Wiki Workshop (10th edition) – May 11, 2023

Building Authoring Tools for Multimedia Content with Human-in-the-loop Relevance Annotations

Jheng-Hong Yang,[1] Carlos Lassance,[2] Rafael S. Rezende,[2]
Krishna Srinivasan,[4] Miriam Redi,[3] Stéphane Clinchant,[2] and Jimmy Lin[1]
[1] University of Waterloo [2] Naver Labs Europe [3] Wikimedia Foundation [4] Google Research

Abstract

We present a new dataset and a dedicated track designed to facilitate the development of Authoring Tools for Multimedia Content Creation (AToMiC). AToMiC is built on top of the existing Wikipedia-based Image Text dataset and includes three components: image–text associations, texts, and images. To bring research groups together to discuss their work on this new large test collection, we have collaborated with the National Institute of Standards and Technology (NIST) to host a dedicated information retrieval track at the TREC 2023 conference. We aim to invite participants from different backgrounds to join the discussion in order to consolidate generalized knowledge across a wide variety of techniques, much wider than only a few research groups could tackle. To foster collaboration, we have made the resources of AToMiC publicly available at `https://github.com/TREC-AToMiC/AToMiC`.

Keywords: Authoring Tools, Image–Text Retrieval, Cross-modal Retrieval, Multimedia Analysis

Introduction

Our goal is to address the challenge of assisting authors in creating multimedia content that enhances the appeal of web pages that are primarily textual in nature, such as descriptive articles or travel blog posts. To accomplish this, authors can add multimedia content such as images and videos to complement the text. We are developing tools to assist content creators in this task, which we refer to as Authoring Tools for Multimedia Content (AToMiC).

To encourage evaluation, share baselines, and foster a community around this challenge, we will organize the AToMiC Track at the 2023 Text Retrieval Conference (TREC). Given recent advances in vision–language pretrained models (Radford et al., 2021; Jia et al., 2021; Li et al., 2021; Li et al., 2022; Gu et al., 2018; Yao et al., 2022; Singh et al., 2022; Bleeker and Rijke, 2022), we believe this is an opportune time for such a track, which we think will attract interest not only from the information retrieval community but also from the natural language processing, computer vision, and multimedia communities.

The TREC 2023 AToMiC Track tasks are operationalized in the context of English articles and images (from 108 languages) in Wikipedia. We propose two evaluation tasks: the image suggestion task and the image promotion task. The objective of the image suggestion task is to suggest an image that can enhance a particular section of an article and make it more engaging. In contrast, the image promotion task entails identifying the most relevant information in a Wikipedia article to describe or explain an image. For instance, in the current Wikipedia article on the National Institute of Standards and Technology (NIST),[1], images in the "History" section complement the textual description. However, no image exists in the "World Trade Center collapse investigation" section. In the image suggestion task, the information need of an editor is to identify an appropriate image[2] that can be added to this section to complement the text suitably. While in the image promotion task, we seek to find context information about a specific image, e.g., the collapse of World Trade Center,[3] that could help us write useful captions that captures the essential elements about the event.

Our objective is not limited to a known-item retrieval task. Rather, we aim to develop Authoring Tools for Multimedia Information Content (AToMiC) that assist future editors in creating new and better content for Wikipedia. To provide training data, our track design uses the existing content–image pairings in Wikipedia, and we intend to obtain relevance annotations with the help of NIST. Our evaluation focuses on systems that can perform both tasks. For example, a relevant image may exist in the Wikimedia Commons database but has not yet been included in a specific article. Alternatively, there may be no existing image that perfectly matches the context of a section, but the system can propose images that are similar or related. As a result, these tasks require systems to comprehend both the textual content of the article and the

[1] `https://en.wikipedia.org/wiki/National_Institute_of_Standards_and_Technology`

[2] `https://en.wikipedia.org/wiki/NIST_World_Trade_Center_Disaster_Investigation#/media/File:Fire_test;_World_Trade_Center_(5887635739).jpg`

[3] `https://en.wikipedia.org/wiki/File:JohnsonKV_DSC_0104.jpg`

© Copyright held by the owner/author(s), published under Creative Commons CC BY 4.0 License

