澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

科技學院
Faculdade de Ciências e Tecnologia
Faculty of Science and Technology

Relation Extraction Method Based on Pre-trained LLM BERT

University of Macau

WU QIWEI, ZHANG YUCI, QIANG ZIYAN

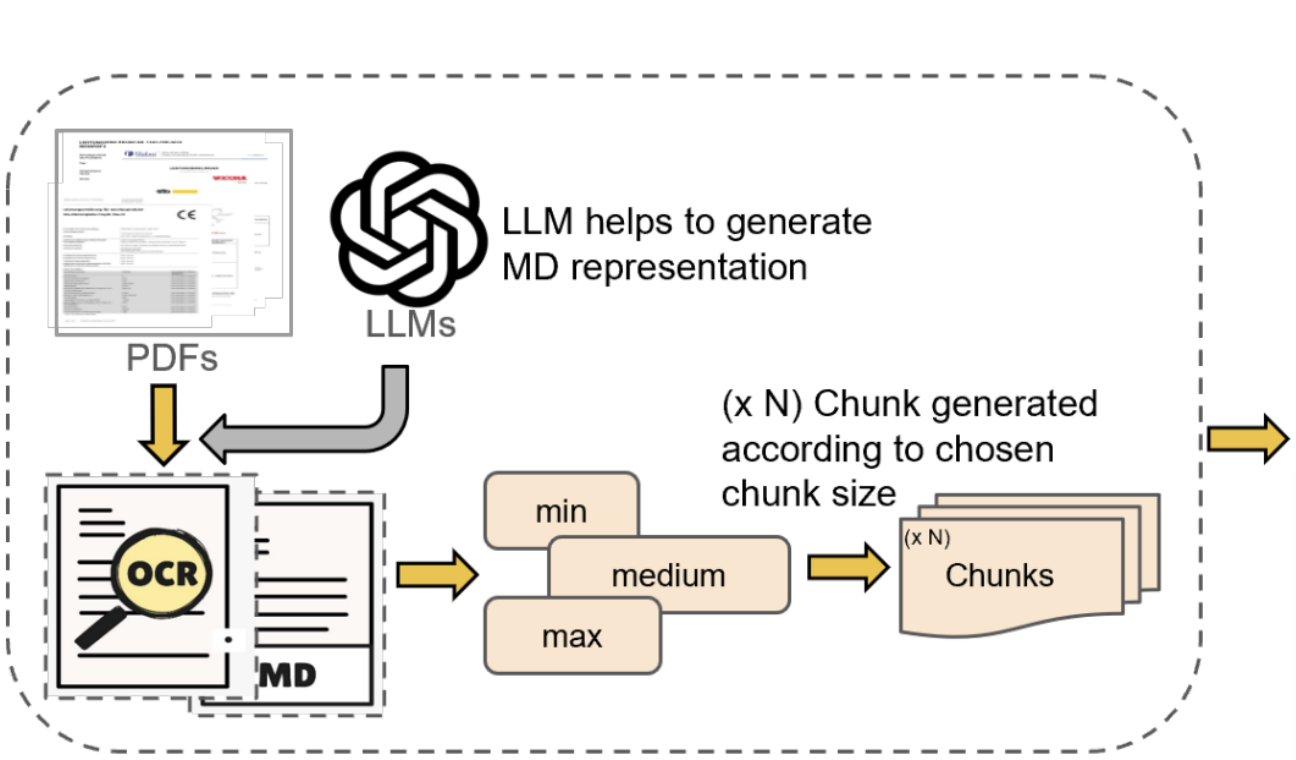
MC56730,MC56685,MC56461

Content

- Introduction
- Related Work
- Task Analysis
- Method Description
- Results Analysis
- Conclusion

