The CRECIL Corpus: a New Data Set for Extraction of Relations between Characters in Chinese Multi-party Dialogues

Yuru Jiang*, Yang Xu*, Yuhang Zhan*, Weikai He*, Yilin Wang* , Zixuan Xi*,
Meiyun Wang*, Xinyu Li*, Yu Li*, and Yanchao Yu†
* jiangyuru@bistu.edu.cn † y.yu@napier.ac.uk

Introduction

We describe a new freely available Chinese multi-party dialogue corpus for automatic extraction of dialogue-based character relationships:

- extracted from the original TV scripts of a Chinese sitcom called “I Love My Family".
- contains ***complex family-based*** human daily spoken conversations in Chinese
- introduced human annotation scheme for both global Character relationship map and character reference relationship – between 140 entities.

We also carried out a data exploration experiment by deploying a BERT-based model to extract character relationships on the CRECIL corpus and another existing relation extraction corpus (DialogRE[1]).

- extracting character relationships is more challenging in CRECIL than in DialogRE.

Referential Relationship

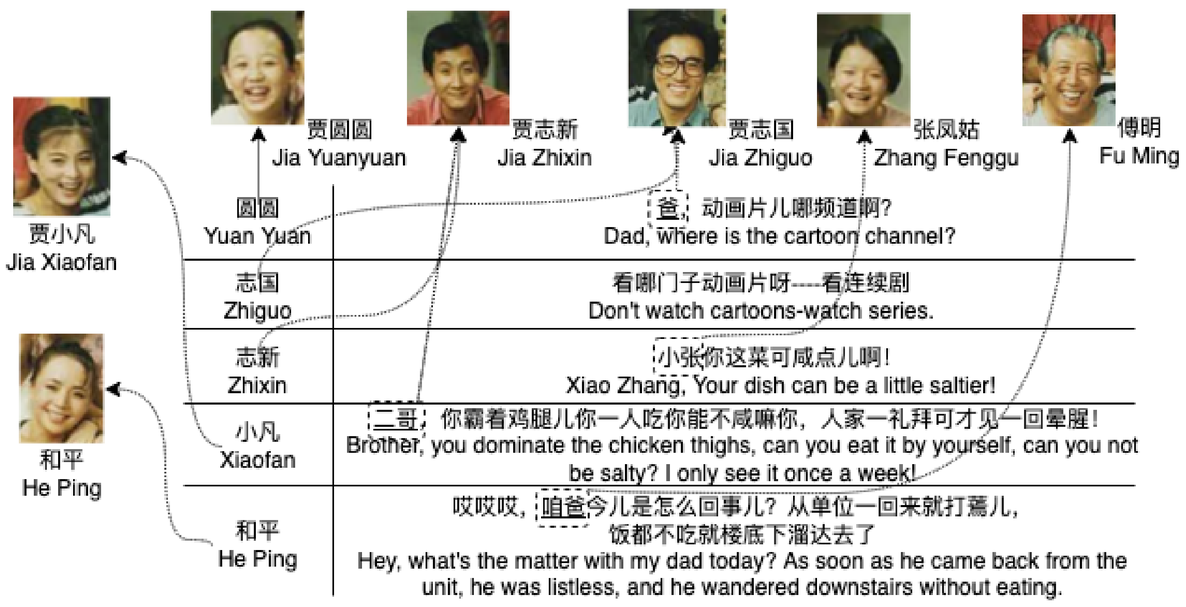

Figure 2: Schematic diagram of referential relationship labelling in the CRECIL corpus

