Multimodal Named Entity Recognition and Relation Extraction with Retrieval-Augmented Strategy

Xuming Hu
Tsinghua University
hxm19@mails.tsinghua.edu.cn

ABSTRACT

Multimodal Named Entity Recognition (MNER) and Multimodal Relation Extraction (MRE) are tasks in information retrieval that aim to recognize entities and extract relations among them using information from multiple modalities, such as text and images. Although current methods have attempted a variety of modality fusion approaches to enhance the information in text, a large amount of readily available internet retrieval data has not been considered. Therefore, we attempt to retrieve multimodal content related to images, objects, and entire sentences from the internet and use this retrieved content as input for cross-modal fusion to improve the performance of entity and relation extraction tasks in the text.

CCS CONCEPTS

• **Computing methodologies** → **Information Extraction**.

KEYWORDS

Multimodal Named Entity Recognition, Multimodal Relation Extraction, Retrieval-Augmented Strategy

ACM Reference Format:
Xuming Hu. 2023. Multimodal Named Entity Recognition and Relation Extraction with Retrieval-Augmented Strategy. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3539618.3591790

1 MOTIVATION

Motivated by the limitations of existing efforts, which predominantly focus on modeling the visual and textual content of input, we aim to enhance the performance of multimodal named entity recognition and multimodal relation extraction tasks. While prior work, such as text-based methods [2, 3, 5], graph-based method [7], hierarchical visual prefix fusion network [1], and fine-grained multimodal alignment approach with Transformer [4], have demonstrated progress in this area, there is still room for improvement.

Inspired by retrieval-augmented multimodal relation extraction proposed by Wang et al. [6], we propose a novel approach that goes beyond simple text retrieval. Our method retrieves not only textual evidence, but also visual and textual evidence related to the object, sentence, and entire image. By employing a unique strategy to combine evidence from the object, sentence, and image levels, we facilitate improved reasoning across modalities, which we believe will lead to better performance in MNER and MRE tasks.

2 CHALLENGES AND FUTURE RESEARCH DIRECTIONS

The challenges in retrieval-augmented strategy for multimodal relation extraction and named entity recognition tasks include: (1) Retrieval efficiency: Efficiently extracting relevant information from internet poses a significant challenge. (2) Noisy retrieval content: Ensuring the retrieved content is of high quality and free from noise. (3) Inconsistency between retrieved content and original input: Ensuring that the retrieved content aligns well with the initial image and textual information is essential. (4) Fusion of retrieved content with original input: Developing an effective method for integrating retrieved content with the original textual and visual input to enhance reasoning capabilities.

To address these challenges, future research can explore the following directions: (1) Improved retrieval algorithms: Develop more efficient and accurate algorithms for retrieving relevant textual and visual evidence from large knowledge bases. (2) Noise reduction techniques: Investigate methods for identifying and reducing noise in the retrieved content to improve the quality of the data used for reasoning. (3) Consistency enforcement: Devise strategies to ensure that the retrieved content is consistent with the original input, thereby enhancing the relevance and usefulness of the retrieved data. (4) Advanced fusion techniques: Explore innovative approaches for effectively combining the retrieved content with the original textual and visual information to enable more accurate and robust multimodal reasoning.

REFERENCES

[1] Xiang Chen, Ningyu Zhang, Lei Li, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022. Good Visual Guidance Make A Better Extractor: Hierarchical Visual Prefix for Multimodal Entity and Relation Extraction. In *Findings of NAACL*. 1607–1618.
[2] Xuming Hu, Lijie Wen, Yusong Xu, Chenwei Zhang, and Philip Yu. 2020. SelfORE: Self-supervised Relational Feature Learning for Open Relation Extraction. In *Proc. of EMNLP*. Online, 3673–3682.
[3] Xuming Hu, Chenwei Zhang, Fukun Ma, Chenyao Liu, Lijie Wen, and S Yu Philip. 2021. Semi-supervised Relation Extraction via Incremental Meta Self-Training. In *Findings of EMNLP*. 487–496.
[4] Lei Li, Xiang Chen, Shuofei Qiao, Feiyu Xiong, Huajun Chen, and Ningyu Zhang. 2023. On Analyzing the Role of Image for Visual-enhanced Relation Extraction. In *In Proc. of AAAI (Student Abstract)*.
[5] Shuliang Liu, Xuming Hu, Chenwei Zhang, Shu'ang Li, Lijie Wen, and Philip S. Yu. 2022. HiURE: Hierarchical Exemplar Contrastive Learning for Unsupervised Relation Extraction. In *Proc. of NAACL-HLT*. 5970–5980.
[6] Xinyu Wang, Jiong Cai, Yong Jiang, Pengjun Xie, Kewei Tu, and Wei Lu. 2022. Named Entity and Relation Extraction with Multi-Modal Retrieval. In *Proc. of EMNLP*.
[7] Changmeng Zheng, Junhao Feng, Ze Fu, Yi Cai, Qing Li, and Tao Wang. 2021. Multimodal Relation Extraction with Efficient Graph Alignment. In *Proc. of ACM MM*, Heng Tao Shen, Yueting Zhuang, John R. Smith, Yang Yang, Pablo César, Florian Metze, and Balakrishnan Prabhakaran (Eds.). ACM, 5298–5306.

SIGIR '23July 23–27, 2023Taipei, Taiwan

ACM ISBN 978-1-4503-9408-6/23/07.
https://doi.org/10.1145/3539618.3591790