Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1346–1352
Marseille, 11–16 May 2020

Development and Validation of a Corpus for Machine Humor Comprehension

Yuen-Hsien Tseng, Wun-Syuan Wu, Chia-Yueh Chang, Hsueh-Chih Chen, Wei-Lun Hsu
National Taiwan Normal University
No. 162, Sec. 1, Heping East Rd., Taipei 10610, Taiwan
MOST AI Biomedical Research Center
{samtseng, 40420213l, 80701003E, chcjyh, 60715004e}@ntnu.edu.tw

Abstract

This work developed a Chinese humor corpus containing 3,365 jokes collected from over 40 sources. Each joke was labeled with five levels of funniness, eight skill sets of humor, and six dimensions of intent by only one annotator. To validate the manual labels, we trained SVM (Support Vector Machine) and BERT (Bidirectional Encoder Representations from Transformers) with half of the corpus (labeled by one annotator) to predict the skill and intent labels of the other half (labeled by the other annotator). Based on two assumptions that a valid manually labeled corpus should follow, our results showed the validity for the skill and intent labels. As to the funniness label, the validation results showed that the correlation between the corpus label and user feedback rating is marginal, which implies that the funniness level is a harder annotation problem to be solved. The contribution of this work is two folds: 1) a Chinese humor corpus is developed with labels of humor skills, intents, and funniness, which allows machines to learn more intricate humor framing, effect, and amusing level to predict and respond in proper context (https://github.com/SamTseng/Chinese_Humor_MultiLabeled). 2) An approach to verify whether a minimum human labeled corpus is valid or not, which facilitates the validation of low-resource corpora.

Keywords: Humor Corpus, Humor Framing, Humor Intent, Corpus Validation, Multi-label Classification, Traditional Chinese

1. Introduction

Humor is an important element in inter-personal communication. In commercial applications, a humorous communication process can often dispel user complaints (Bellegarda, 2014; Binsted, 1995). In education, previous research (Bryant & Zillmann, 1989; Mcghee & Frank, 2014) has found that proper use of humor in the classroom can attract students' attention, improve in-class interaction, and help students learn better with fun. In more high-stake situations such as in delivering a speech in public, humor can help speakers reduce anxiety and improve their performance. In addition, advertising, entertainment, and other commercial sectors are also the domains of humor applications.

As human-dialogue systems, chatbots, or conversational user interface become versatile in this resurgent AI (Artificial Intelligence) age, the introduction of humor to the human-machine communication become important for the above reasons. This requires the system to recognize human humor or generate humorous utterance. To start building such a system, a humor corpus, commonly in the form of a joke collection, is needed, especially a manual labeled corpus for machines to learn to identify or even to generate humorous dialogues.

There have been a number of joke corpora collected in past studies. However, most corpora are just collections of joke stories, hilarious snippets, puns, or one-liners. Some with negative examples for humor recognition by machine learning. However, to have a deeper understanding of how a text snippet is amusing and what effect it may cause, joke framing and humor intent are important knowledge, in addition to the capability of distinguishing jokes/non-jokes.

In this paper, we introduce a joke corpus with manually labeled humor skills, intents, and funniness. Although it is a traditional Chinese corpus, it borrows some ideas from humor psychology to define the humorous characteristics not only applied to Chinese.

Currently, this corpus contains 3,365 jokes collected from over 40 sources. Each joke was labeled with five levels of funniness, eight skill sets of humor, and six dimensions of intent or motivation. To validate the manual labels, we trained SVM (Support Vector Machine) and BERT (Bidirectional Encoder Representations from Transformers) with half of the corpus (labeled by one annotator) to predict the skill and intent labels of the other half (labeled by the other annotator). The results showed that BERT performs better than SVM, although both results are still unsatisfactory due to the difficulty of the problem. In addition, both perform better than they are trained with randomly re-assigned labels. These two facts verify the validity of the costly labeling results. As to the funniness level, we validated the labels by developing a retrieval-based chatbot, called IceBreaker, to allow college students to tell a context-relevant joke in front of their audience during their term project oral presentation and to get their feedback. Our funniness level barely coincides with the rating given by 76 volunteer users, indicating the difficulty to reach the consensus of the funniness labels of the corpus, for which we have some observation and explanation.

