Introducing the TREC 2023 AToMiC Track

Draft of November 18, 2022 for Discussion at TREC Workshop

Jheng-Hong Yang,[1, 2] 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

1 INTRODUCTION

Our work tackles the challenge of providing assistance to authors in multimedia content creation. To enhance the appeal of web pages that are primarily textual in nature (e.g., a descriptive article about a topic, a travel blog post, etc.), authors can add appropriate multimedia content to complement the text (e.g., images, video, etc.) [1, 10, 14]. We aim to build tools that help content creators in this task, which we dub **A**uthoring **To**ols for **M**ulti**m**edia **C**ontent (AToMiC). To provide evaluation resources, 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 [2–6, 9, 12, 15], we believe this is a particularly propitious time for such a track, which we believe can draw interest from not only the information retrieval community, but also the natural language processing, computer vision, and multimedia communities.

The tasks in the TREC 2023 AToMiC Track are operationalized in the context of Wikipedia, presently focused on images and English articles. Concretely, we propose two tasks: In the *image suggestion task*, given an existing section in an article that does not currently contain an image, an editor's information need is as follows: What image can be added *to this section* to yield a more engaging article? The inverse of the image suggestion task is the *image promotion task*: Given an image, what sections of Wikipedia articles can it be attached to yield a more engaging article?

As a working example, consider the current article on Wikipedia about the National Institute of Standards and Technology (NIST).[1] The "History" section contains images illustrating various activities the institute has engaged in, which feels appropriate in complementing the textual description. Now, consider the section titled "World Trade Center collapse investigation" shown in Fig. 1: For the image suggestion task, an editor asks, "I wonder what images I might insert here that would complement the text appropriately?" This user need represents a *topic* in TREC parlance, i.e., the information need expressed implicitly when composing an article with texts and images. A relevant image given the context might be this image[2] shown as the top candidate in Fig. 1. The image promotion task can be viewed as the inverse. Given an image, a Wikipedia editor asks: Where can I attach this image in order to make a particular article more compelling? The goal of the AToMiC Track is to evaluate systems that can perform both tasks.

It is important to note that for the image suggestion task, a relevant image may or may not exist. Similarly, for the image promotion task, a relevant image may or may not exist. Thus, both are *not* (merely) a known item retrieval task. While existing content–image pairings in Wikipedia provide training data in our track design, it is our intention to develop authoring tools for multimedia information content (AToMiC) that *help* future editors forge *new* and better contents in Wikipedia.

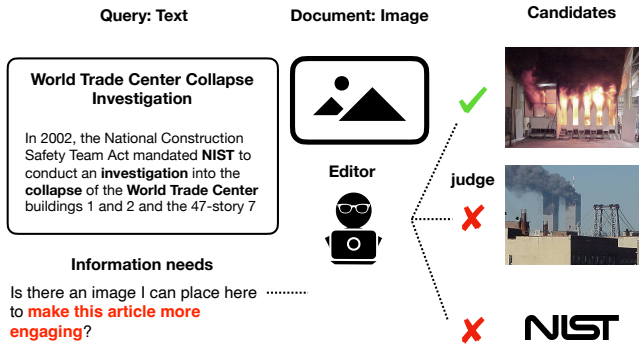

Figure 1: A section of the Wikipedia article about the National Institute of Standards and Technology (NIST) with the heading "World Trade Center Collapse Investigation". This section currently does not have an associated image, but an editor wonders: "Is there an image I can place here to make this article more engaging?" This represents the image suggestion task in the TREC AToMiC Track. The goal of the system is to suggest potentially relevant images.

2 PROBLEM DEFINITION

In standard *ad hoc* retrieval, we assume the existence of a corpus C comprised of a collection of documents $\{d_1, d_2 \ldots d_n\}$. In response to a user's information need represented as query q, the system's goal is to return a top-k ranked list of documents that maximizes some metric of quality such as nDCG, MRR, or MAP.

