

AMELI: Enhancing Multimodal Entity Linking with Fine-Grained Attributes

Barry Menglong Yao[♣] Sijia Wang[♣] Yu Chen[♡] Qifan Wang[♡] Minqian Liu[♣]
Zhiyang Xu[♣] Licheng Yu[♡] Lifu Huang[♣]
[♣]Virginia Tech [♡]Meta AI

{barryyao, sjiawang, minqianliu, zhiyangx}@vt.edu

{hugochen, wqfcr, lichengyu}@meta.com

Abstract

We propose attribute-aware multimodal entity linking, where the input consists of a mention described with a text paragraph and images, and the goal is to predict the corresponding target entity from a multimodal knowledge base (KB) where each entity is also accompanied by a text description, visual images, and a collection of attributes that present the meta-information of the entity in a structured format. To facilitate this research endeavor, we construct AMELI, encompassing a new multimodal entity linking benchmark dataset that contains 16,735 mentions described in text and associated with 30,472 images, and a multimodal knowledge base that covers 34,690 entities along with 177,873 entity images and 798,216 attributes. To establish baseline performance on AMELI, we experiment with several state-of-the-art architectures for multimodal entity linking and further propose a new approach that incorporates attributes of entities into disambiguation. Experimental results and extensive qualitative analysis demonstrate that extracting and understanding the attributes of mentions from their text descriptions and visual images play a vital role in multimodal entity linking. To the best of our knowledge, we are the first to integrate attributes in the multimodal entity linking task¹.

1 Introduction

Entity linking aims to disambiguate and link entity mentions within a text to their corresponding entities in knowledge bases. While earlier research (Onoe and Durrett, 2020; Zhang et al., 2021b; Tan and Bansal, 2019; Tang et al., 2021; Yang et al., 2019; Ganea and Hofmann, 2017a; Ravi et al., 2021; Ayoola et al., 2022a,b) predominantly focus on linking entities based on text, recent studies have started to extend it to multi-modality where

both mentions and entities in knowledge bases are described with text and visual images (Zhang et al., 2021a; Zhou et al., 2021; Li and Wang, 2021; Zheng et al., 2022; Dost et al., 2020; Wang et al., 2022b; Adjali et al., 2020; Wang et al., 2023). However, all these studies view each entity in the knowledge base as an atomic symbol while ignoring the meta-information, such as various attributes of each entity, which, we argue, is especially important in disambiguating entities in a multimodal context.

In this work, we focus on multimodal entity linking (MEL) which requires understanding fine-grained attributes of mentions from both text and images and linking them to the corresponding entities in a target multimodal knowledge base where each entity is also illustrated with text, images, and a set of attributes. Figure 1 shows an example where each entity, such as *ASUS ROG Laptop - White* in the target knowledge base is described with a set of attributes, such as *Screen Size*, *System Memory*, *Graphics*, and to disambiguate and link a particular mention, e.g., *ASUS laptop* to the target entity, we need to carefully detect the attributes of the mention from its text and image descriptions and compare it against each entity. Such attribute-aware multimodal entity linking is critical to E-commerce domains, e.g., analyzing user opinions from social media posts about particular products. Yet, it is relatively less studied in the entity linking literature.

To support research toward attribute-aware multimodal entity linking, we introduce AMELI, which consists of (1) a multimodal knowledge base that includes 34,690 product entities collected from the Best Buy² website and each entity is described with a product name, a product description, a set of attribute categories and values, e.g., “Color: Black”, and several images; and (2) a multimodal entity linking benchmark dataset that contains 16,735

¹The programs, model checkpoints, and the dataset are publicly available at <https://github.com/VT-NLP/Ameli>.

²<https://www.bestbuy.com/>





Review	Products			
 <p>The screen size is 14" which I think is the perfect size for a laptop. This ASUS laptop has a formidable performance with AMD Ryzen 9 CPU, NVIDIA Geforce 2060 Max Q, Ram 16GB, 1tb SSD, 1080p+120Hz display.</p> <p>helpful unhelpful</p>	<p>Category: PC Gaming -> Gaming Laptops Name: ASUS ROG Laptop - 16 GB</p> 	<p>Category: PC Gaming -> Gaming Laptops Name: ASUS ROG Laptop - Eclipse Grey</p> 	<p>Category: PC Gaming -> Gaming Laptops Name: ASUS ROG Laptop - White</p> 	
	<p>Description: Game like a pro on Windows 11 with this ROG Zephyrus G14. ... Enjoy a fast 120Hz refresh rate, 16GB of DDR4 RAM ...</p>	<p>Description: The AMD Ryzen 9 processor and 16GB of memory ... This 14-inch IPS Level Full HD ASUS notebook PC has a 1000 GB SSD ...</p>	<p>Description: ASUS ROG Zephyrus Gaming Laptop. The AMD Ryzen 9 processor and 16GB of RAM let you run graphics-heavy games smoothly. ... This ASUS notebook PC has 1000 GB SSD.</p>	
	<p>Attribute: Graphics : NVIDIA GeForce RTX 2060 Solid State Drive Capacity : 512 gigabytes System Memory (RAM) : 16 gigabytes Screen Size : 15.6 inches</p>	<p>Attribute: Graphics : NVIDIA GeForce RTX 2060 Solid State Drive Capacity : 1000 gigabytes System Memory (RAM) : 16 gigabytes Screen Size : 14 inches</p>	<p>Attribute: Graphics : NVIDIA GeForce RTX 2060 Solid State Drive Capacity : 1000 gigabytes System Memory (RAM) : 16 gigabytes Screen Size : 14 inches</p>	

Figure 1: An example for our attribute-aware multimodal entity linking. Left: review text and image; Right: product title, image, description, and attributes. To link the mention *ASUS laptop* to the target entity, we need to be aware of the attributes, e.g., memory and SSD capacity, and image features, e.g., color.

data instances while each instance contains a text description for a particular entity mention and several images. The goal is to interpret the multimodal context and attributes of each mention and map it to a particular entity in the multimodal knowledge base. AMELI is challenging as many entities in the knowledge base are about similar products with subtle differences in a few attributes, and thus, the model needs to correctly detect all the attributes from the multimodal context of each mention in order to link it to the target entity.