The contribution of this work is two folds: 1) a traditional Chinese humor corpus is developed with labels of humor skills, intents, and funniness, which allow machines to learn more intricate humor framing, effect, and amusing level to predict and respond in proper context. 2) An approach to verify whether a minimum human labeled corpus is valid or not, which facilitates the validation of low-resource corpora.

The rest of the paper is organized as follows: In Section 2, we introduce existing humor datasets. Section 3 describes the development of the labeled corpus. Section 4 validate the corpus by comparing SVM and BERT as classifiers for skill and intent label prediction. Section 5 shows the validity of the funniness label in terms of regression and user feedback. Finally, we draw concluding remarks in Section 6. (The corpus and the validating tools are at: https://github.com/SamTseng/Chinese_Humor_MultiLabeled.)

2. Literature Review

In the field of natural language processing (NLP), human-computer interaction, and artificial intelligence, studies on humor identification and humor generation have been conducted for at least two decades. The goal of these studies is to explore humorous computational models (Bergen & Coulson, 2006; Ritchie, 2009), to enhance human-computer communication and user experience (Morkes, Kernal, & Nass, 1999; Nijholt, 2006), or to assist people with communication disabilities to enhance their interpersonal interactions (Ritchie, Manurung, Pain, Waller, & O'Mara, 2006).

Like other NLP tasks, the study of computational humor needs relevant corpora for machine (or even human) to understand what to learn.

Mihalcea and Strapparava (2006a, 2006b) used 10 English jokes (one-liners) as seed queries to search the web pages containing jokes (their URLs must contain oneliner, one-liner, humor, humour, joke, funny, etc.). From which, more jokes (for example, those listed in the <li> HTML tag like a seed joke) were extracted, and finally 16,000 jokes were obtained. To validate the joke collection, they randomly reviewed 200 of them, and only about 9% among the 200 were considered as noise (non-jokes). Examples of this collection include: "Change is inevitable, except from a vending machine". In addition, they also collected negative examples (non-jokes) with similar textual features (text length and used terms) as learning corpus for machine classification (identification) of jokes.

The above 16,000 one-liners corpus is static in terms of language usage. In contrast, those from social network platforms are dynamic (e.g., event-relevant). The ability to identify humorous sentences from these sources is an issue worth of studying. Zhang and Liu (2014) used the texts on Twitter. They collected 1,000 humorous tweets by automatic downloading and manual judgement. With the similar approach, they also collect 1,000 jokes (not in the tweets) from http://textfiles.com/. Examples include: "when nothing goes right... go left".

According to the transcripts of the TED speech, L. Chen and Lee (2017) semi-manually selected 4,726 humorous sentences (those trigger laughter), and randomly selected one sentence from the seven sentences before and after the funny sentence as a negative example for the joke classification task.

In contrast to the binary humor/joke classification, the International Workshop on Semantic Evaluation (SemEval) held the Learning a Sense of Humor evaluation task in 2017 (Potash, Romanov, & Rumshisky, 2017). From the tweets of a comedy competition TV program, a total of 112 topics and 12,734 tweets were collected and organized during a period of about eight months, and then the participants were asked to compare the humor levels. Hence, the humor is situational, and some even require external knowledge. For example: "The host of Singled Out #BadJobIn5Words" is more humorous than "Donut receipt maker and sorter #BadJobIn5Words".

In addition to English corpora, humor datasets have been developed in other languages. Castro, Chiruzzo, Rosa, Garat, and Moncecchi (2018) presented a corpus of 27,000 tweets written in Spanish and crowd-annotated by their humor value and funniness score, with about four annotations per tweet. The inter-annotator agreement Krippendorff's alpha value is 0.5710.

Blinov, Bolotova-Baranova, and Braslavski (2019) collected a large amount (about 150,000) of Russian jokes from various public online sources. They also collected 150,000 non-jokes for machine classification. To verify that the automatically created collection contains valid jokes and non-jokes, 1,000 random jokes and 1,000 random non-jokes were assessed through crowd sourcing. This resulted in 1,877 examples being labeled by at least three assessors and only 238 of them (12.7%) being observed as opposite assessments, i.e. 'not a joke' and 'a joke'.