AToMiC has a different setup. For TREC 2023, we have prepared two different collections: a collection of texts $C_T = \{t_1, t_2 \ldots t_n\}$ (t stands for text) and a collection of images $C_M = \{m_1, m_2 \ldots m_n\}$ (m stands for media). An example t from C_T is shown in Figure 2. An example m from C_M is shown in Figure 3. Note, critically, that each m is comprised of the image itself (i.e., pixel values) *as well as* metadata (if available), which might include a caption describing the image. More details about the dataset construction process can be found in Section 3.

Image Suggestion Task. An information need (for convenience, a *query*), denoted q, is simply a text t drawn from C_T, i.e., $q \in C_T$. That is, an editor of Wikipedia examines a *a specific section* of an article and wishes to locate an appropriate image. Given q, 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

```
{
"id": "wit-train-topic-5390014",
"page_title: "Half Dome",
"section_title": "Ascents",
"page_description": "Half Dome is a granite dome at the eastern end of
Yosemite Valley in Yosemite National Park, California. It is a well-known
rock formation ...",
"section_description": "As late as the 1870s, Half Dome was described as
'perfectly inaccessible' by Josiah Whitney of the  California Geological
Survey. The summit was ...."
}
```


Figure 2: A text t from C_T, which contains text from the section "Ascents" in the Wikipedia article on "Half Dome".


```
{
"id": 296128,
"image_url": "5/5d/Half_Dome--cables.jpeg",
"file_name: "train_1214099.png",
"caption_reference_description": "Hikers use cables to ascend Half Dome,
"caption_alt_text_description": "",
"caption_attribution_description": "Hikers walk up the east face of
Half Dome, aided by a pair of cables. Olga Joplin in foreground.
This is an image of a place or building that is listed on the National
Register of Historic Places in the United States of America.
Its reference number is 12000494"
}
```


Figure 3: An image m from C_M. Given the text t in Figure 2 as a query in the image suggestion task, this would be a relevant image. (In fact, this image is already associated with that particular section in the Wikipedia article.)

system's task is to return a top-k ranked list of images, drawn from C_M that maximizes some metric of quality. Relevance in this context is operationalized as "this image would be appropriate to attach to this section". So, for the query represented by the text in Figure 2, the image in Figure 3 would be relevant. Note that in this case, the image is already associated with the section "Ascents" in the Wikipedia article on "Half Dome".

Image Promotion Task. This task is the inverse of the image suggestion task. The query q is an image drawn from C_M and the collection to be searched is C_T. A system is expected to return a top-k ranked list of texts, where relevance is operationalized as "this is a section of a Wikipedia article that would be an appropriate attachment point for this image". Although this is not the focus of our efforts, the image promotion task can also provide an authoring tool for crafting captions for images. In this specific case, the users can write a caption: "Getting ready for Half Dome hike" to make it relevant to the text section shown in Figure 2, or "A mountaineer

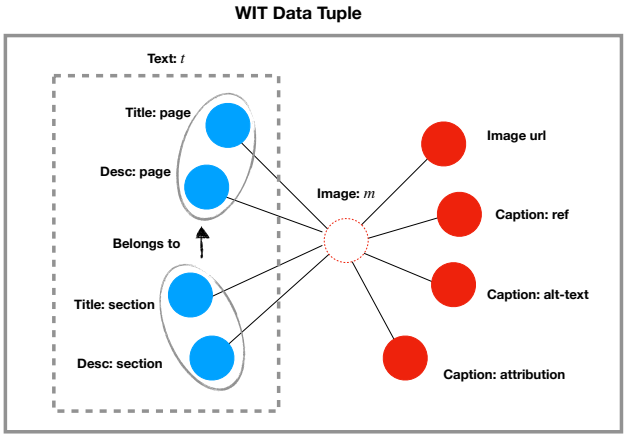

Figure 4: A conceptual graph of the WIT tuple. For simplicity, we omit other contextual metadata such as language identifiers, page URLs, image size, etc.

proceeds across the ridge" to make the image relevant to the mountaineering article in Wikipedia.[3] In sum, the image promotion task requires retrieval systems to return relevant contexts for users to conduct downstream tasks, such as descriptive image captioning [7] and open-knowledge visual question answering [8, 11], that could benefit from them.