BERT

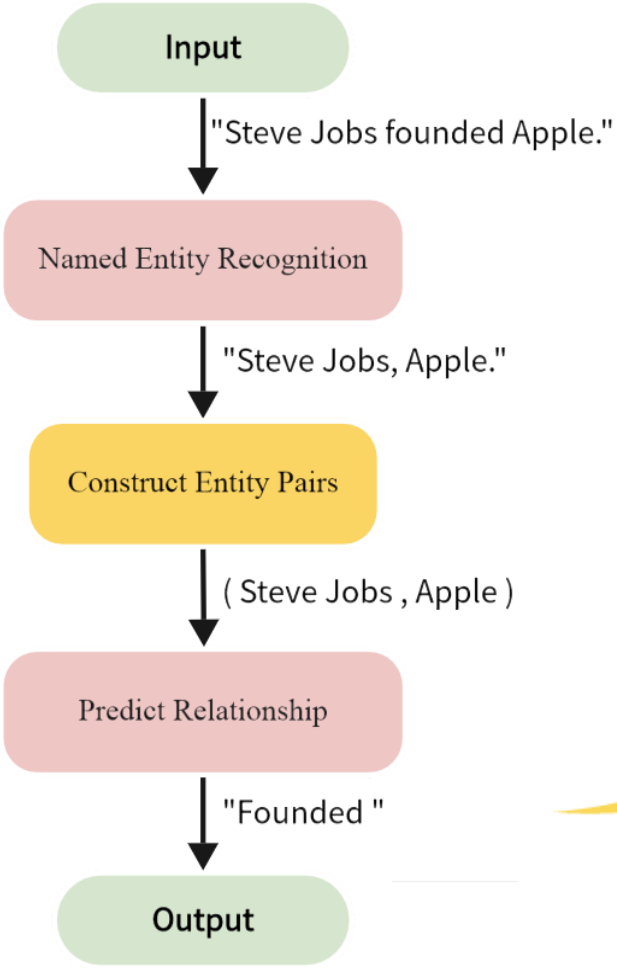
Introduction



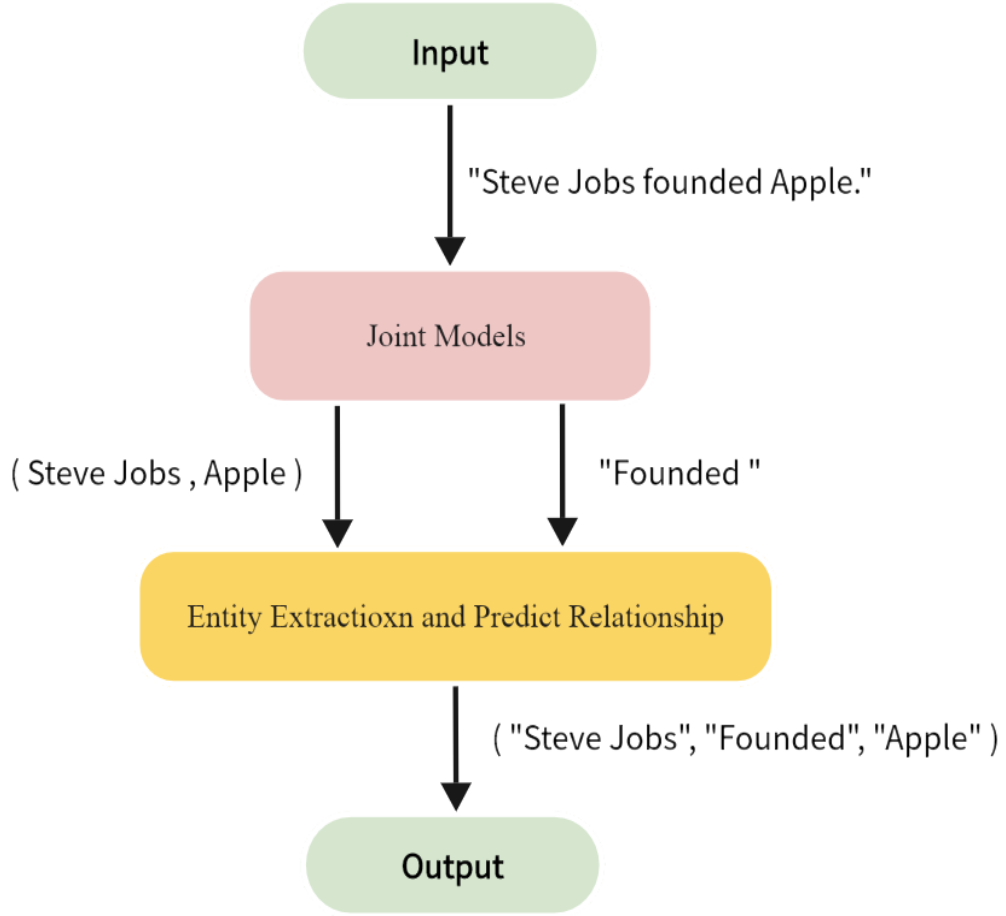
Named Entity Recognition

Relation Extraction

Introduction

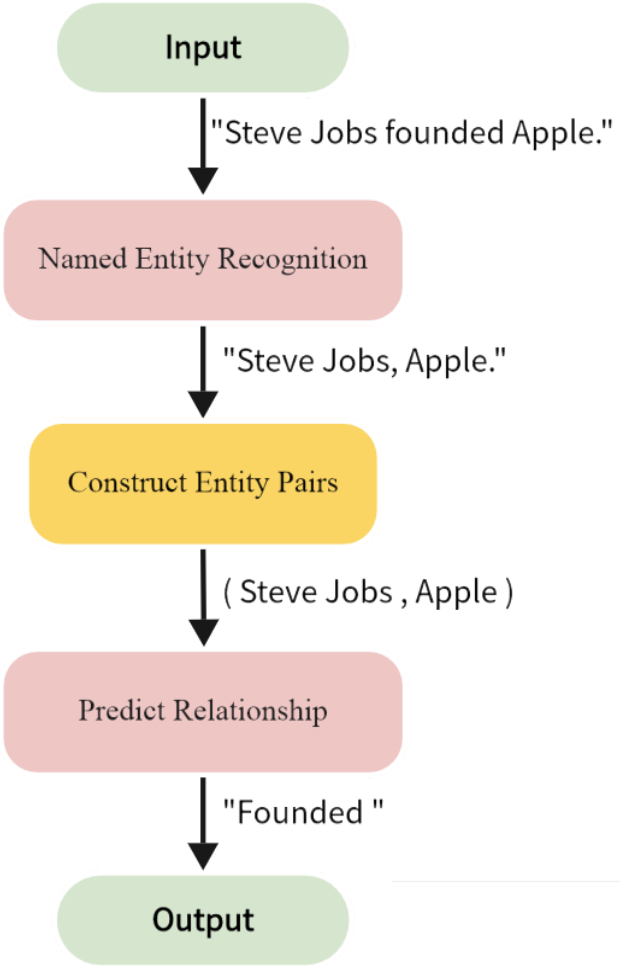


Pipeline-Based Methods

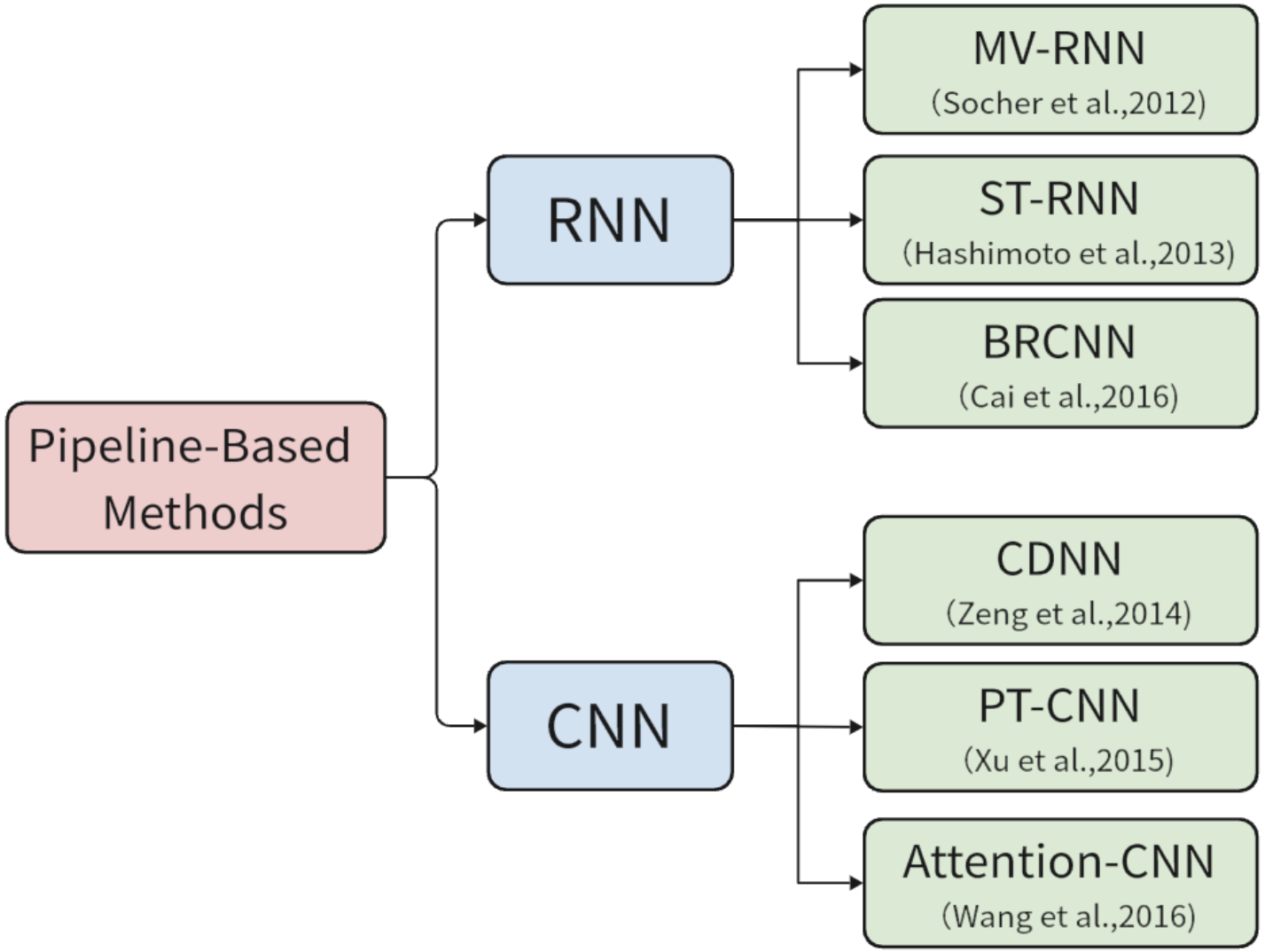


Joint Learning-Based Methods