We conduct baseline experiments with several entity linking methods and propose a new framework consisting of a Natural Language Inference (NLI) based text disambiguation model to compare the mention description and attributes of candidate entities from the knowledge base and an image disambiguation model based on contrastive learning. Though our proposed approach significantly outperforms all the strong baselines, the experimental results still show a large gap between machine (51.54% F-score) and human performance (74.0% F-score). The contributions of this work can be summarized as follows:

- To the best of our knowledge, AMELI is the first benchmark dataset to support attribute-aware multimodal entity linking, and we are the first to integrate attribute features to improve the multimodal entity linking task.
- We propose a new disambiguation approach that considers the multimodal context of mentions as well as attributes of candidate entities, which significantly outperforms all the previous strong baselines on AMELI. Ablation studies further demonstrate the benefit and necessity of incorporating attribute information for multimodal entity

linking.

2 Related Work

Previous research on textual entity linking has established various benchmark datasets (Guo and Barbosa, 2018; Logeswaran et al., 2019; Hoffart et al., 2011; Cucerzan, 2007; Milne and Witten, 2008) and state-of-the-art neural models (Wu et al., 2019; Logeswaran et al., 2019; Ayoola et al., 2022c; Peters et al., 2019; Ganea and Hofmann, 2017b; Kolitsas et al., 2018; Cao et al., 2021; Lai et al., 2022; Cao et al., 2020; De Cao et al., 2022). However, these approaches cannot be directly adapted to multimodal entity linking due to the fundamental differences in input modalities and challenges.

Multimodal entity linking has recently been explored in various contexts such as social networks (Zhang et al., 2021a; Moon et al., 2018; Zhou et al., 2021; Li and Wang, 2021), domain-specific videos (Venkatasubramanian et al., 2017) and general news domains (Wang et al., 2022b). These studies focus on reducing noise in the abundant visual input of social networks (Zhang et al., 2021a; Li and Wang, 2021), learning distinguishable entity representations by contrastive learning (Wang et al., 2022b; Moon et al., 2018; Gan et al., 2021), or directly generating target entity names (Wang et al., 2023; Shi et al., 2023). Compared to these studies, our research considers the unique attributes along with visual and textual inputs. Table 1 compares AMELI with other existing entity linking datasets.

Many studies have been proposed to extract attribute values of products from their titles and descriptions by formalizing it as a sequence tagging task (Yan et al., 2021; Guo et al., 2018; Xu et al., 2019) or a question-answer problem (Yang et al., 2022; Wang et al., 2020). Several recent stud-

Dataset \ Feature	Attribute	Mention Images	Mention Text	Entity Images	Entity Text
Zhou et al. (2021)	X	✓	✓	✓	✓
Wikidiverse (Wang et al., 2022b)	X	✓	✓	✓	✓
WIKIPerson (Sun et al., 2022a)	X	✓	X	✓	✓
OVEN-Wiki (Hu et al., 2023)	X	✓	X	✓	✓
ZEMELD (Zheng et al., 2022)	X	✓	✓	✓	✓
MEL_Tweets (Adjali et al., 2020)	X	✓	✓	✓	✓
M3EL (Gan et al., 2021)	X	✓	✓	✓	✓
Weibo (Zhang et al., 2021a)	X	✓	✓	✓	✓
SnapCaptionsKB (Moon et al., 2018)	X	✓	✓	✓	✓
VTKEL (Dost et al., 2020)	X	✓	✓	X	✓
Guo and Barbosa (2018)	X	X	✓	X	✓
Zeshel (Logeswaran et al., 2019)	X	X	✓	X	✓
AMELI (Ours)	✓	✓	✓	✓	✓

Table 1: Comparison between AMELI and other related datasets.

ies (Lin et al., 2021; Zhu et al., 2020; Wang et al., 2022a) incorporate visual clues, such as product images or visual objects, into textual descriptions and extract attribute values based on their fused representations. In this study, we explore the potential of leveraging attribute values extracted from noisy user reviews to improve multimodal entity linking and achieve this by implicitly inferring attribute values through Natural Language Inference (NLI).

3 Dataset Construction

Data Source Our goal is to build (1) a multimodal knowledge base where each entity is described with text, images, and attributes, and (2) an entity linking benchmark dataset where each mention in a given context is also associated with text and several images and can be linked to a specific entity in the multimodal knowledge base. To construct these two benchmark resources, we use Best Buy³, a popular retailer website for electronics such as computers, cell phones, appliances, toys, etc., given that it consists of both multimodal product descriptions organized in a standard format and user reviews in both text and/or images which can be further used to build the entity linking dataset. As shown in Figure 1, each product in Best Buy is described with a *product name*, a list of *product categories*, a *product description*, a set of *attribute categories and values* as well as several *images*⁴. Additionally, users can post reviews in text and/or images under each product, while each review can be rated as helpful or unhelpful by other users. We develop scripts based on Requests⁵ to collect all

the above information. Each product webpage also requires a button click to display the attributes, so we further utilize Selenium⁶ to mimic the button click and collect all the attributes and values for each product. In this way, we collect 38,329 product entities and 6,500,078 corresponding reviews.

Data Preprocessing Many reviews are not suitable for the multimodal entity linking task for various reasons. Considering this, we designed several rules to preprocess the collected reviews: (1) Remove reviews and products without images; (2) Remove reviews with more than 500 tokens, since most of the state-of-the-art pretrained language models can only deal with 512 tokens; (3) Remove a review if it is only labeled as “unhelpful” by other users since we observe that these reviews usually do not provide much meaningful information; (4) Validate the links between reviews and their corresponding products and remove the invalid links. There are invalid links because Best Buy links each review to all variants of the target product. For example, for the review of *ASUS laptop* shown in Figure 1, the target product *ASUS ROG laptop - White* has several other variants in terms of color, memory size, processor model, etc., while Best Buy links the review to all variants of the target product. Since we take each product variant as an entity in our multimodal knowledge base, we detect valid links between reviews and product variants based on a field named *productDetails*, which reveals the gold target product variant information of the review in Best Buy’s search response. After obtaining the valid link for each review, we remove invalid links between this review and all other products. (5) Remove truncated images uploaded by users as these images cause “truncated image error”

³<https://www.bestbuy.com/>

⁴For simplicity, we show one image for each review or product in the figure, but there could be multiple associated images for both of them.