Gu, Tseng, Hsu, Wu, and Chen (2019) has developed a collection of 3,691 jokes in traditional Chinese. The corpus is classified manually into 9 topical categories for retrieval in proper context.

3. Development of the Labeled Corpus

The study of computational humor often starts with jokes, which is one important form of creating humor in textual (or verbal) communication. To help building a machine to comprehend a joke so as to respond properly, it would be beneficial to know the various ways a joke is framed (uttering skills of a joke) and various effects when a joke is told (motivation or intent). These two aspects, as well as the joke quality (amusing level), are the main concerns when we develop the joke corpus.

In this study, the development of a traditional Chinese humor corpus follows the steps: 1) collecting, cleaning, and cataloguing a set of jokes; 2) adopting two classification schemes for joke framing and joke intent, respectively, and a multi-level funniness label; 3) manually labeling the jokes based on the schemes.

3.1 Collection of Jokes

To diversify the joke contents, we search and evaluate quite a few joke sources and, during a period of eight weeks, collect 3,828 jokes from 41 sources, which include 27 public websites (2777 jokes), 11 joke collection books (895 jokes), and 3 free Apps (156 jokes). As the jokes are accumulating, it is possible to collect duplicates from different sources. We then applied a full text matching technique, based on bag of words, TFxIDF term weighting and Cosine similarity, to detect near duplicates for removal.

3.2 Classification of Joke Framing, Intent, and Funniness

This subsection describes the classification schemes to help analyze how a joke is written, what effect it might be (e.g., what social functions it serves), and the amusing quality of a joke (to what extent a joke is considered humorous).

3.2.1 Joke Framing

There are various ways to "frame a joke" which refers to applying some humor skills when creating a joke. Although not exhaustively and not exclusively, the classification scheme that we adopted for joke framing is based on past studies of H. C. Chen, Chan, Dai, Liao, and Tu (2017), and most skill categories are relevant to incongruity in psychology. The scheme includes eight categories: 1) Double meanings; 2) Exaggeration; 3) Anthropomorphism; 4) Bridge-inference; 5) Illogic; 6) Irony; 7) Imitation; and 8) Others.

1) Double meaning means there are more than one way to interpret a text, in which one hide under the others commonly known. Once it was found, it creates a humorous feeling. Double meaning has the sub-types by homonyms, puns, semantics, grammar, and phrasing. Examples include: "An Apple (phone) a day keeps a Doctor (scholar degree) away." 2) Exaggeration refers to maximizing the level of situation or description in order to

impress people or express creativity. An example may read: “A fatty bachelor always replies someone's question about his marriage status: I'd rather wear a belly than a ring.” 3) Anthropomorphism refers to interpretation of non-human things or events in terms of human characteristics. Example: “0 considers herself as an elegant one. When she met 8, she criticized 8 as a phony fat one with a belt”. 4) Bridge inference: hides the relationship of a setup and a punch line so as to surprise the readers/listeners with the punch line in a playful way. Example: “Wife: you seldom drink outside. Why do you drink a lot at home? Husband: I was told that alcohol keeps me brave (the wife is too terrible to face without alcohol)”. 5) Illogic: Using a logical way in mistaken situations in order to mock a person’s silly or foolish behaviors, like: “My wife always encourage me to do my best. So, I do my best waiving every chore whatever she told me to do (including sex)”. 6) Irony: Describe a negative/positive situation opposite to the expectation, like: “I have nothing but money”. 7) Imitation: Make a similar snippet by following the logic of a setup snippet. Example: “A good spouse is a harbor for you to rest in a storm; a bad spouse is a storm in a harbor”. 8) Others: those that cannot be classified into the above seven categories, like some proverbs, quotations, or synaesthesia (e.g., “What surgery *technique* can turn eyes into ears? => Lip Reading *Technique*”). From the above examples, a joke may belong to multiple categories, such as “I have nothing but money” may also belong to Bridge inference.

3.2.2 Joke Intents

Humorous content may contain elements that challenge social norms or taboos, attack or taunt on people or things, or, on the contrary, comfort others. These contents could trigger an emotional journey (e.g., to release self-restriction or to increase personal sense of superiority), which in turn resolves the internal pressure and produces a pleasant feeling. Based on these perspectives proposed by H. C. Chen et al. (2017), the motivations or intents of jokes are classified into six categories: 1) Affinity; 2) Self-Improvement; 3) Attack; 4) Self-Depression; 5) Taboo; and 6) Others.

