Keeping Their Words: Direct and Indirect Chinese Quote Attribution from Newspapers

Kuan-Lin Lee
Department of Psychology, National Chengchi University
104702016@nccu.edu.tw

Yu-Chung Cheng
Department of Journalism, National Chengchi University
yuchungc@nccu.edu.tw

Pai-Lin Chen
Department of Journalism, National Chengchi University
pailinch@nccu.edu.tw

Hen-Hsen Huang*
Department of Computer Science,
National Chengchi University
hhhuang@nccu.edu.tw

ABSTRACT

Quote attribution plays an important role in emerging research topics such as fact checking, stance detection, and argument mining. This work explores Chinese quote attribution from newspapers. Both direct and indirect quotes are addressed by a text-encoder based sequence labeling model. We create a dataset for empirical analysis. Experimental results show the effectiveness of our model.

CCS CONCEPTS

• **Computing methodologies** → **Information extraction**; **Discourse, dialogue and pragmatics**.

KEYWORDS

quote extraction, quote attribution, computational journalism

ACM Reference Format:
Kuan-Lin Lee, Yu-Chung Cheng, Pai-Lin Chen, and Hen-Hsen Huang. 2020. Keeping Their Words: Direct and Indirect Chinese Quote Attribution from Newspapers. In *Companion Proceedings of the Web Conference 2020 (WWW '20 Companion), April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3366424.3382716

1 INTRODUCTION

Quote attribution is a natural language processing (NLP) task aimed at extracting quotes and attributing those quotes to their speakers [4], determining "who said what" [3]. Public figures' statements reported in newspapers show important information including their opinions, stances, plans, and achievements. Various research topics including fact checking, stance detection, and argument mining can be explored based on the quotes [2, 8].

However, only few previous studies on the task of quote extraction and attribution [3–6]. Quote extraction is not a straightforward task. Even direct quote extraction cannot be completely tackled with a deterministic pattern matching algorithm because of the ambiguous usage of quotation marks [6]. Moreover, indirect quote extraction is even more challenging, since the lack of explicit marks identifying the boundary of a quote [5].

This work focuses on Chinese quote attribution. Similar to English, the Chinese opening quote mark—‘ “ ’ or ‘ 「’ —and the closing quote mark—‘ ” ’ or ‘」’—explictly denote the beginning and the end of a direct quote, respectively. In practical, however, a quote may be indirect or even partly direct and partly indirect. The following example shows an instance. The phrase "陳妻" (Chen's wife) presents the speaker, the underlined fragment is indirect, and the fragment enclosed in quote marks is direct. The two fragments form a complete quote of the speaker. In addition to the quote itself, quote attribution also identifies the speaker of the quote as mandatory information.

> 陳妻否認丈夫的指控，指稱三個兒子從小都是由她照顧，丈夫很少在家，且在外結交女友，甚至將她打傷，「丈夫沒有權利要求離婚啊」。
> Chen's wife denied her husband's allegations, alleging that her three sons had been taken care of by her since their childhood. Her husband was rarely at home, had a girlfriend outside, and even injured her. "[My] husband has no right to ask for a divorce."

To the best of our knowledge, this is the first study in Chinese news quote attribution. We analyze the characteristics of Chinese quotation and establish a dataset by inviting domain experts to annotate news articles from different newspapers. Then, we propose a neural network-based model for extracting the direct and the indirect quotes and identifying their speakers given a news article. The contributions of this paper are threefold as follows.

(1) We explore an important task that provides essential information for a wide range of emerging research topics.
(2) The challenging issues in Chinese quotation are addressed.
(3) A robust model is proposed for Chinese quote attribution. Experimental results show its effectiveness.

2 DATASET CONSTRUCTION

We collect a total of 2,000 news articles published by five newspapers in Taiwan in November 2018. Four experts in the journalism domain are invited to label the quotes and their speakers from the news articles. All the annotators will label common 10 articles for inter-rater agreement measurement. The agreements of quotes and speakers are 0.86 and 0.91, respectively. The resulting dataset contains a total of 4,716 quotes.

*Corresponding author

WWW '20 Companion, April 20–24, 2020, Taipei, Taiwan

ACM ISBN 978-1-4503-7024-0/20/04.
https://doi.org/10.1145/3366424.3382716

3 METHODOLOGY

We formulate Chinese quote extraction and speaker identification as a single sequence labeling task at the character level. Specifically, our model is trained to classify each Chinese character as one of the following five labels:

- **S**: The speaker of a quote.
- **B**: The beginning of a quote.
- **E**: The end of a quote.
- **I**: Inside of the quote.
- **O**: Not a part of a quote/speaker.