⁵<https://requests.readthedocs.io/en/latest/>

⁶<https://www.selenium.dev/>

during loading with standard image tools such as Pillow⁷. (6) Remove reviews containing profanity words based on the block word list provided by *Free Web Header*⁸. (7) Review images can also contain irrelevant objects or information; for example, a review image for a fridge can also contain much information on the kitchen. We apply the object detection model (Liu et al., 2023b) to detect the corresponding object using the entity name as prompt and save the detected image patch as the cleaned review image. We remove an image if the entity object can not be detected from it. Both original images and cleaned images are included in our dataset. (8) We also notice that many reviews do not contain enough context information from the text and images to link the product mention to the target product entity correctly. For example, in Figure 2, the target product is a *Canon camera*. However, the review image does not show the camera itself, and the review text does not contain any specific information about the camera. To ensure the quality of the entity linking dataset, we further design a validation approach (explained in Appendix A) to filter out reviews that do not contain enough context information.

Mention Detection We identify entity mentions from the reviews based on their corresponding products to construct the entity linking benchmark dataset. To achieve this, we design a pipelined approach to detect the most plausible product mention from each review. Given a review and its corresponding product, we first extract all product name candidates from the product title and category by obtaining their root word and identifying a fraction of the root word to be product name candidates with spacy⁹. For each n -gram span ($n \in \{1, 2, 3, 4, 5, 6\}$) in the review text, if it or its root form based on lemmatization matches with any of the product name candidates, we will take it as a candidate mention. Each review text may contain multiple mentions of the target product. Therefore, we compute the similarity between each candidate mention and the title of the target product based on SBERT (Reimers and Gurevych, 2019) and choose the one with the highest similarity to be the product mention. This approach achieves

⁷<https://pillow.readthedocs.io/en/stable/installation.html>

⁸<https://www.freewebsiteheaders.com/bad-words-list-and-page-moderation-words-list-for-facebook>

⁹<https://spacy.io/usage/linguistic-features#noun-chunks>

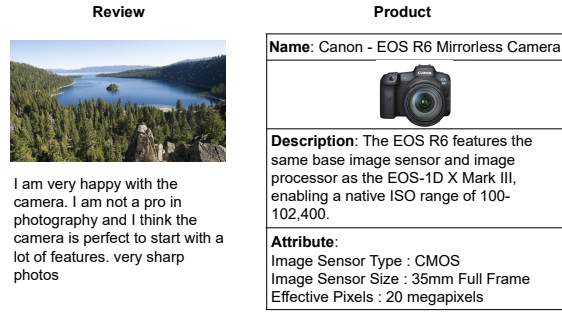


Figure 2: Example of Uninformative Reviews.

Data	Train	Dev	Test
# Reviews	12,148	1,846	2,741
# Review Images	21,780	3,369	5,323
Avg. # of Image / Review	1.79	1.83	1.94
Avg. # of Attributes / Review	1.22	1.62	3.54
# Products in KB		34,690	
# Product Images		177,873	
# Product Categories		986	
Avg. # of Image / Product		5.13	
Avg. # of Attributes / Product		23.01	

Table 2: Dataset statistics of AMELI.

an accuracy of 91.9% based on manual assessment on 200 reviews. Thus, we further apply it to detect product mentions for all reviews and remove the reviews that do not contain product mentions. We then ask one annotator to manually verify and correct all detected mentions in the Test set.

Train / Dev / Test Splits After all the pre-processing and filtering steps, we obtain 34,690 entities for the multimodal knowledge base and 17,431 reviews for the entity linking benchmark dataset. We name it AMELI and split the reviews into training (Train), development (Dev), and test (Test) sets based on the percentages of 70%, 10%, and 20%, respectively.

Note that since we utilize automatic strategies to detect mentions from reviews and filter out uninformative reviews, there is still noise remaining in the AMELI though the percentage is low. Thus, we ask humans to verify the Test set of AMELI. However, it is not trivial for humans to compare each mention with thousands of entities in the target knowledge base. To facilitate entity disambiguation by humans, for each review, we design two strategies to automatically retrieve strong negative candidate entities from the knowledge base: (1) as we know the target product of each review, we first retrieve the top- N ¹⁰ most similar entities to the target project from the KB as negative candidates.

¹⁰We set $N = 10$ as we observe that the top-10 retrieved candidates have covered the most confusing negative entities.

Here, the similarity between two products is computed based on the cosine similarity scores of their title representations produced by SBERT (Reimers and Gurevych, 2019); (2) Similarly, we also retrieve the top- N similar entities to the target product based on the cosine similarity scores of their image representations learned by CLIP (Radford et al., 2021). We combine these $2N$ negative candidates together with the target product entity as the set of candidate entities for each review and ask 12 annotators to choose the most likely target entity. Most annotators reach an accuracy of around 80%, while the overall accuracy is 79.73%, as shown in Table 5 in Appendix B. We remove the review if any of the annotators cannot select the target entity correctly. In this way, we obtain 2,741 reviews as the Test set. For each review in the Test set, we further ask one annotator to label all the attributes (Gold Attributes) of each mention. Table 2 shows the detailed statistics for each split of AMELI. Table 6 in Appendix C shows the category distribution of products in the multimodal knowledge base of AMELI.