1) Affinity: turn the table or kick away the embarrassment with friendly and kindly words, make others feel comfort, or say something funny and relaxing to make everyone happy, in order to being closer to each other or ease the conflict in a group. Example: “Everyone wants peace in the world; I only want the world of you.” 2) Self-Improvement: Switch the viewpoint or self-encourage by accepting the ridiculous situation in order to cheer oneself and face the problem. It is a kind of humor coping strategies. Example: “If being good looking is a crime, then I am so guilty.” 3) Attack: Make oneself happy by laughing at the shortcomings of others, say something mean about others’ fault and difficulty, or make others uncomfortable to lower their status in a group. Example: “I have waited my dishes for an hour; is the chef sloth (rhymed with slow and implying laziness)?” 4) Self-Depression: Say or do something to mock oneself to please others, like: “I have no intent to commit a crime of being (naturally born) ugly.” 5) Taboo: Mocking something related to sex, death, excrement, forbidden behaviors, or thoughts. Example: “One day, a stack of black stools met a stack of white stools, and the black stool asked: Why do you look so white and so beautiful? The white one was very angry and replied: I'm not a shit! I'm ice cream!” 6) Others: those that cannot be classified into the above five categories. As above, this is a multi-label classification problem as a joke may belong to multiple intent categories.

3.2.3 Joke Funniness

In addition to the above classification schemes, the funniness level of a joke, ranging from 1 (least funny) to 5 (most funny), was adopted to reveal the amusing quality of the collected jokes.

3.3 Joke Labeling

Two annotators majored in Chinese linguistics were hired to label the data based on the above schemes and examples written in a manual. When they labeled the jokes, they also corrected typos, split long series of jokes into multiple ones, format the jokes for better reading and humorous effect, and removed similar ones that were not detected automatically (due to a stringent similarity threshold). This results in 3,365 jokes, where one labeled 1,691 and the other labeled 1674 of them. The labeling job took the part-time annotators about four months to finish.

4. Validation of the Joke Corpus

Because of the labor-intensive labeling job, only two annotators were recruited and each joke was labeled by only one. To verify the validity of the labeled corpus, we make the following assumptions for which any valid dataset should hold: 1) A machine learning model trained by manually labeled examples should outperform the same model trained by the corresponding randomly re-labeled examples; 2) Better classifiers shown in most corpora should still perform better for the corpus to be verified, in general cases. These two assumptions were applied to the multi-label tasks of joke framing and joke intent. For the funniness level, we validate its label in a real-case application described in Section 5.

4.1 Basic Statistics of the Corpus

The corpus is split into a training set and a test set. Those labeled by one annotator are regarded as training examples and those that were labeled by the other are regarded as the testing examples. As such, the training set has 1,691 jokes and the test set has 1,674 ones. The joke length distribution is shown in Figure 1 (by bins of joke length 10), where the Y axis is the number of training and testing jokes, respectively, and the X axis contains two series of numbers: the first line is the length of jokes (in Chinese characters or English words), while the second line is the sum of the numbers of the training and testing jokes at the corresponding joke length. Note, there are about 30 jokes whose length ranging from 500 to 2026.

Table 1, 2, and 3 show the number of jokes in each skill, intent, and funniness category for the training and test sets, respectively. It can be seen that for the skill scheme, most categories have about even number of jokes in the training and test sets. Only 3 categories show unbalanced distribution of jokes. For the intent scheme, most categories have more jokes in the training set. Only Others has the opposite distribution.

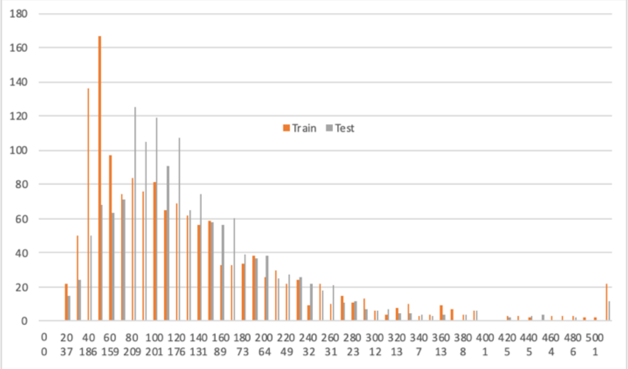

Figure 1: The length distributions of training and test sets.

