Supporting Comedy Writers: Predicting Audience’s Response from Sketch Comedy and Crosstalk Scripts

Maolin Li
Independent Research*
maolin.li.cs@gmail.com

Abstract
Sketch comedy and crosstalk are two popular types of comedy. They can relieve people’s stress and thus benefit their mental health, especially when performances and scripts are high-quality. However, writing a script is time-consuming and its quality is difficult to achieve. In order to minimise the time and effort needed for producing an excellent script, we explore ways of predicting the audience’s response from the comedy scripts. For this task, we present a corpus of annotated scripts from popular television entertainment programmes in recent years. Annotations include a) text classification labels, indicating which actor’s lines made the studio audience laugh; b) information extraction labels, i.e. the text spans that made the audience laughed immediately after the performers said them. The corpus will also be useful for dialogue systems and discourse analysis, since our annotations are based on entire scripts. In addition, we evaluate different baseline algorithms. Experimental results demonstrate that BERT models can achieve the best predictions among all the baseline methods. Furthermore, we conduct an error analysis and investigate predictions across scripts with different styles.  

1 Introduction
Comedy plays a major role in people’s lives in that it relieves stress and anxiety (Williams et al., 2005; Sarıtaş et al., 2019). There are two popular types of comedy: sketch comedy and crosstalk. A sketch comedy usually presents a short story and is performed by multiple comedians in various short scenes; while in a crosstalk performance, which is similar to a talk show, there are usually two performers telling humorous stories behind a desk. Although these two types of comedy are different, both of them are performed based on scripts. A script breaks down a story into pieces along with the details that describe which performer should take what action or say which lines at a specific point (Blake, 2014). Therefore, the quality of the script is critical and it directly influences whether the audience enjoys the performance.

However, it is difficult for script writers to ensure a high-quality comedy script and be productive. Firstly, writers have to assess if audiences will react as expected, in particular laughing at specific points. It is necessary to rehearse multiple times to continuously improve the script, which is time-consuming and can be costly. Secondly, to develop laughter triggers, writers need to identify the potential points from the script where there are possibilities for performers to use funny body moves, tone or tell amusing stories to make the audience laugh. Thirdly, the more times a script is publicly performed, the less laughter it can bring, since the audience have become too familiar with it. Thus, it is essential for comedy writers to explore new laughter triggers constantly.

Since natural language processing (NLP) has been widely and successfully applied to a number of fields (Carrera-Ruvalcaba et al., 2019; Rao and McMahan, 2019), we investigate how NLP methods can support comedy writers to produce high-quality scripts more efficiently. This paper specifies this challenge as a new task, i.e. the prediction of the audience’s response to sketch comedy and crosstalk scripts. To address this challenge, we explore the use of two different NLP methodologies: 1) Text Classification: we predict whether or
not an actor’s lines\(^2\) can make audiences laugh. In other words, we formulate the task of predicting as a binary text classification problem. 2) Information Extraction: we predict the text spans from an actor’s lines indicating the specific words that trigger an audience’s laughter.

**Contributions** Firstly, we introduce a Chinese corpus of annotated comedy scripts collected from popular TV entertainment programmes. Our annotations include both text classification and information extraction labels. Tables 1 and 2 present annotation examples. The corpus can be used to build an intelligent system to benefit the script writing for comedy writers. It may also be useful for dialogue system research and discourse analysis. Secondly, we evaluate a number of NLP methods and the results demonstrate that BERT models (Devlin et al., 2019) are able to achieve the best prediction performance among all methods. We also further conduct an error analysis which may be useful for further improving the performance. Lastly, we experimentally show that our corpus can also be used to predict laughter triggers for scripts which have very different styles compared to training data.

## 2 Related Work

Our work is closely related to humour detection, which has been widely studied for many years in natural language processing. Mihalcea and Strapparava (2006); Yang et al. (2015); Chen and Soo (2018); Blinov et al. (2019) investigated if a text fragment is a one-liner.\(^3\) Zhang and Liu (2014); Ortega-Bueno et al. (2018); Chiruzzo et al. (2019) explored the humour classification task on tweets. Castro et al. (2018) collected humour values and funniness scores of Spanish tweets by using crowdsourcing. Chiruzzo et al. (2019) proposed a regression task that predicts the humour score for a tweet. Li et al. (2020) collected Chinese Internet slang expressions and combined them with a humor detecting method to analyse the sentiment of Weibo\(^4\) posts. It should be noted that the examples in all of the corpora used or constructed in the above-mentioned studies are independent of each other. Since our corpus is based on entire scripts, the annotated lines and text spans might also benefit the researchers who are interested in modelling long-context-aware algorithms to understand humour. Apart from the studies on short text fragments, Bertero (2019) and Hasan et al. (2019) created corpora from television (TV) sitcoms such as *The Big Bang Theory*\(^5\) and *TED talks*\(^6\) respectively. Their goal is to predict whether or not a sequence of texts will trigger immediate laughter. Yang et al. (2015); Zhang et al. (2019) extracted the key words such as *sung*, *sign language* and *pretty handy* from jokes, which are similar to our information extraction annotations.