4 Approach

4.1 Problem Formulation

We formulate the task as follows: given a user review r consisting of a text t_r , several images $\bar{\mathbf{V}}_r = \{v_r^0, \dots, v_r^q\}$, and an entity mention m_r , e.g., “coffee maker”, we aim to link the mention to a unique entity in the target knowledge base (KB). Each entity e_j in the KB is described with a text description d_{e_j} , a title \hat{t}_{e_j} , several images $\bar{\mathbf{V}}_{e_j} = \{v_{e_j}^0, \dots, v_{e_j}^h\}$, and a set of attributes $\bar{\mathbf{A}}_{e_j} = \{a_{e_j}^0, \dots, a_{e_j}^s\}$. Note that the entity title is also one of the attributes. Following previous work (Sevgili et al., 2022), we solve this task through a two-step pipeline: *Candidate Retrieval*, which retrieves top- K candidate entities $\{e_0, \dots, e_K\}$ from the KB, and *Entity Disambiguation*, which selects the gold entity e^+ from the K candidates $\{e_0, \dots, e_K\}$. Note that e^+ may not be in $\{e_0, \dots, e_K\}$ due to the retrieval error.

4.2 Candidate Retrieval

As shown in Figure 3, we retrieve a set of candidate entities from the KB based on textual and visual similarity. For efficiency purposes, we aim to first generate a lookup embedding for each review and entity based on their textual descriptions and visual images, so that the representations can be cached to enable efficient retrieval.

Text Cosine Similarity We apply SBERT (Devlin et al., 2018) to take the review text t_r and entity text t_{e_j} ¹¹ as input, respectively, and output their representations \mathbf{T}_r and \mathbf{T}_{e_j} ¹² based on the CLS token. Then we compute a textual cosine similarity score $s_t^R(m_r, e_j)$ for each pair.

$$s_t^R(m_r, e_j) = \text{cosine}(\mathbf{T}_r, \mathbf{T}_{e_j}) \quad (1)$$

The SBERT model is fine-tuned based on the InfoNCE loss (Van den Oord et al., 2018):

$$\mathcal{L}(\mathbf{T}_r, \mathbf{T}_{e^+}, \mathcal{T}) = -\log \frac{\exp[\cos(\mathbf{T}_r, \mathbf{T}_{e^+})]}{\sum_{\mathbf{T}_{e_j} \in \mathcal{T}} \exp[\cos(\mathbf{T}_r, \mathbf{T}_{e_j})]} \quad (2)$$

where e^+ is the gold entity of mention m_r , \mathcal{T} is text representations of candidate entities for m_r , including the gold entity e^+ , standard negative entities whose product categories are different from the gold entity, and in-batch negative entities that are candidate entities to other reviews in the same batch¹³.

Image Cosine Similarity To incorporate visual similarity, we employ CLIP (Radford et al., 2021) to obtain image representations, followed by a cosine similarity computation. Since multiple images exist for each review and entity, the CLIP model is fine-tuned based on the InfoNCE loss computed for each review image.

$$\mathcal{L}(\mathbf{V}_r^q, \mathbf{V}_{e^+}^h, \mathcal{T}) = -\log \frac{\exp[\cos(\mathbf{V}_r^q, \mathbf{V}_{e^+}^h)]}{\sum_{\mathbf{V}_{e_j}^i \in \mathcal{T}} \exp[\cos(\mathbf{V}_r^q, \mathbf{V}_{e_j}^i)]} \quad (3)$$

where v_r^q is one review image, $v_{e_j}^i$ is one entity image, and \mathcal{T} is image representations of candidate entities for m_r , including the gold entity e^+ , standard negative entities whose product categories are different from the gold entity, and in-batch negative entities. The image cosine similarity score between mention m_r and entity e_j is the maximum cosine similarity between their image sets $\bar{\mathbf{V}}_r = \{v_r^0, \dots, v_r^q\}$ and $\bar{\mathbf{V}}_{e_j} = \{v_{e_j}^0, \dots, v_{e_j}^h\}$.

$$s_v^R(m_r, e_j) = \max_{v_r^q \in \bar{\mathbf{V}}_r, v_{e_j}^i \in \bar{\mathbf{V}}_{e_j}} \text{cosine}(\mathbf{V}_r^q, \mathbf{V}_{e_j}^i) \quad (4)$$

Candidate Selection A weighted sum is applied to the textual and visual cosine similarity scores

¹¹We append entity title, description, and attributes as the entity textual information for candidate retrieval phase because this combination achieves better performance than other combinations in our preliminary experiments, as shown in Table 7 in Appendix D.

¹²We use bold symbols to denote vector representations.

¹³We remove any duplicate sentences within the same batch

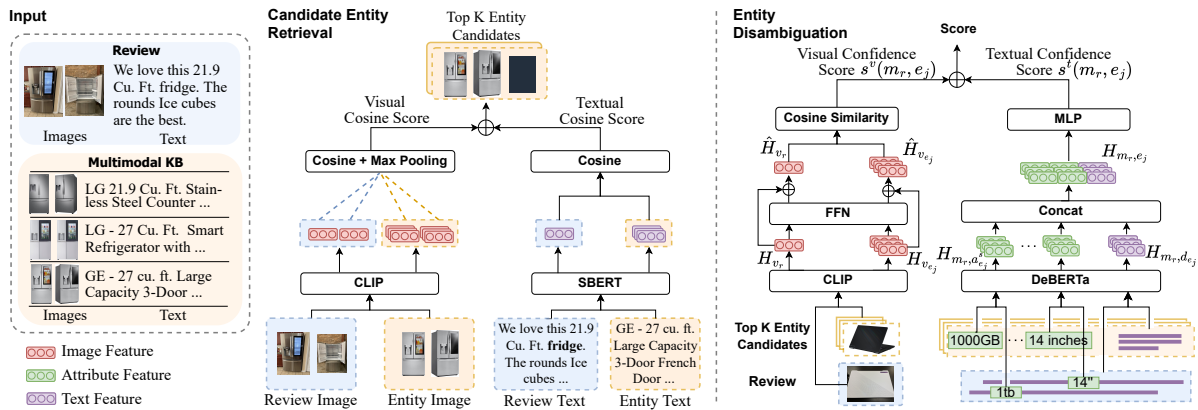


Figure 3: Candidate retrieval and entity disambiguation pipeline. We first retrieve the most relevant candidates using cosine similarity with regard to both textual descriptions and images and then predict the gold entity with the NLI-based text disambiguation and contrastive-learning-based image disambiguation modules.

to obtain the merged similarity scores $s^R(m_r, e_j)$, based on which, we select the top- K ranked entities as candidates.

$$s^R(m_r, e_j) = \lambda \cdot s_t^R(m_r, e_j) + (1 - \lambda) \cdot s_v^R(m_r, e_j) \quad (5)$$

where λ is a coefficient searched on the Dev set.