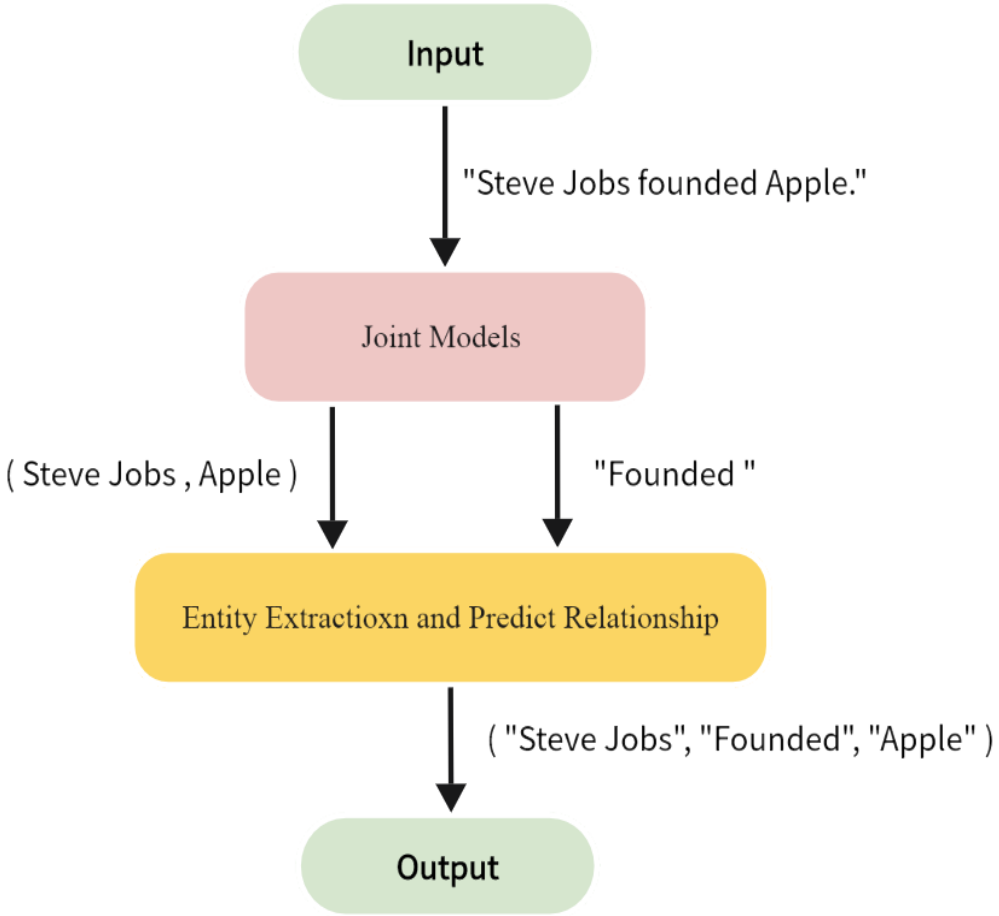
Related Work: Pipeline-Based Methods



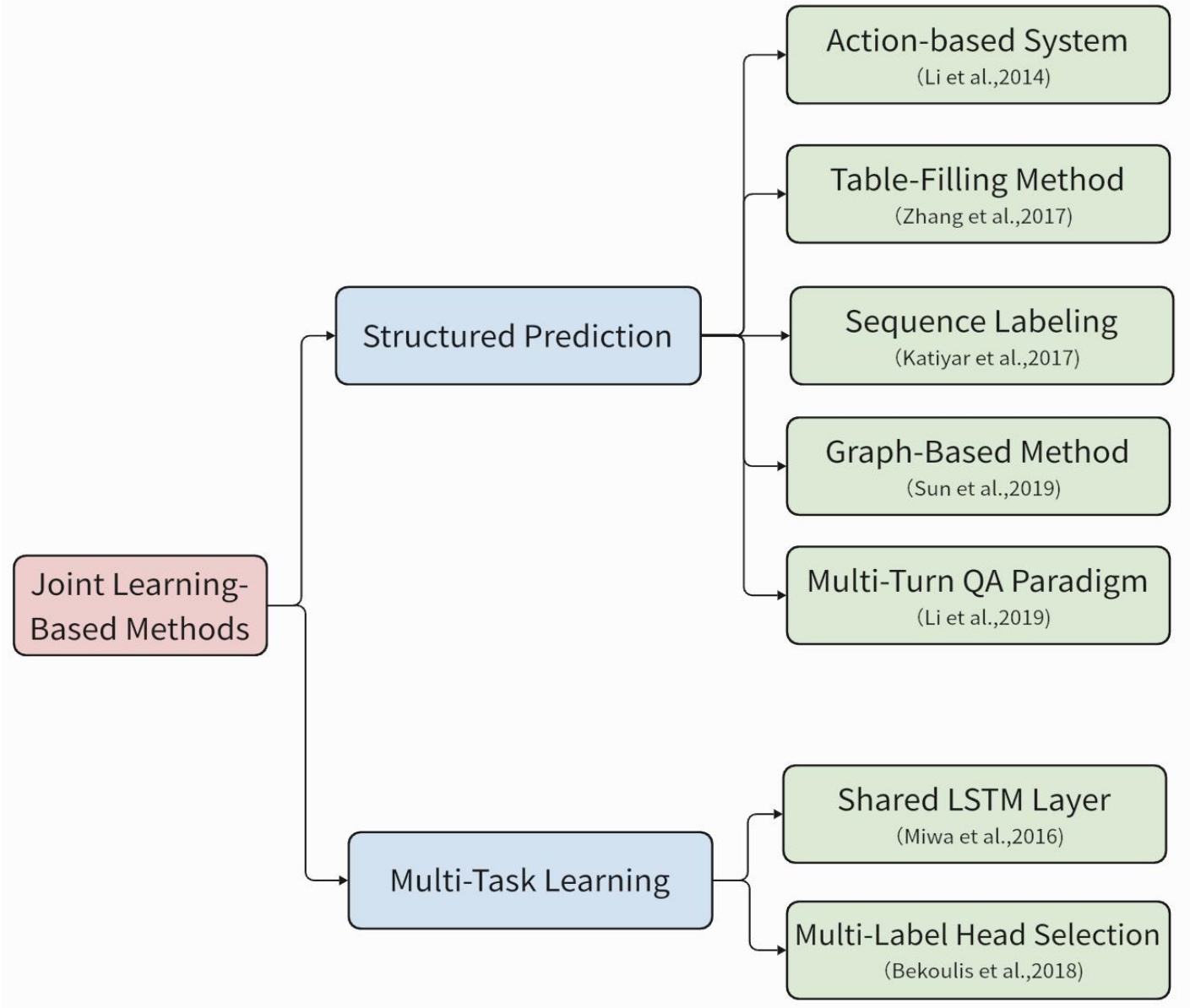
Pipeline-Based Methods



Related Work: Joint Learning-Based Methods



Joint Learning-Based Methods



Task Analysis

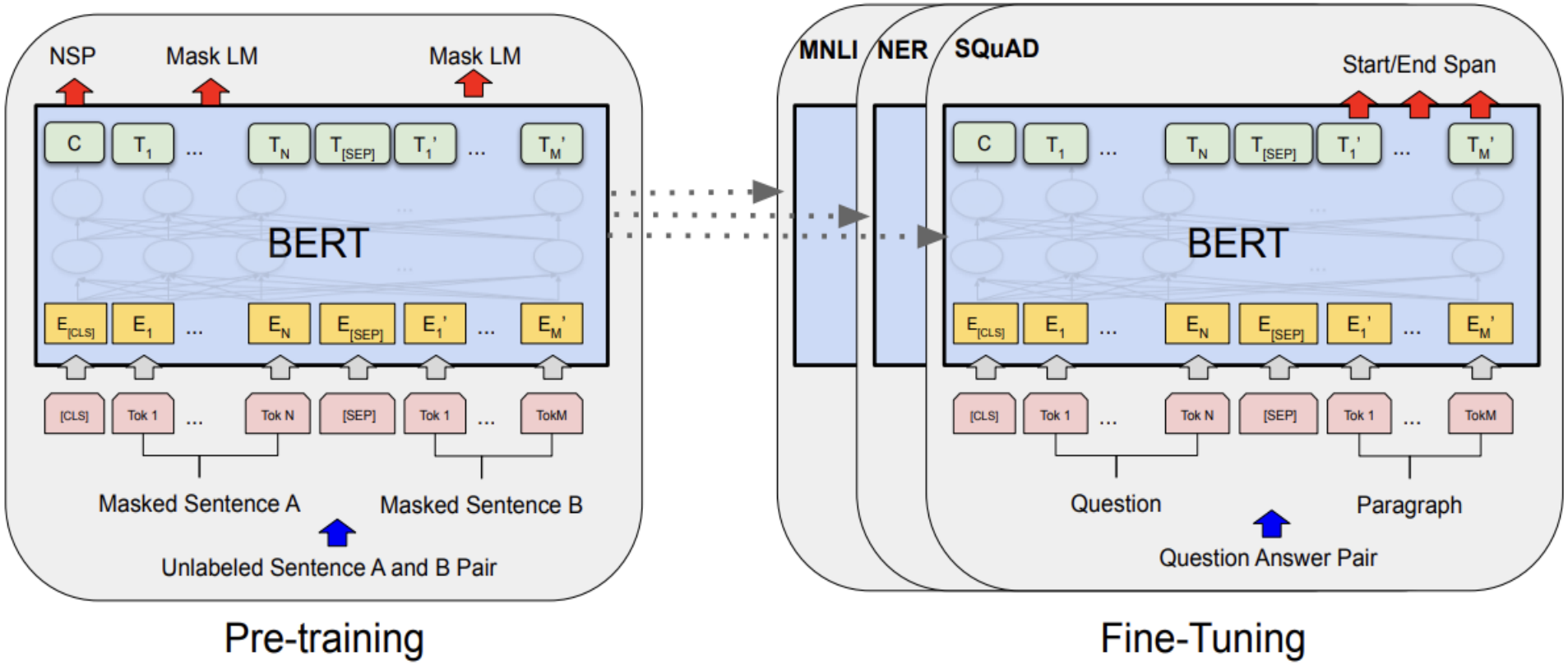
- **Objective:**