\(^2\)The lines are from the dialogue of a comedy performance. Each line consists of an actor’s name and the sentences this actor speaks in performance.

\(^3\)A one-liner is a joke that is delivered in a single line which only contains a few words.

\(^4\)Weibo is a Chinese micro-blogging website similar to Twitter: \(https://www.weibo.com/\)

\(^5\)https://the-big-bang-theory.com/

\(^6\)https://www.ted.com/talks
3 Corpus

3.1 Data Collection

Source Selection In order to ensure the high-quality of scripts, we carefully selected thirty performances (the total duration is approximately 473 minutes), including both sketch comedies and crosstalks, of which the leading roles are famous Chinese comedians. These performances were played on well-known Chinese TV entertainment programmes such as Chinese New Year Gala and Ace VS Ace\(^7\). Since there were many people in the audience present for the recording of these performances, the annotators can judge whether the audience laughed based on the performance videos. Please refer to the appendix for the full list of performances which gives details of their titles, leading comedians and sources. Lastly, we manually typed up actors’ lines for each performance and completed thirty scripts. Although there may be differences between our scripts and the real scripts used by comedians in terms of format or content, we assume that our scripts contain the key information about the real scripts, i.e., the actors’ lines. Therefore the corpus can be useful for the development of intelligence-assistant comedy script writing systems.

Diversity We also took the comedy style into consideration. In order to ensure the diversity and its balance: a) The performances were selected from three main different types of sources\(^8\) as shown in Table 3, including the topic descriptions of selected performances. It can be observed that the corpus has a wide range of topics. b) As a preliminary study, we selected six popular Chinese comedians who have various and distinctive styles, and we chose five representative performances of each comedian.

Corpus Statistic Table 4 illustrates the statistics and Figure 1 shows the laughter rates of each script. The highest line-level and character-level rates are

\(^7\)https://es.wikipedia.org/wiki/Ace_vs_Ace

\(^8\)The three sources are: Chinese New Year Galas—the annual televised Chinese New Year celebrations which are the most viewed TV shows in China. The shows consist of various performances including sketch comedies and crosstalks; Reality Shows—the programmes that show the unscripted actions of participants such as playing games and talking. We selected the shows in which comedians were involved; Comedy Competition Shows—the programmes where different comedians present their comedy performances to a studio audience and the winners are selected based on the audience’s votes.

| Source                     | Topics                                                                 |
|----------------------------|------------------------------------------------------------------------|
| Chinese New Year Galas     | Love stories and blind dates between old people;                        |
|                            | Reflecting social phenomena to call for a better society (e.g. avoid judging people by their appearances, do not spoil children, care more about lonely seniors, the woman builds a good relationship with her mother-in-law, spend more time with children, be wary of scams); |
|                            | Funny family stories during spring festival;                            |
| Reality Shows              | Stories happened in ancient times;                                     |
|                            | Stories about young people (e.g. encounter ex-boyfriends or ex-girlfriends, relationships between best friends, blind dates); |
|                            | Reflecting social phenomena to call for a better society (e.g. give seats to vulnerable people); |
| Comedy Competition Shows   | Love stories;                                                           |
|                            | Hot topics (e.g. support the COVID-19 frontline fighters);               |
|                            | Funny stories that happened among friends and in families;              |
|                            | Reflecting social phenomena to call for a better society (e.g. be wary of scams, care more about orphans in orphanages); |

Table 3: Topics of the selected comedies.

| Statistics                | Value         |
|---------------------------|---------------|
| # of Comedy Scripts       | 30            |
| Year Range                | 2014—2020     |
| Total Duration            | 473.44 mins   |
| Average Duration          | 15.78 mins    |
| # of Actors’ Lines        | 6087          |
| Laughter Rate (Line-Level)| 28.62%        |
| # of Characters           | 120451        |
| Laughter Rate (Character-Level)| 8.16%    |

Table 4: Corpus statistics. \# of Actors’ Lines and Characters correspond to the total number of lines and characters in our corpus respectively. Laughter Rate is the rate of lines/characters that trigger laughter.