4.3 Entity Disambiguation

As shown in the right part of Figure 3, our disambiguation model comprises an NLI-based text module and a contrastive learning-based image module.

Preprocess We first apply four methods to extract attributes for each mention from its review text and images: (1) *OCR*: since there may exist text inside review images such as brand names, we apply an off-the-shelf OCR tool¹⁴ to recognize texts within each review image; (2) *String Match*, that identifies attribute values from review text based on the attribute values of the top- K retrieved candidate entities, e.g., if a candidate entity has an attribute value like “16 gigabyte” which also occurs in the review text, we will take it as a value for the attribute category “Memory”; (3) *ChatGPT* (OpenAI, 2022): in many cases, the review may contain the attribute value which is described in a slightly different form, such as “16 GB”, which cannot be identified by *String Match*. So, we further leverage ChatGPT and formalize our attribute value extraction task as a multiple-choice QA (Robinson et al., 2022) problem by treating each attribute category as a question and the corresponding attribute values from the top- K candidate entities as options, as detailed in Figure 5 in Appendix E. Limited by

¹⁴<https://github.com/JaidedAI/EasyOCR>

the computational cost, we apply ChatGPT on 11 common attribute categories, including “Brand, Color, Model Number, Product Title, Screen Size, Processor Brand, Processor Model, System Memory (RAM), Graphics, Solid State Drive Capacity, Processor Model” (4) *GPT-2* (Radford et al., 2019): for attribute categories not covered by *ChatGPT* method, we further apply GPT-2 to generate attribute values for all attribute categories in a zero-shot text completion manner, with the prompt template “Attribute Value Extraction:\n #Review_text \n #Attribute_key:”, where Attribute Value Extraction is the text prompt, #Review_text is the corresponding review text and #Attribute_key is the attribute to be extracted, as shown in Figure 6 in Appendix E. For all the approaches discussed above, we only keep the attribute values that match any attribute value of top- K candidate entities. The resulting attribute value set is denoted as System Attribute. We then filter out candidate entities whose attribute values do not match the System Attribute of each mention. Since we don’t manually label the attributes of mentions in the Train and Dev datasets, we clean the System Attribute to obtain Gold Attribute by removing the attributes that do not match with the attributes of the gold entity product.

Text-based Disambiguation Our text-based disambiguation module is based on Natural Language Inference (NLI) with the motivation that the review text should imply the product attribute if it is mentioned in the review. For example, given the review “I was hoping it would look more pink than it does, it’s more of a gray-toned light pink.

Not a dealbreaker. I still like this bag”, it should imply the attribute value of the target product, e.g., “The color of this bag is pink”, while contradicting attribute values of other products, e.g., “The color of this bag is black”. Thus, for each review with a mention m_r and text t_r , given a set of candidate entities $\{e_0, \dots, e_j\}$ with their descriptions $\{d_{e_0}, \dots, d_{e_j}\}$ and attribute values $\{a_{e_0}^0, \dots, a_{e_j}^0\}, \dots, \{a_{e_0}^s, \dots, a_{e_j}^s\}$, as there could be many attributes of candidate entities that are not mentioned in the review, we first select a subset of attribute values for the candidate entities based on the attribute categories covered in the System Attribute of mention m_r . Then, we pair each entity attribute or the entity description with the review description and feed each pair into a DeBERTa (He et al., 2023) encoder, which is fine-tuned and has shown promising performance on general NLI tasks, to obtain their contextual representations

$$H_{m_r, d_{e_j}} = \text{DeBERTa}(d_{e_j}, [m_r : t_r]) \quad (6)$$

$$H_{m_r, a_{e_j}^s} = \text{DeBERTa}([m_r : t_r], a_{e_j}^s) \quad (7)$$

where $:$ denotes the concatenation operation. For each entity with multiple attribute values, we concatenate all the contextual representations obtained from DeBERTa and feed it through MLP to predict the final NLI score:

$$H_{m_r, e_j} = [H_{m_r, a_{e_j}^0} : H_{m_r, a_{e_j}^1} \dots, H_{m_r, a_{e_j}^s} : H_{m_r, d_{e_j}}] \quad (8)$$

$$s^t(m_r, e_j) = \text{MLP}(H_{m_r, e_j}) \quad (9)$$

During training, we optimize the text-based disambiguation module based on the cross-entropy objective:

$$\mathcal{L}^t(m_r, e^+) = -\log \frac{\exp(s^t(m_r, e^+))}{\sum_{j=0}^{K-1} \exp(s^t(m_r, e_j))} \quad (10)$$

where e^+ is the gold entity, and K is the number of retrieved candidate entities.

Image-based Disambiguation Given the review image v_r ¹⁵ and entity images for a set of candidate entities $\{v_{e_0}, \dots, v_{e_j}\}$, we feed them into CLIP to obtain their image representations $\{H_{v_r}, H_{v_{e_0}}, \dots, H_{v_{e_j}}\}$. Inspired by previous studies (Zhang et al., 2022; Gao et al., 2021; Sun et al., 2022a), we feed these through a feed-forward layer

¹⁵Following (Wang et al., 2022b; Sun et al., 2022a), we select one image for each review and entity during the disambiguation, based on the cosine similarity score of their CLIP representations, which also showed better performance than using all images in our preliminary experiments.

and residual connection to adapt the generic image representations to a task-oriented semantic space

$$\hat{H}_{v_{e_j}} = H_{v_{e_j}} + \text{ReLU}(H_{v_{e_j}} \cdot W_1^e) \cdot W_2^e \quad (11)$$

$$\hat{H}_{v_r} = H_{v_r} + \text{ReLU}(H_{v_r} \cdot W_1^r) \cdot W_2^r \quad (12)$$

where W_1^r and W_2^r are learnable parameters for review representation learning, W_1^e and W_2^e are learnable parameters for entity representation learning.