  Identify and classify different types of medical relationships from medical text in Chinese.

- **There are 10 relationship categories:**
  Clinical Manifestation, Drug Treatment, Synonym,Etiology,Complication, Pathological Subtype, Laboratory Test, Adjuvant Treatment, Related, and Imaging Test.

Example	
ID	2437
sentence	Food poisoning @ Fever may be caused by extragastrointestinal infection or overlapping infection.
(h)	Food Poisoning
(t)	stress ulcer
(r)	Etiology

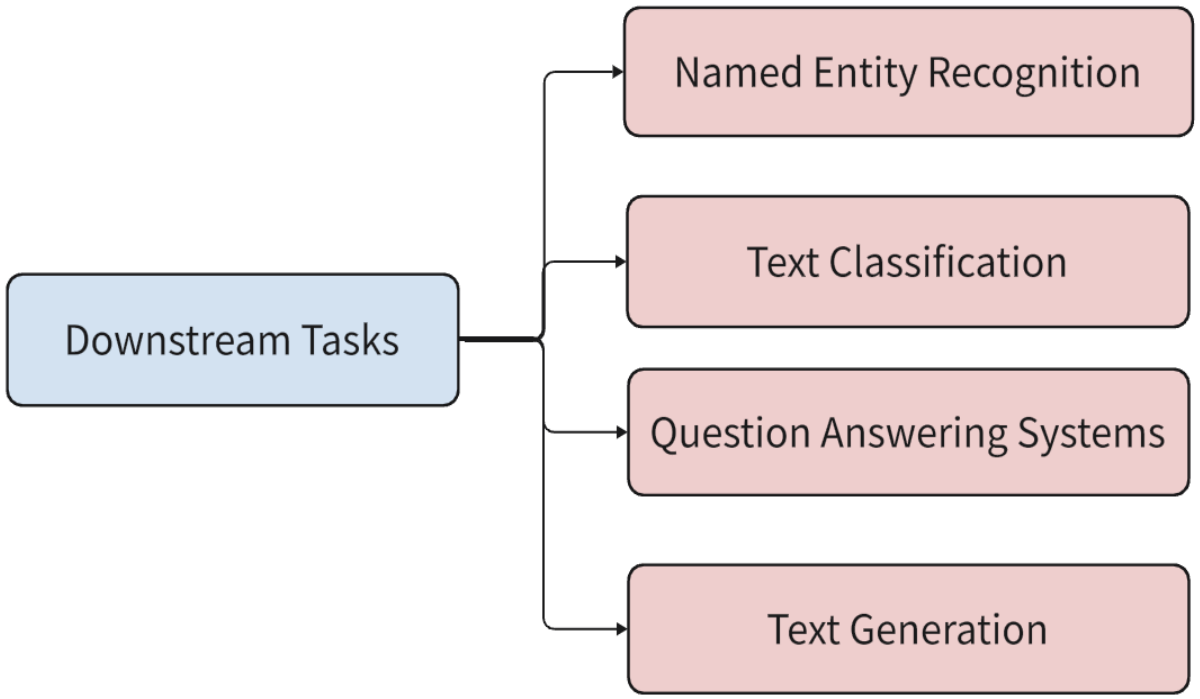
Table 1: Data Example. Each piece of data includes: Sentence, Head Entity (h), Tail Entity (t),and Relationship (r).