![Figure 1: Each script’s laughter rate in our corpus.](image)

45.39% and 13.12%, while the lowest rates are 16.03% and 3.49%. We note that the character-level laughter rates vary in different scripts. This may be due to density of laughter triggers of a line or the topic of the script.

3.2 Annotation

The annotation was completed on Doccano platform (Nakayama et al., 2018) and the annotators are two native Chinese speakers. The annotations were produced based on the studio audiences’ responses as observed in the videos, and are not based on the annotators’ responses.

Annotation Instruction Annotating text classification labels is easy; annotators are requested to simply assigned label 1 to the lines that make audiences laugh, and 0 to the others. With regard to
the information extraction annotations, annotators are requested to identify text spans which are usually phrases. The span consists of the words that immediately made the audience laugh after the comedians said them. For example, as indicated in Table 2, the span incredibly big was annotated. In this case, only annotating big would be considered as an incorrect annotation, because the comedian was using incredibly to strongly emphasise big which was her first impression of a man’s house in a blind date. Only annotating incredibly would also be incorrect, because the main reason why the audience laughed was because the comedian said the house looked big. 9

Annotation Process The annotation process was as follows: Firstly, the annotators conducted discussions about the conflicting annotations after several attempts to annotate the same three scripts. Secondly, once agreements about how to solve the conflicts had been reached, they started to annotate their assigned scripts. Afterwards, since information extraction annotation is more complex than that of classification annotations, we measured its quality by computing three types of inter-annotator agreement. We asked the annotators to annotate the same six scripts having different styles and then calculated the Overall Percent Agreement (OPA), Fleiss’s kappa (Fleiss, 1971) and Randolph’s kappa (Randolph, 2005). We found that the agreement rates were high (OPA 98.09%, Fleiss’s Kappa 0.85, Randolph’s Kappa 0.96). This is due to the fact that the discussions about solving conflicts were in-depth and the laughter triggers were usually clear in the lines.

4 Baselines and Results Discussion

In order to understand how well the machine learning methods work on our corpus, we evaluate the performances of a number of models on 5-fold cross-validation random splits of the scripts in our corpus and report the average results in this section. 10 All the BERT models were pre-trained by using a mixture of large Chinese corpora. 11 Please refer to the appendix for the results of each fold, statistics of splits, computing infrastructure, each model’s running time, parameter details and hyper-parameter settings.

Baselines Tables 5 and 6 respectively present the results of text classification and information extraction. BERT-base has the best F1-scores among all the methods. We also note that the classification recall of RCNN (Lai et al., 2015) is much higher than other methods. Therefore, we suggest using this model if users prefer a classifier with a high recall. In addition, we observe that the scores are not high, especially for the information extraction task. The reason may be if the audience laughter highly depends on the conversation contexts which were not considered by baselines. Therefore, taking a longer conversation context of a line into consideration is a worthy research direction.

Prediction Errors Tables 7 and 8 present examples incorrectly predicted by the BERT-base. The first 3 examples describe how the model failed to predict the laughter triggers, while the last 3 examples show false positive predictions. Incorporating the context information of lines may further reduce these errors.

Cross-Style Performance We further investigate the performance of predicting laughter triggers on scripts with a totally different style compared to

| Model                        | P   | R   | F   | Acc  |
|------------------------------|-----|-----|-----|------|
| CNN (Kim, 2014)              | 42.51 | 64.14 | 51.07 | 66.29 |
| RCNN (Lai et al., 2015)      | 41.21 | 68.52 | 50.89 | 63.54 |
| BiLSTM (Liu et al., 2016)    | 41.17 | 65.94 | 52.18 | 66.69 |
| + Attention (Zhou et al., 2016) | 39.97 | 59.39 | 47.44 | 63.94 |
| FastText (Joshi et al., 2017) | 40.64 | 68.26 | 56.12 | 63.72 |
| DPCNN (Johson and Zhang, 2017) | 42.46 | 61.32 | 50.76 | 66.32 |
| Transformer (Yang et al., 2017) | 42.60 | 64.71 | 51.24 | 66.18 |
| BERT-tiny (Jiao et al., 2019) | 47.56 | 53.38 | 49.91 | 66.27 |
| BERT-small (Ture et al., 2019) | 47.29 | 56.21 | 51.21 | 70.78 |
| BERT-base (Declin et al., 2019) | 47.60 | 56.64 | 52.61 | 70.94 |