We apply the following contrastive loss during training based on the cosine similarity scores.

$$\mathcal{L}^v(m_r, e^+) = -\log \frac{\exp(\cos(\hat{H}_{v_r}, \hat{H}_{v_{e^+}}))}{\sum_{e_j \in B} \exp(\cos(\hat{H}_{v_r}, \hat{H}_{v_{e_j}}))} \quad (13)$$

where B is the set of all entities in the current batch since we utilize in-batch negatives to improve our model’s ability to distinguish between gold and negative entities.

Inference During inference, we combine the NLI score $s^t(m_r, e_j)$ from the text-based disambiguation module, the cosine similarity score $s^v(m_r, e_j)$ from the image-based disambiguation module and the retrieval score from the candidate retrieval model, and predict the entity with the highest weighted score $s(m_r, e_j)$ as the target

$$s^v(m_r, e_j) = \cos(\hat{H}_{v_r}, \hat{H}_{v_{e_j}}) \quad (14)$$

$$s(m_r, e_j) = \lambda_1 s^t(m_r, e_j) + \lambda_2 s^v(m_r, e_j) + (1 - \lambda_1 - \lambda_2) s^R(m_r, e_j) \quad (15)$$

where λ_1, λ_2 are coefficients tuned on the Dev set.

5 Experiments and Analysis

5.1 Candidate Retrieval

For each review, we retrieve the top- K candidate entities from the target KB and evaluate the retrieval performance based on Recall@ K ($K = 1, 10, 20, 50, 100$). As shown in Table 3: (1) the multimodal retrieval outperforms the single-modality retrieval, demonstrating that both text and image information complement each other. (2) All models have obtained significant improvements (e.g., an average improvement of Recall@10 is 25.3%) after fine-tuning, which indicates the effectiveness of fine-tuning on our dataset. (3) After fusing image and text cosine similarity scores, our model achieves 95% of Recall@100, which shows that most relevant entities can be retrieved from the multimodal knowledge base.

Modality	Method	Recall@1	Recall@10	Recall@20	Recall@50	Recall@100
T	Pre-trained SBERT	19.52	46.63	57.06	71.18	82.52
V	Pre-trained CLIP	14.45	39.00	47.25	59.25	68.77
T+V	Pre-trained CLIP/SBERT	27.14	59.76	67.75	77.49	82.60
T	Fine-tuned SBERT	32.65	66.65	76.87	87.34	93.32
V	Fine-tuned CLIP	28.06	62.82	71.76	80.48	86.25
T+V	Fine-tuned CLIP/SBERT	48.12	85.84	90.26	93.69	95.11

Table 3: Performance of candidate retrieval. The modality of T and V represents the textual context and visual context, respectively.

Modality	w Attribute	Method	Disambiguation F1 (%)	End-to-End F1 (%)
-	No	Random Baseline	10.00	8.58
V	No	V2VEL (Sun et al., 2022b)	19.27	16.78
T	No	V2TEL (Sun et al., 2022b)	19.57	17.07
T+V	No	V2VTEL (Sun et al., 2022b)	31.37	30.22
T+V	No	LLaVA (Liu et al., 2023a)	23.33	20.03
T+V	No	GHMFC (Amigo et al., 2022)	12.52	12.11
T+V	Filter	GHMFC* (Amigo et al., 2022)	23.25	21.78
T+V	No	Wikidiverse (Wang et al., 2022c)	12.95	10.93
T+V	Filter	Wikidiverse* (Wang et al., 2022c)	24.57	20.48
T+V	No	Our Approach_w/o_Attribute	52.53	44.85
T	System	Our Approach_w/o_Image	44.40	38.52
V	System	Our Approach_w/o_Text	42.61	36.64
T+V	System	Our Approach	60.30	51.54
T+V	Gold	Our Approach	73.08	62.87
T+V	No	Human	80.00	74.00

Table 4: Performance of entity disambiguation. Gold stands for the Gold Attribute mentioned in the review, System stands for the System Attribute predicted by our methods, while Filter applies a straightforward elimination of candidate entities whose entity attributes do not align with the predicted review attributes.

5.2 Entity Disambiguation

We further evaluate the entity disambiguation performance based on the micro F1-score under the (1) *End-to-End* setting, where models predict the target entity from the top- K ($K = 10$) retrieved entities, and (2) *Disambiguation* setting, where models are evaluated on a subset of testing instances if their gold entities exist in the top- K ($K = 10$) retrieved candidates. We compare our approach with a *Random Baseline* which chooses the target product randomly and several high-performing baselines for multimodal entity linking as detailed in Appendix F.

As shown in Table 4, our approach outperforms all baseline methods and reaches 51.54% of *End-to-End* F1 score. One reason for the low performance is the error propagation from the Candidate Retrieval phase to Disambiguation. Our model can reach 60.30% of F1 score under the *Disambiguation* setting when the gold entity exists in the retrieved candidate set.

To evaluate the impact of each modality on entity disambiguation, we design ablated models of our approach by removing text, image, or attributes from the model input. The results show that each

modality can benefit the disambiguation, while the attribute information brings a considerable performance improvement. A possible reason for this performance gap is that attributes provide a strong, direct signal for the coreference between the review context of each mention and its gold entity. In addition, during the text-based disambiguation, we use System Attribute to select a subset of attribute values for the candidate entities. However, the System Attribute may contain incorrect attributes or miss some attributes of the mention that are also contained in the review. To evaluate its impact on text-based disambiguation, for the Test set, we use Gold Attribute labeled by humans, which yields significantly higher F1 scores, e.g., 73.08% F1 for the *Disambiguation* setting and 62.87% for the *End-to-End* setting.

Finally, we also set up a human performance for entity disambiguation by randomly sampling 50 reviews, with 10 candidate entities for each review, for the *Disambiguation* setting and 50 reviews for the *End-to-End* setting, and ask two annotators to execute manual entity-linking. Based on Fleiss κ (Fleiss, 1971), the agreement score between the two annotators is 0.69 for the *Disambiguation* set-

ting and 0.71 for the *End-to-End* setting. We consider a human prediction accurate only if both annotators provide the true label. As we can see in Table 4, there is a considerable gap between our model and *Human Performance*.