Joke Skill	Train	Test	All
Double meanings	778	261	1039
Exaggeration	71	60	131
Anthropomorphism	140	40	180
Bridge inference	316	291	607
Illogic	438	486	924
Irony	47	35	82
Imitation	148	110	258
Others	110	478	588

Table 1: Number of jokes in each skill category.

Joke Intent	Train	Test	All
Affinity	73	19	92
Self Improvement	27	16	43
Attack	266	143	409
Self Depression	47	17	64
Taboo	355	229	584
Others	972	1254	2226

Table 2: Number of jokes in each intent category.

Funniness	Train	Test	All
1	47	316	363
2	251	616	867
3	742	571	1313
4	604	125	729
5	47	46	93

Table 3: Number of jokes in each funniness level.

Table 4 shows the number of jokes having multiple labels for the skill and intent classification scheme. As can be seen, most jokes are single-labeled. But for machine prediction, this is still a multi-label classification problem.

Another characteristic worth to note is that the highly skewed category distribution in the intent scheme. The Others category has far more jokes than the others. If measured with the inverse of Simpson diversity index (Simpson, 1949), denoted as 1/S, where $S = \sum_{i=1}^{n} s_i^2$ and s_i is the share (proportion of the number of jokes among all jokes) of category i, then the intent scheme has 1/S=2.13, meaning only 2.13 categories in average (among 6 categories) were used to label all the jokes. Whereas, this index for the skill scheme is 5.24 (among 8 categories) and is 3.6 (among 5 levels) for the funniness.

Joke Skill		Joke Intent	
No. of Label	No. of Jokes	No. of Label	No. of Jokes
1	2964	1	3301
2	355	2	57
3	41	3	1
4	3	0	6
0	2		

Table 4: Number of jokes having multiple labels.

4.2 Machine Learning for Category Prediction

For the multi-label classification tasks of joke framing and joke intent, we applied the scikit-learn Python library (Pedregosa et al., 2011) for performing traditional classification as baseline and a latest high-performing deep learning model for comparison.

There are quite a few approaches in the literature that transform the multi-label problem into multiple single-label problems such that the existing single-label algorithms can be used. The single-label algorithm used in our experiments is Support Vector Machine (SVM) (Cortes & Vapnik, 1995) with linear kernel function. The approaches that utilize SVMs are Binary Relevance (denoted as BR-SVM), Classifier Chains (CC-SVM), and Label Powerset (LP-SVM). In BR-SVM, an ensemble of SVM binary classifiers is trained, one for each class. Each SVM predicts either the membership or the non-membership of one category. The union of all categories that were predicted is taken as the multi-label output. This approach is easy to implement; however, it ignores the possible correlations between class labels. In CC-SVM, a chain of binary classifiers SVM_1, SVM_2, …, SVM_M is constructed, where a SVM_i uses the predictions of all the classifier SVM_j , where j < i. By this way, the method can take into account possible label correlations. The LP-SVM does take possible correlations between class labels into account, because it considers each member of the power set of labels in the training set as a single label. Thus, this method needs worst case 2^M classifiers, where M is the number of all categories. In consequence, some label combinations will have very few positive examples. Fortunately, most examples in our corpus are labeled with only 1 or 2 categories, as shown in Table 4. Therefore, the disadvantage is not severe and is worth of a try.

BERT (Bidirectional Encoder Representations from Transformers) (Devlin, Chang, Lee, & Toutanova, 2018) is a deep learning model that is able to learn to classify texts with state-of-the-art performance. Although in recent months there are other models that may perform even better, like XLNet (Yang et al., 2019) and ERNIE (Sun et al., 2019), the widespread popularity of BERT in supporting Chinese enable us to apply tools, like simpletransformers (Rajapakse, 2019), to accomplish our experiments with ease.