Table 5: Text classification performance. P, R, F and Acc. are Precision, Recall, F1-score and Accuracy respectively.

| Model                        | P   | R   | F   | Acc  |
|------------------------------|-----|-----|-----|------|
| BOW (Rabiner and Huang, 1986) | 22.19 | 7.43 | 11.04 |     |
| CRF (Lafferty et al., 2001)  | 28.56 | 6.11 | 10.01 |     |
| BiLSTM (Huang et al., 2015)  | 31.21 | 1.64 | 3.09  |     |
| BiLSTM-CRF (Lample et al., 2016) | 30.33 | 9.81 | 14.48 |     |
| HMM (Rabiner and Juang, 1986) | 21.47 | 24.59 |     |     |
| CRF (Lafferty et al., 2001)  | 25.86 | 6.11 | 10.01 |     |
| BiLSTM (Huang et al., 2015)  | 31.21 | 1.64 | 3.09  |     |
| BiLSTM-CRF (Lample et al., 2016) | 30.33 | 9.81 | 14.48 |     |
| BERT-tiny (Jiao et al., 2019) | 26.26 | 19.89 | 22.57 |     |
| BERT-small (Ture et al., 2019) | 28.82 | 17.52 | 21.56 |     |
| BERT-base (Declin et al., 2019) | 30.15 | 21.47 | 24.59 |     |

Table 6: Information extraction performance. Relaxed metrics are used. The exact-match metric is over-strict because the length of text spans in this corpus is much longer than general named entities. The computation of these metrics can be found in (Nguyen et al., 2017).
We study the prediction of laughter triggers from comedies and crosstalk scripts. Secondly, we evaluate a number of baselines and find that BERT models achieve the best performance. We note that the information extraction performance was very low, indicating that this task is particularly challenging. We also conduct an error analysis of incorrect predictions. The errors suggest the incorporation of rich context information may further improve the performance. Therefore, it is worth investigating a model which can take such information into consideration. Furthermore, it is also worth extending the corpus to a multimodal one by aligning scripts to corresponding audios or videos, because certain intonations or scenes can also make audiences laugh. The multimodal corpus can also benefit the creation of silent comedy. Enriching the corpus by including scripts in other languages may also be a potential direction. Lastly, the encouraging cross-style prediction performance shows the usefulness of our corpus for predicting new scripts with different styles.

Moreover, it is also interesting to explore human performances by asking annotators to make predictions based purely on the scripts of unwatched comedies, and to investigate if the script writers find the model predictions insightful. We hope this study will benefit script writing by inspiring the community to develop intelligent systems for comedy writers and other artists in the field. The corpus might also be useful for researchers who are working on related or similar tasks, such as discourse analysis and humorous response generation for dialogue systems.

Acknowledgements

We would like to express our sincere appreciation to all comedians and backstage teams for their hard work to make audiences happy. We also sincerely appreciate the valuable comments of all anonymous reviewers. My deepest thanks should also be given to my best friend who inspired me to conduct such research and whose encouragement was important for me to be able to complete the work.
References

Dario Bertero. 2019. Conversational humor recognition and generation through deep learning. PhD dissertation, Hong Kong University of Science and Technology.

Marc Blake. 2014. How to be a Comedy Writer, volume 1. Andrews UK Limited.

Vladislav Blinov, Valeria Bolotova-Baranova, and Pavel Braslavski. 2019. Large dataset and language model fun-tuning for humor recognition. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4027–4032, Florence, Italy. Association for Computational Linguistics.

Ernesto Carrera-Ruvalcaba, Johnson Ekedum, Austin Hancock, and Ben Brock. 2019. Leveraging natural language processing applications and microblogging platform for increased transparency in crisis areas. SMU Data Science Review, 2(1):6.

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, Melbourne, Australia. Association for Computational Linguistics.

Peng-Yu Chen and Von-Wun Soo. 2018. Humor recognition using deep learning. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 113–117, New Orleans, Louisiana. Association for Computational Linguistics.