6 Remaining Challenges

We randomly sample 50 reviews linked to incorrect entities under the `System Attribute` setting from the `Test` and identify the following key challenges for the entity disambiguation task¹⁶.

Attribute Extraction: User reviews often contain informal language, idiomatic expressions, and diverse writing styles. This linguistic variability makes it challenging to accurately extract specific attribute values as different users might use other terms to describe the same attribute. Furthermore, our knowledge base encompasses over 30,000 attribute values. Determining the attribute referenced within a given review poses a challenging inference task. For 10% of the errors, our method fails to extract some key attributes. For example, given review #1 “*Plus their are 10 programmable buttons and rated to 50 million clicks with omron switches now you can’t beat that.*” in Figure 7 in Appendix G, we can distinguish the gold entity with 10 buttons from the candidate entity with 17 buttons after extracting the attribute “`Number of Buttons (Total): 10`”. Recognizing brand logos or integrating a better OCR model to detect text within images will also increase the quality of `System Attribute`, as shown in review #2. More analysis on attribute extraction module is detailed in Appendix H.

Reasoning over Attributes: 18% of the errors can be fixed if the model pays more attention to appropriate attributes or conducts reasoning based on the attribute. For example, the review #3 in Figure 7 in Appendix G claims “*i bought this because you can use it on your phone too*”. As a result, we can skip the candidate entity with the attribute “`Compatible Platform(s): Windows, Mac, PlayStation 4, PlayStation 5`” since it does not support phones. In some cases, `System Attribute` contains the key review attributes to distinguish the gold entity from the candidate entity. However, the model is fed with abundant multimodal context and fails to focus on the distinguishable attribute. For

example, in review #4, the model fails to take care of the attribute “`Carafe Capacity`”.

Fine-grained Image Matching: In 32% of the errors, the gold entity and candidate entity can be distinguished based on fine-grained image texture. For example, in review #5 in Figure 7 in Appendix G, the delicate pattern in the computer case acts as the main hint to link to the gold entity. Since these inconspicuous patterns can be pretty elusive to spot, visual attributes will be helpful to guide attention in some cases. For example, in review #6, the difference between the gold and candidate entities is whether there is an Ice and Water Dispenser on the fridge surface. With the visual attribute “`Ice and Water Dispenser Location: External`,” the model can focus on the image patch on which the Ice and Water Dispenser is normally located.

Candidate Retrieval: 26% of the disambiguation errors are due to the gold entity not being in the top 10 retrieved candidates, a.k.a. error propagation from the candidate retrieval phase. We notice the following retrieval error patterns by comparing the gold entity with the top 10 retrieved candidates. (1) Similar to the disambiguation phase, attribute-based match and fine-grained image match can help distinguish the gold entity from candidate entities. (2) One unique error in the retrieval phase happens when one of the review images is irrelevant to the gold entity, thus introducing noise when computing the average image similarity score.

7 Conclusion

We propose attribute-aware multimodal entity linking, which requires features extracted from images, text descriptions, and structured attributes to disambiguate and link each mention to the corresponding entity in a target knowledge base. To support this research, we construct AMELI, consisting of a multimodal knowledge base that contains 34,690 product entities described with text, images, and fine-grained attributes, and a multimodal review dataset that contains 16,735 review instances while each review is also associated with a text description and an image. We experiment with several high-performance entity-linking approaches, including a new approach that incorporates attributes of entities for disambiguation. Experimental results show that the attributes indeed significantly enhance the model performance, but still, there is a large gap between the machine and human performance.

¹⁶We analyze the `System Attribute` setting instead of the `Gold Attribute` since the gold review attribute may not always be available in the real world application.

Limitations

Advanced Approach to Incorporate Attributes

In this research endeavor, we propose an innovative approach incorporating attributes into the disambiguation process using a Natural Language Inference (NLI)-based framework. However, we acknowledge that this approach may not fully harness the potential of attributes. Attribute-aware encoding (Wei et al., 2021; Saini et al., 2022), attribute-based zero-shot learning (Lampert et al., 2013), and attribute-aware retrieval (Wei et al., 2021; Dong et al., 2023) can be promising directions for future work.

Ethics Statement

We carefully follow the ACM Code of Ethics¹⁷ and have not found potential societal impacts or risks so far. To the best of our knowledge, this work has no notable harmful effects and uses, environmental impact, fairness considerations, privacy considerations, security considerations, or other potential risks. Our dataset does not contain sensitive personally identifiable information such as name, address, or phone number.

Since our dataset contains user reviews, some claims may be offensive and we remove reviews containing profanity words as specified in Sec 3.

Acknowledgements

This research is based upon work supported by Meta AI. We also extend our gratitude to Daniel Hajjaligol for his assistance with the initial web crawling, to Meghana Holla for her assistance in refining the initial version of this paper, and to Jingyuan Qi, Ying Shen, Zoe Zheng, Pritika Ramu, Samhita Reddivalam, Sai Gurrapu, Indrajeet Kumar Mishra, Mingchen Li, and Mohammad Beigi for their invaluable contributions to dataset annotation and human experimentation.

References

Omar Adjali, Romaric Besançon, Olivier Ferret, Hervé Le Borgne, and Brigitte Grau. 2020. Multimodal entity linking for tweets. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part I*, pages 463–478. Springer.

¹⁷<https://www.aclweb.org/portal/content/acl-code-ethics>

Enrique Amigo, Pablo Castells, Julio Gonzalo, Ben Carterette, J Shane Culpepper, Gabriella Kazai, Peng Wang, Jiangheng Wu, and Xiaohang Chen. 2022. *Multimodal Entity Linking with Gated Hierarchical Fusion and Contrastive Training*. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 938–948.

Tom Ayoola, Joseph Fisher, and Andrea Pierleoni. 2022a. Improving entity disambiguation by reasoning over a knowledge base. *arXiv preprint arXiv:2207.04106*.