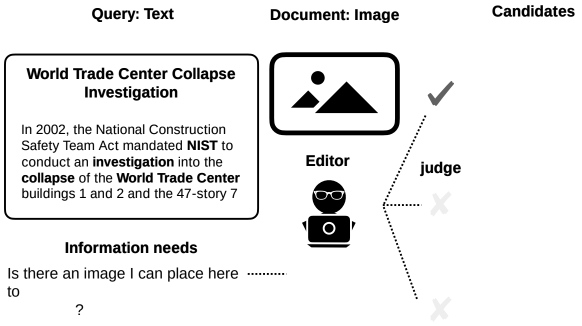
Query: Text
Document: Image
Candidates
World Trade Center Collapse Investigation
In 2002, the National Construction Safety Team Act mandated **NIST** to conduct an **investigation** into the **collapse** of the **World Trade Center** buildings 1 and 2 and the 47-story 7
Editor
judge
Information needs
Is there an image I can place here to
?

Figure 1: A section of the Wikipedia article about the National Institute of Standards and Technology (NIST) titled “World Trade Center Collapse Investigation” with no associated image. An editor’s information need is to find a suitable image to make this section more engaging.

visual content of the images, as well as their connections.

The AToMiC Track at TREC 2023 aims to foster research and development in multimedia content creation by providing a platform to evaluate and compare systems that address image suggestion and promotion tasks. We aspire to promote the development of new authoring tools that can assist content creators in enhancing the visual appeal of their web pages. Additionally, we anticipate that the AToMiC Track will encourage collaboration and knowledge-sharing among researchers in information retrieval, natural language processing, computer vision, and multimedia communities. Ultimately, our objective is to advance the state of the art in multimedia content creation and make it easier and more accessible to produce engaging and informative web pages.

Dataset Overview

AToMiC Collections. To construct our collections, we further filter the WIT (Srinivasan et al., 2021) data tuples and separate them into two disjoint sets. First, we consider a subset that only contains the English domain in Wikipedia. In addition, we separate them by grouping the article-specific attributes (e.g., titles and descriptions) as a text “document” C 2) and other image-specific attributes (e.g., captions) as an image “document” < 2 ". After removing invalid image URLs and duplicates (based on string matching), we arrive at a text collection j)j 10M and an image collection j "j 10M.

Sparse relevance labels. Sparse relevance labels, or qrels in TREC terminology, can be extracted from existing section-image associations in Wikipedia using WIT. We have extracted all the available pairs from WIT to create our qrels. However, it is important to note that these qrels only contain pseudo-positive judgments based on known associations. We have further divided these qrels into training, validation, and test sets that are aligned with the WIT splits.[4] Since the qrels are sparse and binary, metrics such as mean reciprocal rank (MRR) and recall would be appropriate.

References

[Bleeker and Rijke2022] Maurits Bleeker and Maarten de Rijke. 2022. Do lessons from metric learning generalize to image-caption retrieval? In *Proc. of ECIR*, pages 535–551.

[Gu et al.2018] Jiuxiang Gu, Jianfei Cai, Shafiq R Joty, Li Niu, and Gang Wang. 2018. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *Proc. of IEEE/CVF CVPR*, pages 7181–7189.

[Jia et al.2021] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proc. of ICML*, pages 4904–4916.

[Li et al.2021] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Proc. of NeurIPS*, 34:9694–9705.

[Li et al.2022] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proc. of ICML*.

[Radford et al.2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proc. of ICML*, pages 8748–8763.

[Singh et al.2022] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. Flava: A foundational language and vision alignment model. In *Proc. of IEEE/CVF CVPR*, pages 15638–15650.

[Srinivasan et al.2021] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proc. of SIGIR*, page 2443–2449.

[Yao et al.2022] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2022. FILIP: Fine-grained interactive language-image pre-training. In *Proc. of ICLR*.

[4] `https://github.com/google-research-datasets/wit/blob/7b15d12d374d660ae3101f973f45f0909f174661/DATA.md`

© Copyright held by the owner/author(s), published under Creative Commons CC BY 4.0 License