Luis Chiruzzo, S Castro, Mathias Etcheverry, Diego Garat, Juan José Prada, and Aiala Rosá. 2019. Overview of haha at iberlef 2019: Humor analysis based on human annotation. In Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019). CEUR Workshop Proceedings, CEUR-WS, Bilbao, Spain.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for 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, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Md Kamrul Hasan, Wasiur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. 2019. UR-FUNNY: A multimodal language dataset for understanding humor. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2046–2056, Hong Kong, China. Association for Computational Linguistics.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. arXiv preprint arXiv:1508.01991.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2019. Tinybert: Distilling bert for natural language understanding. arXiv preprint arXiv:1909.10351.

Rie Johnson and Tong Zhang. 2017. Deep pyramid convolutional neural networks for text categorization. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 562–570, Vancouver, Canada. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the 18th International Conference on Machine Learning, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, AAAI’15, page 2267–2273. AAAI Press.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 260–270, San Diego, California. Association for Computational Linguistics.

Da Li, Rafał Rzepka, Michał Ptaszynski, and Kenji Araki. 2020. Hemos: A novel deep learning-based fine-grained humor detecting method for sentiment analysis of social media. Information Processing & Management, 57(6):102290.
Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Recurrent neural network for text classification with multi-task learning. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI ’16, page 2873–2879. AAAI Press.

Rada Mihalcea and Carlo Strapparava. 2006. Learning to laugh (automatically): Computational models for humor recognition. *Computational Intelligence, 22*(2):126–142.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for humans. Software available from https://github.com/doccano/doccano.

An Thanh Nguyen, Byron Wallace, Junyi Jessy Li, Ani Nenkova, and Matthew Lease. 2017. Aggregating and predicting sequence labels from crowd annotations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 299–309, Vancouver, Canada. Association for Computational Linguistics.

Reynier Ortega-Bueno, Carlos E Muniz-Cuza, José E Medina Pagola, and Paolo Rosso. 2018. Uo upy: 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 203–213, Seville, Spain.

Lawrence Rabiner and B Juang. 1986. *An introduction to hidden markov models*. *IEEE ASSP Magazine, 3*(1):4–16.

JJ Randolph. 2005. *Free-marginal multirater kappa: an alternative to fleiss’ fixed-marginal multirater kappa*. Joensuu University Learning and Instruction Symposium.

Delip Rao and Brian McMahan. 2019. Natural language processing with PyTorch: build intelligent language applications using deep learning. “O’Reilly Media, Inc.”.

Serdar Sartaş, Hasan Genç, Şerafettin Okutan, Ramazan Inci, Ahmet Özdemir, and Gülnaz Kızılkaya. 2019. The effect of comedy films on postoperative pain and anxiety in surgical oncology patients. *Complementary Medicine Research, 26*(4):231–239.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: The impact of student initialization on knowledge distillation. *arXiv preprint arXiv:1908.08962*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Sandy M Williams, Pauline K Arnold, and Jennifer N Mills. 2005. Coping with stress: a survey of Murdoch university veterinary students. *Journal of Veterinary Medical Education, 32*(2):201–212.

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, Lisbon, Portugal. Association for Computational Linguistics.

Dongyu Zhang, Heting Zhang, Xikai Liu, Hongfei Lin, and Feng Xia. 2019. Telling the whole story: A manually annotated Chinese dataset for the analysis of humor in jokes. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6402–6407, Hong Kong, China. Association for Computational Linguistics.

Rexnian Zhang and Naishi Liu. 2014. Recognizing humor on twitter. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM ’14*, page 889–898, New York, NY, USA. Association for Computing Machinery.

Zhe Zhao, Hui Chen, Jinbin Zhang, Xin Zhao, Tao Liu, Wei Lu, Xi Chen, Haotang Deng, Qi Ju, and Xiaoyong Du. 2019. UER: An open-source toolkit for pre-training models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 241–246, Hong Kong, China. Association for Computational Linguistics.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 207–212, Berlin, Germany. Association for Computational Linguistics.

### A Appendices

#### A.1 Computing Resources

Table 11 describes the details of the computing resources used for all of our experiments. These resources are freely available from Paperspace.\(^\text{12}\)

#### A.2 Model Details

Below we present model hyper-parameter values,\(^\text{13}\) and the average running time of one epoch.

\(^{12}\)https://www.paperspace.com/

\(^{13}\)The size of model’s trainable parameters can be found in original papers.
| Name      | Description                                      |
|-----------|--------------------------------------------------|
| CPU       | 8 vCPUs, 30GB RAM                                |
| GPU       | NVIDIA Quadro P5000 Graphics Card, 16GB          |
| Processor | Intel Xeon                                       |
| Clock Speed | 2.60 GHz                                   |

Table 11: Details of computing resources.