For the performance metrics, we show the scores of Micro-F, Macro-F, Macro-ROCAUC (Area Under the Receiver

Operating Characteristic Curve), and Micro-ROCAUC implemented in the scikit-learn library. Note that when the category distribution is skewed (unbalanced), micro scores tend to reflect the performance of a few large categories (those with a large number of examples), while macro scores tend to reflect the performance of a large number of small categories (those with few examples).

Table 5 shows the four metrics scores for the four classifiers trained with manually labeled examples, while Table 6 shows the performance trained with the corresponding examples with randomly re-assigned labels. By randomly re-assigned labels, we mean that each training example has the same number of categories assigned, but randomly shuffled; in the meantime, the shuffled categories should be constrained to maintain the same category size (i.e., the number of examples belong to the original category). To achieve this constraint, we apply a program code at stackoverflow.com posted to answer the question of randomizing a matrix while keeping row and column totals the same. In other words, given a matrix A with N rows (corresponding to N training examples) and M columns (corresponding to M categories) with $A_{ij} = 1$ indicating example *i* belongs to category *j*, and $A_{ij} = 0$ otherwise, the shuffled matrix B has the same sum as A for each row and for each column. Depending on A, B might have some rows that remain the same as in A. We choose those B that has as minimum number of same rows as A as possible. As a result, the shuffled training set for the skill scheme has 362 training examples (21.41%) with their labels unchanged, and for intent scheme the number of unchanged rows is 491 (29.04%).

Compared to Table 5, Table 6 shows that for each classifier under each metric, its performance drops when learned from the randomly re-assigned training labels, which meets the assumption 1. In Table 5, BERT performs better under each metric, which meets assumption 2. Therefore, for the skill scheme, the manual labels are valid for this corpus.

Skill	Micro-F	Macro-F	Macro-ROCAUC	Micro-ROCAUC
BR-SVM	0.2231	0.1414	0.5314	0.5559
CC-SVM	0.2569	0.1606	0.5327	0.5722
LP-SVM	0.2281	0.1795	0.5541	0.5522
BERT	**0.2829**	**0.2075**	**0.5640**	**0.5867**

Table 5: Performance with manually labeled training set.

Skill	Micro-F	Macro-F	Macro-ROCAUC	Micro-ROCAUC
BR-SVC	0.1538	0.0871	0.5000	0.5198
CC-SVC	0.1787	0.1131	0.5009	0.5273
LP-SVC	0.1915	0.1454	0.5110	0.5261
BERT	0.1805	0.0917	0.4980	0.5269

Table 6: Results of randomly re-assigned training labels.

Similarly, compare to Table 7, Table 8 shows that for each classifier under the macro metrics, its performance drops when learned from the randomly re-assigned training labels, which meets the assumption 1. In Table 7, BERT performs better under the macro metrics, which meets assumption 2. The reason that the assumption 1 and 2 do not hold for the micro metrics is due to the highly skewed distribution of the intent scheme: the Others category has more than 57% training examples and 74% testing examples. From this result, the manual labels are valid for the Intent scheme, but may be improved if Others is further divided into subcategories or re-assigned.

Intent	Micro-F	Macro-F	Macro-ROCAUC	Micro-ROCAUC
BR-SVC	0.4993	0.1994	0.5369	0.6904
CC-SVC	**0.5704**	0.2115	0.5312	**0.7427**
LP-SVC	0.5256	0.2232	0.5440	0.7159
BERT	0.5487	**0.2880**	**0.6102**	0.7235

Table 7: Performance with manually labeled training set.

Intent	Micro-F	Macro-F	Macro-ROCAUC	Micro-ROCAUC
BR-SVC	0.5198	0.1277	0.4891	0.6998
CC-SVC	0.5602	0.1448	0.4855	0.7359
LP-SVC	0.5277	0.1406	0.4813	0.7170
BERT	0.5180	0.1342	0.4900	0.7013

Table 8: Results of randomly re-assigned training labels.

5. Application of the Corpus for Validation

We have applied three regression techniques, namely Linear Regression, Linear Support Vector Regression, and Support Vector Regression, to learn to predict the funniness level of a joke in our corpus using the metrics: Mean-Square-Error (MSE), Mean-Absolute-Error (MAE), and number of correct predictions after rounding the predicted value. The result is shown in Table 9. Although SVR perform the best in terms of these metrics, it nearly yields a constant prediction value around 3.1, because most jokes' funniness are around 3, 4, and 2, as observed from Table 3.