Method Description: Pre-trained LLM BERT

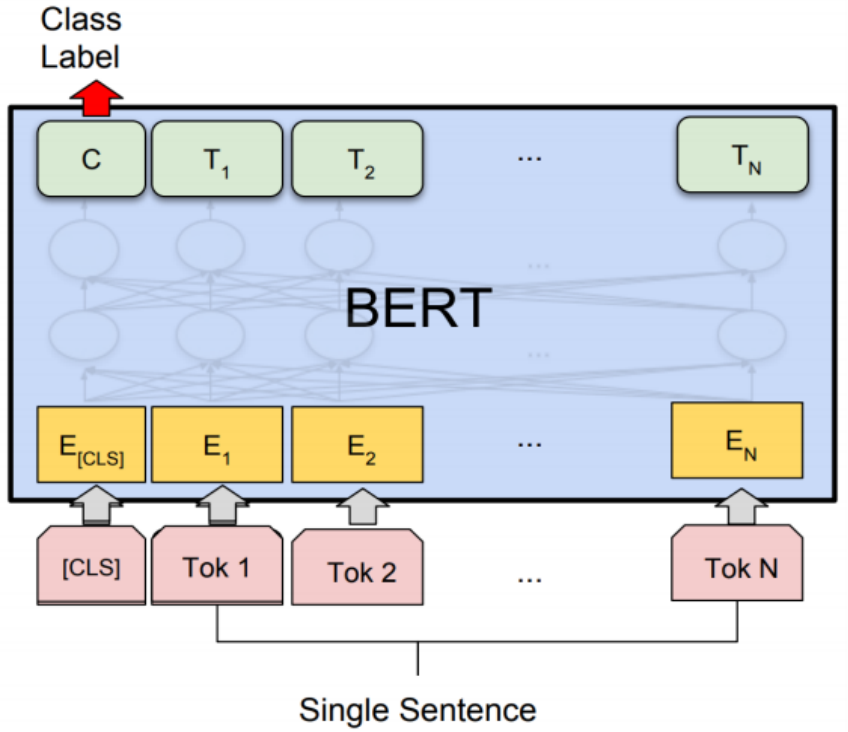


Overall pre-training and fine-tuning procedures for BERT.