A.2.1 Text Classification Models

**CNN (Kim, 2014)** Dropout = 0.5, Number of Epochs = 20, Batch Size = 128, Learning Rate = 0.001, Number of Filters = 256, Filter Sizes = 2,3,4, Average Running Time = 5s.

**RCNN (Lai et al., 2015)** Dropout = 1.0, Number of Epochs = 10, Batch Size = 128, Learning Rate = 0.001, Hidden Size = 256, Number of Layers = 1, Average Running Time = 5s.

**BiLSTM (Liu et al., 2016)** Dropout = 0.5, Number of Epochs = 10, Batch Size = 128, Learning Rate = 0.001, Hidden Size = 128, Number of Layers = 2, Average Running Time = 5.4s.

**+ Attention (Zhou et al., 2016)** Dropout = 0.5, Number of Epochs = 10, Batch Size = 128, Learning Rate = 0.001, Hidden Size = 128 and 64 respectively, Number of Layers = 2, Average Running Time = 5.74s.

**FastText (Joulin et al., 2017)** Dropout = 0.5, Number of Epochs = 20, Batch Size = 128, Learning Rate = 0.001, Hidden Size = 256, Average Running Time = 22.5s.

**DPCNN (Johnson and Zhang, 2017)** Dropout = 0.5, Number of Epochs = 20, Batch Size = 128, Learning Rate = 0.001, Number of Filter = 250, Average Running Time = 5s.

**Transformer (Vaswani et al., 2017)** Dropout = 0.5, Number of Epochs = 20, Batch Size = 128, Learning Rate = 0.0005, Number of Head = 5, Number of Encoder = 2, Average Running Time = 6.528s.

**BERT-tiny (Jiao et al., 2019)** Dropout = 0.1, Number of Epoch = 20, Batch Size = 64, Learning Rate = 0.00002, Size of Embedding = 384, Feedforward Size = 1536, Hidden Size = 384, Number of Head = 6, Number of Layer = 3, Average Running Time = 20.63s.

**BERT-small (Turc et al., 2019)** Dropout = 0.5, Number of Epoch = 20, Batch Size = 64, Learning Rate = 0.00002, Size of Embedding = 512, Feedforward Size = 2048, Hidden Size = 512, Number of Head = 8, Number of Layer = 6, Average Running Time = 36.65s.

**BERT-base (Devlin et al., 2019)** Dropout = 0.1, Number of Epoch = 10, Batch Size = 64, Learning Rate = 0.00002, Size of Embedding = 768, Feedforward Size = 3072, Hidden Size = 768, Number of Head = 12, Number of Layer = 12, Average Running Time = 99s.

A.2.2 Information Extraction Models

**HMM (Rabiner and Juang, 1986)** Uniform Distribution for Initialisation, Average Total Running Time = 8.47s.

**CRF (Lafferty et al., 2001)** LBFGS algorithm, $c_1 = 0.1$, $c_2 = 0.1$, Max Iteration = 100, Average Total Running Time = 11.72s.

**BiLSTM (Huang et al., 2015)** Number of Epoch = 30, Batch Size = 64, Learning Rate = 0.001, Size of Embedding = 128, Average Running Time = 8.43s.

**BiLSTM-CRF (Lample et al., 2016)** Number of Epoch = 30, Batch Size = 64, Learning Rate = 0.001, Size of Embedding = 128, Average Running Time = 9.35s.

**BERT Models** We use the same hyper-parameter settings as used in the text classification models with the exception of Batch Size = 16. The average running time of BERT-tiny, BERT-small and BERT-base for information extraction are 34.00s, 55.73s, and 120s respectively.

A.3 Details of Baseline Experiments

Table 12 shows the statistics of each fold.

| Fold | Text Classification | Information Extraction |
|------|---------------------|-------------------------|
|      | # of Lines | Total | # of Characters | Total |          |          |
| 0    | 277/685 | 962 | 1666/20010 | 216/66 |
| 1    | 290/854 | 1144 | 1703/19354 | 210/57 |
| 2    | 285/709 | 1074 | 1666/19009 | 212/75 |
| 3    | 358/909 | 1267 | 2251/23535 | 257/66 |
| 4    | 459/1181 | 1640 | 2632/28025 | 306/57 |

Table 12: Statistics of each fold in the baseline experiments. The number before slash indicates how many actor’s lines or characters that make the audience laugh. The number after slash indicates the number of lines or characters without causing audiences laugh.
Table 13: Each fold’s text classification baseline experiments and their overall average performance. \( P \), \( R \), \( F \) and \( A \) are Precision, Recall, F1-score and Accuracy respectively.