3 DATASET OVERVIEW

The AToMiC (v0.1) dataset is built upon the Wikipedia-based Image Text (WIT) dataset [13],[4] where the authors provide 37.6M tuples of (image data, text data, contextual data) extracted from Wikipedia pages across 279 languages. A Wikipedia page (article) contains several sections to illustrate an entity or concept in the real world. The WIT dataset contains curated image–text pairs if and only if an image is attached to a section of a page.

Each WIT data tuple is organized around an image and can be represented as a conceptual graph shown in Figure 4. Generally, there are two types of attributes in the conceptual graph: article-specific and image-specific attributes. Article-specific attributes contain the titles and descriptions of a Wikipedia page associated with the image. Image-specific attributes contain image metadata that describes the image content. As a specific example, the section "Ascents" of the entity "Half Dome" is shown in Figure 5; note that there is an image associated with this specific section already. We leverage this structure to separate texts and images into two collections (C_T and C_M) and create sparse relevance labels.

AToMiC Collections. To construct our collections, we further filter the WIT data tuples and separate them into two disjoint sets. First, we consider a subset that only contains the English domain in Wikipedia, resulting in 5.5M tuples. In addition, we separate them by grouping the article-specific attributes (e.g., titles and descriptions) as a text "document" t and other image-specific attributes (e.g.,

[3] https://en.wikipedia.org/wiki/Mountaineering
[4] https://github.com/google-research-datasets/wit

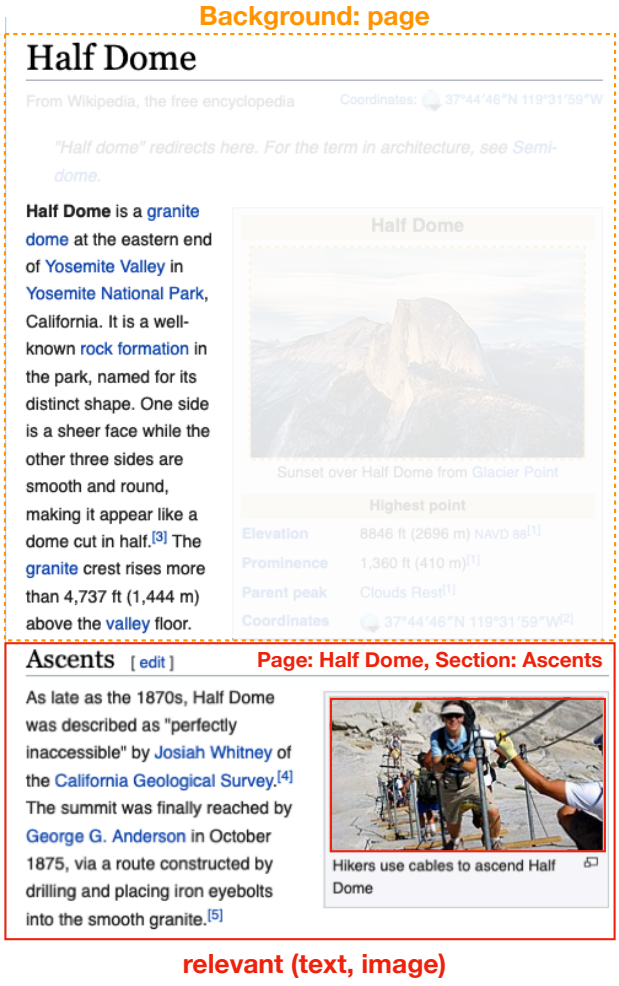

Figure 5: An example of a sparse relevance judgment extracted from WIT extracted from an existing Wikipedia page.

captions) as an image "document" m. After removing invalid image URLs and duplicates (based on string matching), we arrive at a text collection $|C_T| \approx 5.1$M and an image collection $|C_M| \approx 3.8$M.