Text-encoder based models show effectiveness in many NLP tasks. Our sequence labeling model is based on the pre-trained Chinese BERT text-encoder [1]. Given a paragraph in a news article, the text-encoder is employed to generate the context-aware embedding for each Chinese character. In other words, the paragraph consisting of n Chinese characters will be converted to an $n \times 768$ matrix for representing the semantic information in the paragraph. Then, additional dense layers are added, and the last layer is the softmax one that outputs the probabilities of the five labels for each of the n characters. The Chinese BERT text-encoder is fine-tuned during the training.

We train the model as conditional random field (CRF) for finding the optimal label sequence over the entire paragraph [9]. For initializing the transition probabilities of the CRF model, the default method is to randomly sample a value from a uniform distribution. We also try an alternative initialization by computing the prior transition probabilities from the training instances.

4 EXPERIMENTS

The dataset is split into training and test sets. The vanilla BERT model is the baseline. We measure the performances with both strict matching and partial matching evaluation [7]. The metrics include precision, recall, and F-score. Table 1 shows the results of quote extraction. Our best model, BERT-CRF with prior transition probabilities initialized, achieves an F-score of 70% in terms of partial matching. Table 2 shows the results of speaker identification. Our best model achieves an F-score of 75%. Experimental results show the improvement made by CRF. Initializing transition probabilities with prior distribution is also effective.

Table 3 compares the performances of our best model on different newspapers. For quote extraction, the news articles from United Daily are the easiest, and those from Apple Daily are the most difficult. This observation matches our experience because the writing style of United Daily is considered most formal. In contrast, Apple Daily is widely-considered less formal, making the model difficult to extract.

5 CONCLUSION

Quote attribution provides essential information for emerging research topics including fact checking, stance detection, and argument mining. This work is the first one to explore Chinese quote attribution in newspapers. We create a dataset and propose a robust model with the latest NLP technologies for extracting direct and indirect quotes and identifying their speakers.

Table 1: Results of Chinese quote extraction in strict matching and partial matching evaluation, reported in Precision (P), Recall (R), and F-score (F)

Model	Strict			Partial		
	P	R	F	P	R	F
BERT	0.60	0.50	0.54	0.72	0.61	0.66
BERT-CRF (random)	0.59	0.52	0.55	0.72	0.64	0.68
BERT-CRF (prior)	0.58	0.57	0.58	0.70	0.71	0.70

Table 2: Results of speaker identification

Model	Precision	Recall	F1-score
BERT	0.61	0.82	0.70
BERT-CRF (random)	0.62	0.83	0.71
BERT-CRF (prior)	0.69	0.83	0.75

Table 3: Results of the five newspapers, reported in F-score

Newspaper	Strict	Partial	Speaker
China Times	0.59	0.72	0.76
ETtoday	0.56	0.68	0.69
United Daily	0.64	0.75	0.80
Liberty Times	0.57	0.70	0.74
Apple Daily	0.49	0.67	0.75

ACKNOWLEDGMENTS

This research was supported by the Ministry of Science and Technology, Taiwan, under the grants MOST-105-2420-H-004-032-MY3, MOST-108-2634-F-002-008-, and MOST-109-2634-F-002-034-.

REFERENCES

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL 2019*. 4171–4186.
[2] Dan Goldwasser and Xiao Zhang. 2016. Understanding Satirical Articles Using Common-Sense. *Transactions of the Association for Computational Linguistics* 4 (2016), 537–549.
[3] Grace Muzny, Michael Fang, Angel Chang, and Dan Jurafsky. 2017. A Two-stage Sieve Approach for Quote Attribution. In *EACL 2017*. 460–470.
[4] Timothy O'Keefe, Silvia Pareti, James R. Curran, Irena Koprinska, and Matthew Honnibal. 2012. A Sequence Labelling Approach to Quote Attribution. In *EMNLP 2012*. 790–799.
[5] Silvia Pareti, Tim O'Keefe, Ioannis Konstas, James R. Curran, and Irena Koprinska. 2013. Automatically Detecting and Attributing Indirect Quotations. In *EMNLP 2013*. 989–999.
[6] Andrew Salway, Paul Meurer, Knut Hofland, and Øystein Reigem. 2017. Quote Extraction and Attribution from Norwegian Newspapers. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*. 293–297.
[7] Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. 2013. SemEval-2013 Task 9: Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013). In *SemEval 2013*. 341–350.
[8] T. Traylor, J. Straub, Gurmeet, and N. Snell. 2019. Classifying Fake News Articles Using Natural Language Processing to Identify In-Article Attribution as a Supervised Learning Estimator. In *IEEE 13th International Conference on Semantic Computing*. 445–449.
[9] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip H. S. Torr. 2015. Conditional Random Fields as Recurrent Neural Networks. In *ICCV 2015*.