Multiple Downstream Tasks of BERT

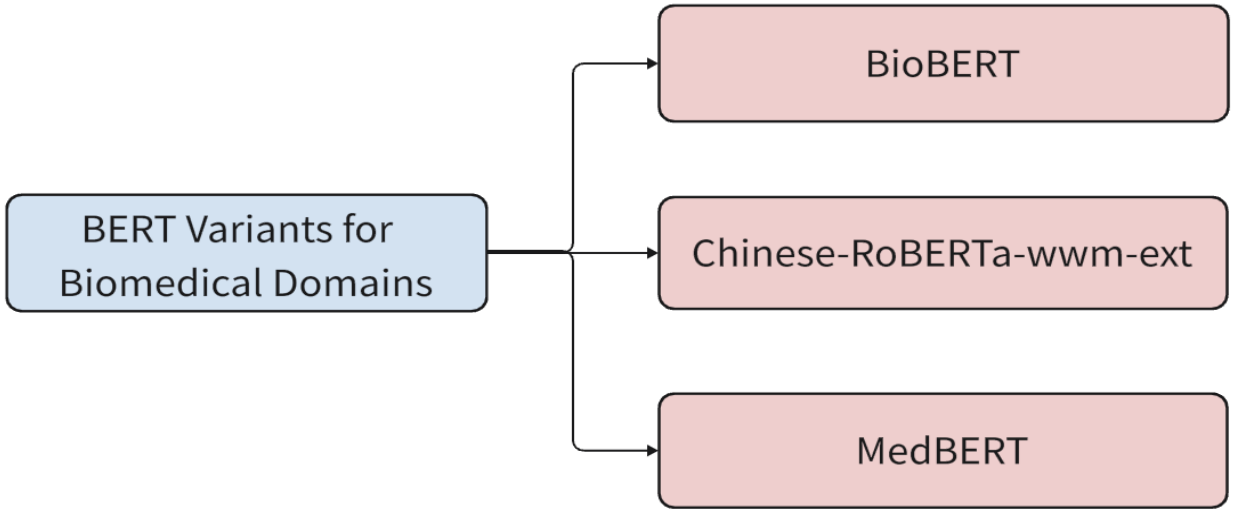


The four main categories of BERT's downstream tasks

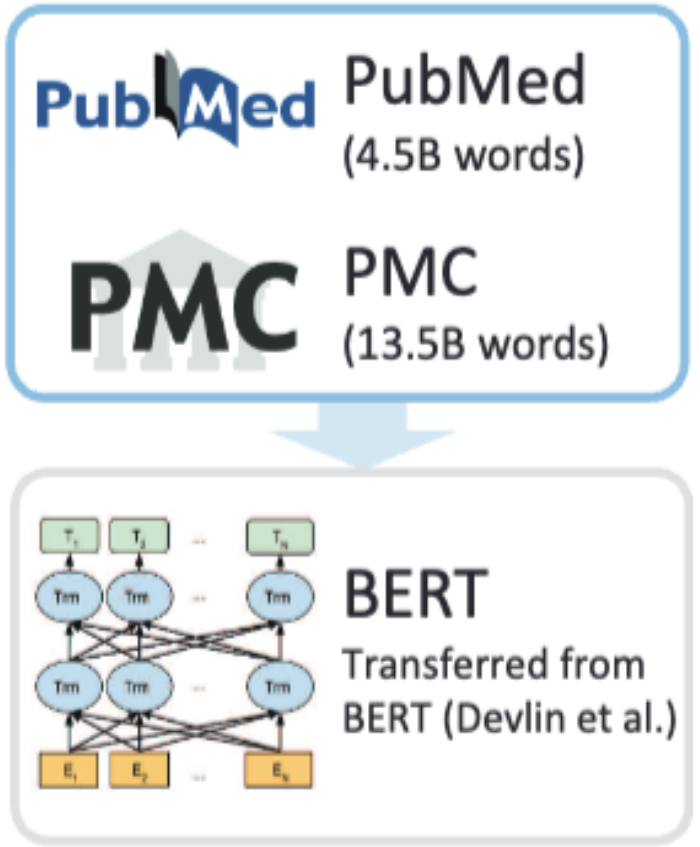


Text Classification

BERT Variants for Different Domains

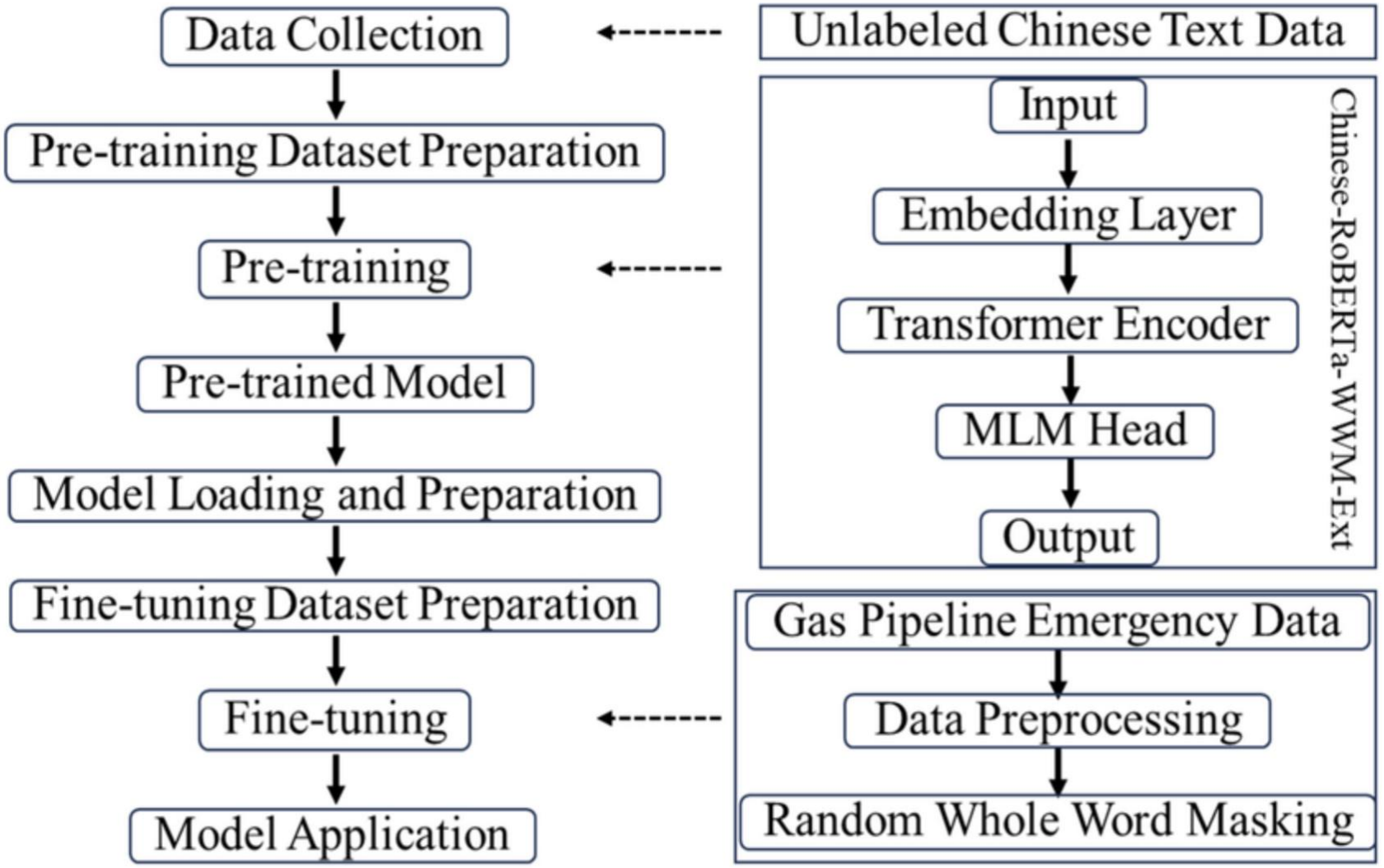


Three variants of BERT in the field of biomedicine



Flowchart

Chinese-RoBERTa-wwm-ext



Encoding schemes and Enhancement Strategy

Method	Pattern
Basic1	*[CLS] {h} [SEP] sentence [SEP] {t}*
Basic2	*[CLS] {h} [SEP] {t} [SEP] sentence*
QA	*[CLS] What is the relationship between {h} and {t}? [SEP] sentence*
Entity1[1]	*[CLS] s1 [E1] {h} [/E1] s2 [E2] {t} [/E2] s3 [SEP]*
Entity2[2]	*[CLS] s1 [Entity1] {h} [/Entity1] s2 [Entity2] {t} [/Entity2] s3 [SEP]*

Encoding schemes.
s1/2/3 denote sentence segments around entity spans.
1.Entity1 is Entity_marked1
2.Entity2 is Entity_marked2

- **We apply weak sample augmentation and weighted cross-entropy：**

$$\mathcal{L} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C}\omega_c\, y_c^{(n)} \log p_c^{(n)}$$

with class weights w emphasizing hard categories.

Results:Pretrained Model Comparison

- Scores of Chinese-RoBERTa-wwm-ext with Different Text Encoding Methods.

Model	BioBERT[1]	Chinese-RoBERTa[2]	MedBERT[3]
Macro-P	46.00%	**90.93%**	89.79%
Macro-R	36.30%	**90.20%**	88.70%
Macro-F1	30.80%	**90.24%**	88.80%
Micro-F1	46.28%	**90.46%**	89.25%

1.BioBERT : Entity_marked1
2.Chinese-RoBERTa : Entity_marked2
3.MedBERT : QA

Chinese-RoBERTa-wwm-ext achieved the highest score across all indicators.

Results:Encoding Strategy Impact

- Chinese-RoBERTa-wwm-ext performance under encoding variants.

	Macro-P	Macro-R	Macro-F1	Micro-F1
Basic1	88.94%	88.60%	86.71%	89.01%
Basic2	89.33%	88.00%	88.09%	88.46%
QA	89.23%	86.80%	87.06%	89.25%
Entity1[1]	88.34%	87.30%	87.60%	87.81%
Entity2[2]	**90.72%**	**90.20%**	**90.24%**	**90.46%**

1.Entity1 is Entity_marked1
2.Entity2 is Entity_marked2

The model achieved the highest overall score by using Entity_marked2 .

Results:Class Enhancement Effects

- Chinese-RoBERTa-wwm-ext respectively utilized weighted Loss and sample augmentation to improve the scores.

	Regular Loss	Weighted Loss	Variation
Macro-P	88.02%	89.62%	+1.60%
Macro-R	88.1%	88.60%	+1.70%
Macro-F1	87.01%	88.71%	+1.70%
Micro-F1	86.90%	89.01%	+2.11%
F1 Range	26.29%	18.01%	-8.28%

Weighted Loss.

	Weighted Loss	Weighted+Aug	Variation
Macro-P	89.62%	90.93%	+1.31%
Macro-R	88.6%	90.20%	+1.60%
Macro-F1	88.71%	90.24%	+1.53%
Micro-F1	89.01%	90.46%	+1.45%
F1 Range	18.01%	9.45%	-8.56%

Sample Augmentation