Funniness	MSE	MAE	Correct Predictions
Linear	1.8555	1.1269	447
Linear SVR	1.6650	1.0614	529
SVR	1.4409	0.9547	571

Table 9: Results of regression methods for funniness.

Due to the above bias, we decide to use human evaluation to verify the funniness level labeled by our annotators. However, it should be noted that humor has at least five characteristics, including subjectivity, regionality, culture, current affairs, and language differences. Each person may respond differently to the same joke due to his/her mood, understanding, or familiarity of the joke they have seen. Therefore, we constrained the human evaluation in a specific scenario for a certain group of people.

As a result, we designed a retrieval-based chatbot, called IceBreaker, for use by college students who would make an oral presentation in their final project. This free chatbot allow users to find relevant jokes to utter at the beginning of their public presentation in order to relax an unduly formal atmosphere, which is basically the propaganda we propagate in various social media channels to solicit college students to use. The chatbot has quick feedback buttons for users to rate the amusing level (from 1 to 3 as our preliminary test showed no good response to use a 5-level funniness in this application scenario) and to report whether it helps to achieve icebreaking effect. (Note: high amusing level did not certainly lead to icebreaking effect, and vice versa in our application experiment, although these two variables are positively highly correlated.)

During a period of two weeks in the end of a semester, 76 users had made 362 valid joke queries (including feedback), and 96 (26.52%) of which achieved the icebreaking effect. The contingency table between the funniness level in the corpus and the feedback amusing level is shown in Table 10. Our Chi-Square Test shows that the independence of these two ratings cannot be rejected at the significance level 0.05 (18.51 > 15.51 = CHIINV(0.05, 8)), but can be rejected at the significance level 0.01 (18.51 < 20/09 = CHIINV(0.01, 8) using the Excel formula). This conflict result shows that the correlation between the corpus label and user feedback is marginal and implies that the funniness level is a harder annotation problem to be solved.

	1	2	3	Total
1	27	3	6	36
2	34	6	11	51
3	90	41	25	156
4	47	30	20	97
5	8	9	5	22
Total	206	89	67	362

Table 10. Corpus funniness (row) vs feedback amusing levels (column).

6. Concluding Remarks

Previous studies in developing humor corpora often sample only a little portion of jokes for binary human judgment (amusing or not). Few has touched the problem of multi-level funniness and aspects such as humor skill and human intent. In this work, we develop a corpus of 3,365 jokes labeled by two annotators based on an 8-category joke skill scheme, a 6-category joke intent scheme, and a 5-level funniness rating. Despite the intricate distinction between the categories, using only one annotator for each joke is possible to yield a valid corpus for the humor skill and intent schemes. However, for the multi-level funniness rating, this is a harder problem for human annotation.

This work has shed light on some kind of low-resource corpora that are costly to label. We have demonstrated how to verify the validity of a multi-labeled corpus with each text labeled by only one annotator.

This work also points some directions for future computational humor research. For example, our results applying SVM and BERT for humor skill and intent prediction are still unsatisfactory and can only be considered as baselines. More linguistic features for humor comprehension should be explored and our corpus could support such possibility. As another example, binary humorous judgement may be viable; however, multi-level humorous rating, as the application of which has been demonstrated in commercial movies like Interstellar, is a delicate problem that needs to be tackled in the future.

7. Acknowledgements

This work was supported in part by the grant MOST 107-2221-E-003-014-MY2, MOST-109-2634-F-002-023 and MOST AI Biomedical Research Center from Ministry of Science and Technology of Taiwan.

8. Bibliographical References

Bellegarda, J. R. (2014). Spoken Language Understanding for Natural Interaction: The Siri Experience. In J. Mariani, S. Rosset, M. Garnier-Rizet, & L. Devillers (Eds.), Natural Interaction with Robots, Knowbots and Smartphones. New York, USA: Springer.

Bergen, B., & Coulson, S. (2006). Frame-Shifting Humor in Simulation-Based Language Understanding. IEEE Intelligent Systems, 21(1), 59-62.