Table 14: Each fold’s information extraction baseline experiments and their overall average performance.

A.4 Details of Cross-Style Experiments

Table 15 shows the statistics of the scripts performed by specific leading comedians.

A.5 Full List of Selected Comedy Performances

Tables 18 and 19 show the full list of performances in our corpus with details.
| Model   | 0       | 1       | 2       | 3       | 4       | 5       | Average |
|---------|---------|---------|---------|---------|---------|---------|---------|
| CNN     | 41.99   | 40.00   | 41.85   | 39.67   | 45.70   | 47.29   | 42.75   |
| BiLSTM  | 38.17   | 42.18   | 39.35   | 36.26   | 45.26   | 46.90   | 41.35   |
| BERT    | 40.73   | 37.41   | 42.38   | 40.81   | 45.30   | 51.40   | 43.01   |
| BERT-tiny | 39.08  | 39.67   | 42.97   | 48.40   | 43.22   | 56.38   | 44.95   |
| RCNN    | 46.93   | 44.01   | 36.94   | 41.30   | 46.50   | 46.41   | 41.57   |
| Transformer | 41.97 | 36.32   | 35.20   | 34.22   | 46.26   | 43.10   | 42.32   |
| FastText| 42.01   | 42.21   | 36.03   | 35.25   | 45.33   | 46.71   | 41.26   |
| DPCNN   | 40.40   | 38.72   | 48.68   | 48.05   | 57.31   | 47.28   |
| + Attention | 42.70  | 38.16   | 45.06   | 38.49   | 48.02   | 46.78   | 43.20   |
| RCNN    | 51.38   | 69.74   | 66.48   | 80.61   | 64.31   | 65.69   |
| BiLSTM  | 51.75   | 47.02   | 62.56   | 70.33   | 75.15   | 75.77   | 60.76   |
| BERT    | 62.22   | 54.40   | 82.05   | 75.15   | 79.39   | 63.76   | 66.82   |
| BERT-tiny | 61.80  | 54.40   | 82.05   | 75.15   | 79.39   | 63.76   | 66.82   |
| RCNN    | 65.87   | 51.38   | 82.05   | 75.15   | 79.39   | 63.76   | 66.82   |
| BiLSTM  | 62.22   | 54.40   | 82.05   | 75.15   | 79.39   | 63.76   | 66.82   |
| BERT    | 62.50   | 44.40   | 42.73   | 48.68   | 48.05   | 57.31   | 47.28   |
| + Attention | 42.70  | 38.16   | 45.06   | 38.49   | 48.02   | 46.78   | 43.20   |

Table 16: Performance of text classification in predicting the scripts performed by specific leading comedians (0: Xiaobao SONG, 1: Yuepeng YUE, 2: Ling JIA, 3: Xiaofei ZHANG, 4: Teng SHEN, 5: Bing JIA).

| Model   | 0       | 1       | 2       | 3       | 4       | 5       | Average |
|---------|---------|---------|---------|---------|---------|---------|---------|
| CNN     | 21.67   | 19.47   | 22.03   | 16.69   | 9.50    | 8.79    | 12.11   |
| CRF     | 22.72   | 20.54   | 39.79   | 21.43   | 18.28   | 28.96   | 25.29   |
| BILSTM  | 26.32   | 21.48   | 31.70   | 24.53   | 32.39   | 37.89   | 30.05   |
| BiLSTM-CRF | 26.32 | 23.35   | 33.08   | 31.86   | 25.71   | 44.02   | 30.72   |
| BERT    | 24.43   | 21.87   | 29.61   | 26.37   | 27.19   | 28.32   | 26.30   |
| BERT-small | 23.93  | 20.19   | 26.01   | 23.85   | 22.80   | 32.81   | 24.93   |
| BERT-base | 20.48  | 19.10   | 23.74   | 24.41   | 26.58   | 33.51   | 24.64   |
| HMM     | 6.91    | 6.26    | 6.96    | 9.30    | 8.79    | 6.68    | 7.52    |
| CRF     | 5.80    | 4.87    | 7.47    | 4.41    | 3.13    | 4.48    | 5.03    |
| BILSTM  | 4.40    | 2.17    | 2.40    | 2.20    | 3.71    | 2.49    | 2.90    |
| BiLSTM-CRF | 7.64  | 14.06   | 17.52   | 7.30    | 3.20    | 2.90    | 8.77    |
| BERT    | 16.88   | 19.89   | 22.26   | 19.15   | 19.06   | 17.98   | 19.20   |
| BERT-small | 22.63  | 31.31   | 29.31   | 32.78   | 32.42   | 15.39   | 27.31   |
| BERT-base | 39.61  | 36.43   | 35.25   | 32.01   | 28.64   | 17.95   | 31.65   |
| HMM     | 10.48   | 9.48    | 10.58   | 12.11   | 10.97   | 10.47   | 10.68   |
| CRF     | 9.24    | 7.87    | 12.59   | 7.31    | 5.34    | 7.76    | 8.35    |
| BILSTM  | 7.60    | 3.94    | 4.46    | 4.04    | 6.65    | 4.67    | 5.23    |
| BiLSTM-CRF | 11.87 | 17.55   | 22.90   | 11.88   | 5.70    | 5.44    | 12.56   |
| BERT    | 19.97   | 20.83   | 25.42   | 22.19   | 22.41   | 22.00   | 22.14   |
| BERT-small | 23.26  | 24.55   | 27.56   | 27.61   | 26.77   | 20.95   | 25.12   |
| BERT-base | 27.00  | 25.06   | 28.37   | 27.70   | 27.57   | 23.37   | 26.51   |