Dialogue-based Character Relationship Triples Generator

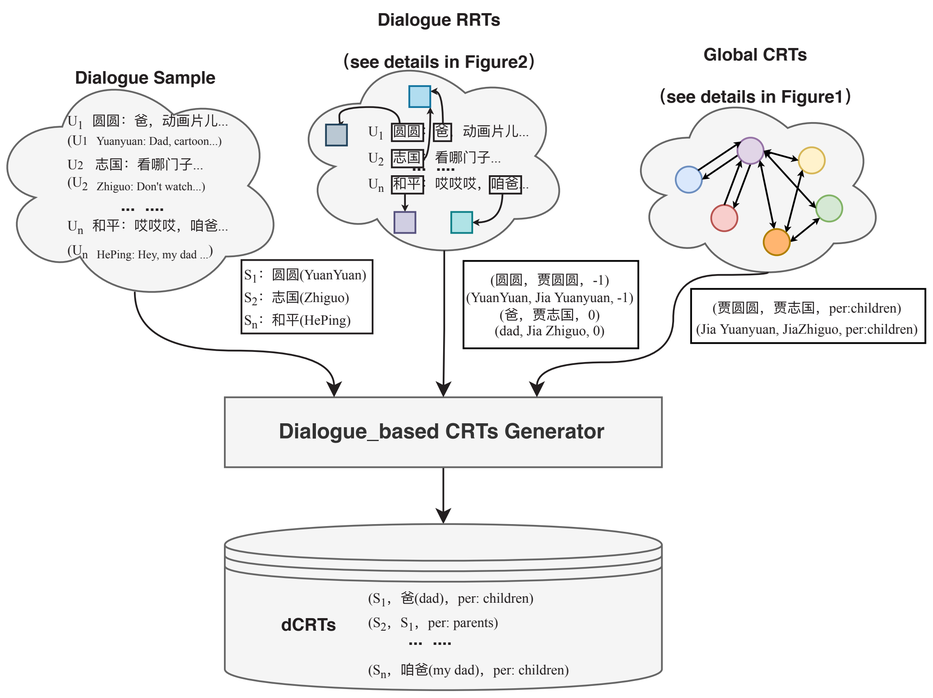

Figure 3: Example of Dialogue-based Character Relationship Triples Generator in the CRECIL corpus.

Comparison with DialogRE

CRECIL	
Average turns length(in tokens)	23.8
Average dialogue length(in tokens)	707.6
Average # of turns	29.7
Average # of speakers	4.1
sAverage # of sentences	39.4
Average # of relational instances	57.4
Average # of no-relational instances	21.6

Table 3.1 : Statistics per dialogue of CRECIL.

DialogRE	
Average dialogue length (in tokens)	225.8
Average # of turns	12.9
Average # of speakers	3.3
Average # of sentences	21.8
Average # of relational instances	4.5
Average # of no-relation instances	1.2

Table 3.2 : Statistics per dialogue of DialogRE.

Global Relationship

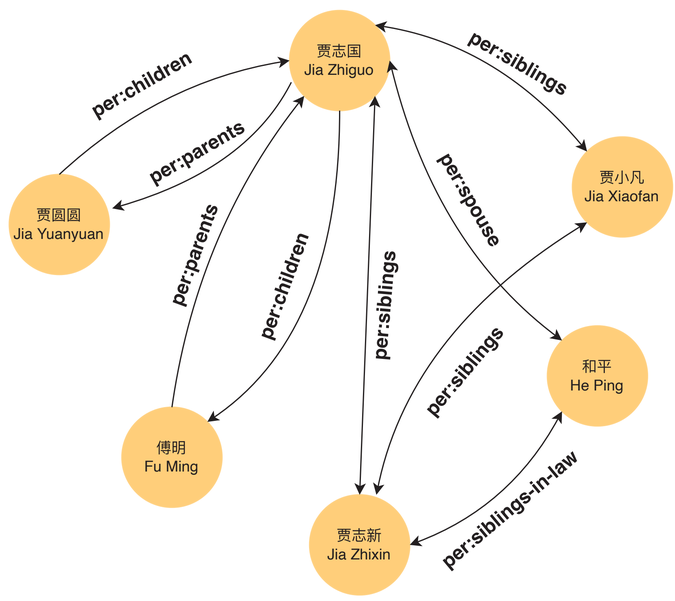

Figure 1: Global Relationship Diagram

Experiment Results

	Alternate Name	Neighbor	Children	Parents	Others
Dev	54.9	64.2	50.0	48.5	51.4
Test	55.7	60.0	47.1	48.6	46.2

Table 1: Comparison between different categories (%) (excluding the ‘unanswerable’ type)

	EN-DialogRE	**CN-DialogRE**	**CRECIL**
Dev	59.4	63.7	56.8
Test	57.9	63.2	54.4

Table 2: Comparison between the CRECIL corpus and the DialogRE corpus (%)

Conclusion & Future Work

- We presented a novel human-annotated dialogue-based relation extraction data set (CRECIL) for multi-party conversations in Chinese.
- We have introduced the Chinese-oriented character relationship categories and labelling rules for annotating the corpus.
- The results demonstrate that extracting character relationships is more challenging in CRECIL than in DialogRE.

Current Work: Explore the character relationship characteristics of Chinese multi-party dialogues and build a better-performance character relationship extraction model

Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant No.61602044).

References

[1] Yu, D., Sun, K., Cardie, C., and Yu, D.(2020). Dialogue-based relation extraction.arXiv preprint arXiv:2004.08056

[2] Yao, Y., Ye, D., Li, P., Han, X., Lin, Y., Liu,Z., Liu, Z., Huang, L., Zhou, J., and Sun,M. (2019). Docred: A large-scale document-level relation extraction dataset. arXiv preprint arXiv:1906.06127.