We further enhanced the model's performance on the weakly classified categories, and achieved significant improvements in overall accuracy and robustness.

Conclusion

- The experimental results show that the model based on Chinese-RoBERT a-wwm-ext ***performs exceptionally well*** in the relation classification task.
- Compared with other pre-trained models, our method has increased the ***Micro-F1 value by 44.18% and 1.21%*** respectively, verifying the significance of ***domain-specific optimization and entity marking methods*** in this task.
- We also effectively alleviated the problem of difficult categories through the ***category enhancement strategy***, further improving the model's performance in categories that are difficult to distinguish.
- This research provides a method for extracting relationships in ***Chinese medical texts***, and it has high practical application value.

References

- *Charangan Vasantharajan, Kyaw Zin Tun, Ho ThiNga, Sparsh Jain, Tong Rong, and Chng Eng Siong.Medbert: A pre-trained language model for biomedical named entity recognition. In 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pages 1482– 1488,2022.*
- *Zexuan Zhong and Danqi Chen. A frustratingly easy approach for entity and relation extraction. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 50– 61,Online, June 2021. Association for Computational Linguistics.*
- *Jue Wang and Wei Lu. Two are better than one: Joint entity and relation extraction with table sequence encoders. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1706– 1721, Online, November 2020. Association for Computational Linguistics.*
- *Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. Pre-training with whole word masking for chinese bert. ArXiv, abs/1906.08101, 2019.*
- *Jinhyuk Lee, WonJin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36:1234 – 1240, 2019.*

This project referred to a total of 34 references. Due to space constraints, we will not list them all in this presentation. For the full list, please refer to the report.

Acknowledgments

- We acknowledge the valuable contributions of the ***medical professionals*** who provided domain expertise and validation support throughout this research.

- We also thank the ***open-source community*** for developing and maintaining the tools and frameworks that made this work possible.

- At the same time, we would like to express our sincere gratitude to the ***CISC7021 teaching team*** for the guidance and computing resources they provided for this research.