Table 17: Performance of information extraction in predicting the scripts performed by specific leading comedians.
Table 18: Full list of the selected comedy performances with their titles, leading comedians, and source.

| Performance Title | Translation | Year | Duration (minutes) | Number of Lines | Laughter Rate (Line-Level) | Number of Characters | Laughter Rate (Character-Level) |
|-------------------|-------------|------|--------------------|-----------------|---------------------------|---------------------|----------------------------------|
| Help Her Up or Not | Teng SHEN etc. | 2014 | 14.75 | 114 | 44.73% | 3481 | 10.25% |
| A Bowl of Yuanxiao | Xiaofei ZHANG etc. | 2018 | 17.58 | 177 | 31.07% | 4465 | 7.57% |
| Manager JIA's Presentation | Bing JIA etc. | 2017 | 23.5 | 228 | 36.4% | 5319 | 9.64% |
| Love Back Home | Ling JIA etc. | 2016 | 8.5 | 81 | 33.33% | 1894 | 7.44% |
| SDTV-1 Chinese New Year Gala | | | | | | | |
| CCTV Chinese New Year Gala | | | | | | | |
| Tomato TV Chinese New Year Gala | | | | | | | |
| TVS Chinese New Year Gala | | | | | | | |
| CCTV Chinese New Year Gala | | | | | | | |
| CCTV Chinese New Year Gala | | | | | | | |
| CCTV Chinese New Year Gala | | | | | | | |
| CCTV Chinese New Year Gala | | | | | | | |
| CCTV Chinese New Year Gala | | | | | | | |

Table 19: Full list of the selected comedy performances with their titles, years, duration, number of lines/characters and laughter rate at line/character-level.

| Performance Title | Translation | Year | Duration (minutes) | Number of Lines | Laughter Rate (Line-Level) | Number of Characters | Laughter Rate (Character-Level) |
|-------------------|-------------|------|--------------------|-----------------|---------------------------|---------------------|----------------------------------|
| Help Her Up or Not | Teng SHEN etc. | 2014 | 14.75 | 114 | 44.73% | 3481 | 10.25% |
| A Bowl of Yuanxiao | Xiaofei ZHANG etc. | 2018 | 17.58 | 177 | 31.07% | 4465 | 7.57% |
| Manager JIA's Presentation | Bing JIA etc. | 2017 | 23.5 | 228 | 36.4% | 5319 | 9.64% |
| Love Back Home | Ling JIA etc. | 2016 | 8.5 | 81 | 33.33% | 1894 | 7.44% |
| SDTV-1 Chinese New Year Gala | | | | | | | |
| CCTV Chinese New Year Gala | | | | | | | |
| Tomato TV Chinese New Year Gala | | | | | | | |
| TVS Chinese New Year Gala | | | | | | | |
| CCTV Chinese New Year Gala | | | | | | | |
| CCTV Chinese New Year Gala | | | | | | | |
| CCTV Chinese New Year Gala | | | | | | | |
| CCTV Chinese New Year Gala | | | | | | | |
| CCTV Chinese New Year Gala | | | | | | | |

52