Sparse relevance labels. From WIT we can extract sparse relevance labels, or qrels in TREC parlance, from existing section–image associations in Wikipedia. As already shown above, again consider a snippet from the Wikipedia page in Figure 5 on "Half Dome". In the section "Ascents", there is already an image attached. This provides a (t, m) relevant pair, where $t \in C_T$ and $m \in C_M$. We have processed WIT to extract all these available pairs to comprise our qrels. To be clear, these qrels at present only contain positive judgments based on these section–image associations. We further divide these qrels into training, validation, and test sets that are aligned with the WIT splits.[5] For clarity, we refer to this as the

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

Split	# texts	# images	# qrels
Training	5,030,748	3,723,512	5,030,748
Validation	38,859	30,365	38,859
Test	30,938	20,732	30,938
Total	5,100,545	3,774,609	5,100,545

Table 1: AToMiC (v0.1) dataset statistics.

AToMiC v0.1 dataset. Dataset statistics are shown in Table 1. Note that the rows—training, validation, and test—capture the number of "queries" in each split. When performing the two proposed tasks (i.e., image suggestion and image promotion), the collection to be indexed and searched should refer to the total number of texts $|C_T| \approx 5.1$M or images $|C_M| \approx 3.8$M.

At present, we remove a qrel if the URL of an image is not available on the Internet. Thus, the number of qrels is less than the 5.5M tuples in the English subset of WIT. The number of images is smaller than the number of qrels because an image can be attached to multiple texts. We believe that the current dataset is sufficient as a starting point for evaluating both the image suggestion task and the image promotion task, although there is one major limitation that we can identify (see Section 5). Since the qrels are sparse, a metric like MRR would be appropriate.

4 TREC EVALUATION

We envision the setup of the TREC evaluation to parallel the relationship between the MS MARCO passage collection and the TREC Deep Learning Tracks. That is, the MS MARCO passage collection provides large-scale training data in the form of sparse judgments, while the TREC Deep Learning Tracks allows researchers to validate their models with a more established evaluation methodology.

In the same way, we envision that researchers will use the AToMiC collection (as described above) for model development, and the TREC 2023 AToMiC Track will provide rich human relevance judgments. Questions of how we sample the topics for both tasks are to be determined.

5 LIMITATIONS

Currently, one major limitation of AToMiC v0.1 is that the collections contain a highly biased sample of the universe of possible texts and images. That is, C_T and C_M contain only text and images that are present in a relevant section–image tuple. For example, consider the page on "Half Dome": sections from that page that do not already have images attached *are not* in C_T. To tackle this issue we are building v0.2 collections that rectify this issue by projecting the v0.1 qrels into a more recent Wikipedia dump. One advantage of doing such a projection is that it adds the possibility of identifying new images that have been added to documents, which would be a natural way of creating the dev set (images not in sections as of v0.1, but that we know are relevant for those sections).

REFERENCES

[1] Sarah Bibi, Dian Shinta Sari, and Muhammad Iqbal Ripo Putra. 2020. The design of multimedia storytelling. In *ELT Forum: Journal of English Language Teaching*, Vol. 9. 16–26.

[2] Maurits Bleeker and Maarten de Rijke. 2022. Do lessons from metric learning generalize to image-caption retrieval?. In *European Conference on Information Retrieval*. Springer, 535–551.

[3] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7181–7189.

[4] 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 *International Conference on Machine Learning*. PMLR, 4904–4916.

[5] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *ICML*.

[6] 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. *Advances in neural information processing systems* 34 (2021), 9694–9705.

[7] Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. 2021. Visual News: Benchmark and Challenges in News Image Captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 6761–6771.

[8] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*. 3195–3204.

[9] 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 *International Conference on Machine Learning*. PMLR, 8748–8763.

[10] Daniele Rama, Tiziano Piccardi, Miriam Redi, and Rossano Schifanella. 2022. A large scale study of reader interactions with images on Wikipedia. *EPJ Data Science* 11, 1 (2022), 1.

[11] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-OKVQA: A Benchmark for Visual Question Answering using World Knowledge. *arXiv preprint arXiv:2206.01718* (2022).

[12] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15638–15650.

[13] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. WIT: Wikipedia-Based Image Text Dataset for Multimodal Multilingual Machine Learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. 2443–2449.

[14] Qian Xu and S Shyam Sundar. 2014. Lights, camera, music, interaction! Interactive persuasion in e-commerce. *Communication Research* 41, 2 (2014), 282–308.

[15] 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 *International Conference on Learning Representations*. https://openreview.net/forum?id=cpDhcsEDC2