction in cartoons: Towards a multimodal theory of humor. *Journal of Pragmatics*, 41(6):1171–1188.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2016. Exploring the realization of irony in twitter data. In *LREC*.

P Wooten. 1996. Humor: an antidote for stress. *Holistic Nursing Practice*, 10(2):49–56.

Diyi Yang, Alon Lavie, Chris Dyer, and Eduard Hovy. 2015. Humor recognition and humor anchor extraction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2367–2376.

Renxian Zhang and Naishi Liu. 2014. Recognizing humor on twitter. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 889–898. ACM.