Binsted, K. (1995). Using humour to make natural language interfaces more friendly. Paper presented at the AI, ALife and Entertainment Workshop, Montreal, Canada.

Blinov, V., Bolotova-Baranova, V., & Braslavski, P. (2019, jul). Large Dataset and Language Model Fun-Tuning for Humor Recognition, Florence, Italy.

Bryant, J., & Zillmann, D. (1989). Chapter 2: Using Humor to Promote Learning in the Classroom. Journal of Children in Contemporary Society, 20(1-2), 49-78. doi:10.1300/J274v20n01_05

Castro, S., Chiruzzo, L., Rosa, A., Garat, D., & Moncecchi, G. (2018). A Crowd-Annotated Spanish Corpus for Humor Analysis. Paper presented at the The Sixth International Workshop on Natural Language Processing for Social Media, Melbourne, Australia.

Chen, H. C., Chan, Y. C., Dai, R. H., Liao, Y. J., & Tu, C. H. (2017). Neurolinguistics of Humor. In S. Attardo (Ed.), The Routledge Handbook of Language and Humor (pp. 282-294). London: Routledge.

Chen, L., & Lee, C. M. (2017). Predicting Audience's Laughter Using Convolutional Neural Network. arXiv:1702.02584. Retrieved from http://arxiv.org/abs/1702.02584

Cortes, C., & Vapnik, V. N. (1995). Support-Vector Networks. Machine Learning, 20(3), 273-297.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv. Retrieved from https://arxiv.org/pdf/1810.04805.pdf

Gu, Y.-C., Tseng, Y.-H., Hsu, W.-L., Wu, W.-S., & Chen, H.-C. (2019). Development and Classification of a Chinese Humor Corpus. Paper presented at the The 20th International Conference on Computational Linguistics and Intelligent Text Processing, La Rochelle, France.

Mcghee, P. E., & Frank, M. (2014). Humor and Children's Development: A Guide to Practical Applications. Oxford, UK: Routledge.

Mihalcea, R., & Strapparava, C. (2006a). Learning to Laugh (Automatically): Computational Models for Humor Recognition. Computational Intelligence, 22(2), 126-142.

Mihalcea, R., & Strapparava, C. (2006b). Technologies That Make You Smile: Adding Humor to Text-Based Applications. IEEE Intelligent Systems, 21(5), 33-39. doi:10.1109/MIS.2006.104

Morkes, J., Kernal, H., & Nass, C. (1999). Effects of humor in task-oriented human- computer interaction and computer-mediated communication: a direct test of

SRCT theory. Human-Computer Interaction, 14(4), 395-435.

Nijholt, A. (2006). Embodied Conversational Agents: “A Little Humor Tool”. IEEE Intelligent Systems, 21(2), 62-64.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

Potash, P., Romanov, A., & Rumshisky, A. (2017). SemEval-2017 Task 6: #HashtagWars: Learning a Sense of Humor. Paper presented at the 11th International Workshop on Semantic Evaluations, Vancouver, Canada.

Rajapakse, T. (2019). simpletransformers. Retrieved from https://github.com/ThilinaRajapakse/simpletransformers

Ritchie, G. (2009). Can Computers Create Humor? AI Magazine, 30(3), 71-81.

Ritchie, G., Manurung, R., Pain, H., Waller, A., & O’Mara, D. (2006). The STANDUP Interactive Riddle-Builder. IEEE Intelligent Systems, 21(2), 67-69.

Simpson, E. H. (1949). Measurement of Diversity. Nature, 163, 688.

Sun, Y., Wang, S., Li, Y., Feng, S., Tian, H., Wu, H., & Wang, H. (2019). ERNIE 2.0: A Continual Pre-training Framework for Language Understanding. Retrieved from http://arxiv.org/abs/1907.12412

Yang, Z., Dai, Z., Yang, Y., Carbonell, J. G., Salakhutdinov, R., & Le, Q. V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. Retrieved from http://arxiv.org/abs/1906.08237

Zhang, R., & Liu, N. (2014). Recognizing Humor on Twitter. Paper presented at the 23rd ACM International Conference on Information and Knowledge Management, Shanghai, China.