Tom Ayoola, Shubhi Tyagi, Joseph Fisher, Christos Christodoulopoulos, and Andrea Pierleoni. 2022b. Refined: An efficient zero-shot-capable approach to end-to-end entity linking. *arXiv preprint arXiv:2207.04108*.

Tom Ayoola, Shubhi Tyagi, Joseph Fisher, Christos Christodoulopoulos, and Andrea Pierleoni. 2022c. *ReFinED: An Efficient Zero-shot-capable Approach to End-to-End Entity Linking*. *arXiv*.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. *Highly Parallel Autoregressive Entity Linking with Discriminative Correction*. *arXiv*.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2020. *Autoregressive Entity Retrieval*. *arXiv*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. *Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality*.

Silviu Cucerzan. 2007. *Large-scale named entity disambiguation based on Wikipedia data*. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 708–716, Prague, Czech Republic. Association for Computational Linguistics.

Nicola De Cao, Ledell Wu, Kashyap Popat, Mikel Artetxe, Naman Goyal, Mikhail Plekhanov, Luke Zettlemoyer, Nicola Cancedda, Sebastian Riedel, and Fabio Petroni. 2022. Multilingual autoregressive entity linking. *Transactions of the Association for Computational Linguistics*, 10:274–290.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jianfeng Dong, Xiaoman Peng, Zhe Ma, Daizong Liu, Xiaoye Qu, Xun Yang, Jixiang Zhu, and Baolong Liu. 2023. From region to patch: Attribute-aware foreground-background contrastive learning for fine-grained fashion retrieval. *arXiv preprint arXiv:2305.10260*.

- Shahi Dost, Luciano Serafini, Marco Rospocher, Lamberto Ballan, and Alessandro Sperduti. 2020. Vtkel: a resource for visual-textual-knowledge entity linking. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, pages 2021–2028.
- Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.
- Jingru Gan, Jinchang Luo, Haiwei Wang, Shuhui Wang, Wei He, and Qingming Huang. 2021. Multimodal entity linking: a new dataset and a baseline. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 993–1001.
- Octavian-Eugen Ganea and Thomas Hofmann. 2017a. [Deep joint entity disambiguation with local neural attention](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2619–2629, Copenhagen, Denmark. Association for Computational Linguistics.
- Octavian-Eugen Ganea and Thomas Hofmann. 2017b. [Deep joint entity disambiguation with local neural attention](#). *arXiv preprint arXiv:1704.04920*.
- Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. 2021. [CLIP-Adapter: Better Vision-Language Models with Feature Adapters](#). *arXiv*.
- Yike Guo, Faisal Farooq, Guineng Zheng, Subhabrata Mukherjee, Xin Luna Dong, and Feifei Li. 2018. [OpenTag](#). *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1049–1058.
- Zhaochen Guo and Denilson Barbosa. 2018. Robust named entity disambiguation with random walks. *Semantic Web*, 9(4):459–479.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. [Deep residual learning for image recognition](#).
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. [Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing](#).
- Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenauf, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 782–792.
- Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. 2023. [Open-domain Visual Entity Recognition: Towards Recognizing Millions of Wikipedia Entities](#). *arXiv*.
- Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.
- Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. 2018. [End-to-end neural entity linking](#). *arXiv preprint arXiv:1808.07699*.
- Tuan Manh Lai, Heng Ji, and ChengXiang Zhai. 2022. [Improving Candidate Retrieval with Entity Profile Generation for Wikidata Entity Linking](#). *arXiv*.
- Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. 2013. Attribute-based classification for zero-shot visual object categorization. *IEEE transactions on pattern analysis and machine intelligence*, 36(3):453–465.
- PengYuan Li and YongLi Wang. 2021. A multimodal entity linking approach incorporating topic concepts. In *2021 International Conference on Computer Information Science and Artificial Intelligence (CISAI)*, pages 491–494. IEEE.
- Rongmei Lin, Xiang He, Jie Feng, Nasser Zalmout, Yan Liang, Li Xiong, and Xin Luna Dong. 2021. Pam: Understanding product images in cross product category attribute extraction. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3262–3270.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. [Visual instruction tuning](#). *arXiv preprint arXiv:2304.08485*.
- Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023b. [Grounding dino: Marrying dino with grounded pre-training for open-set object detection](#). *arXiv preprint arXiv:2303.05499*.
- Lajanugen Logeswaran, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. 2019. [Zero-shot entity linking by reading entity descriptions](#). *arXiv preprint arXiv:1906.07348*.
- David Milne and Ian H Witten. 2008. Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 509–518.
- Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018. Multimodal named entity disambiguation for noisy social media posts. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2000–2008.
- Yasumasa Onoe and Greg Durrett. 2020. Fine-grained entity typing for domain independent entity linking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8576–8583.
- OpenAI. 2022. [Openai: Introducing chatgpt](#).

- Matthew E Peters, Mark Neumann, Robert L Logan IV, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A Smith. 2019. Knowledge enhanced contextual word representations. *arXiv preprint arXiv:1909.04164*.
- 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*, pages 8748–8763. PMLR.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Manoj Prabhakar Kannan Ravi, Kuldeep Singh, Isaiah Onando Mulang, Saeedeh Shekarpour, Johannes Hoffart, and Jens Lehmann. 2021. Cholan: A modular approach for neural entity linking on wikipedia and wikidata. *arXiv preprint arXiv:2101.09969*.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.
- Joshua Robinson, Christopher Michael Rytting, and David Wingate. 2022. Leveraging large language models for multiple choice question answering. *arXiv preprint arXiv:2210.12353*.
- Nirat Saini, Khoi Pham, and Abhinav Shrivastava. 2022. Disentangling visual embeddings for attributes and objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13658–13667.
- Özge Sevgili, Artem Shelmanov, Mikhail Arkhipov, Alexander Panchenko, and Chris Biemann. 2022. [Neural entity linking: A survey of models based on deep learning](#). *Semantic Web*, 13(3):527–570.
- Senbao Shi, Zhenran Xu, Baotian Hu, and Min Zhang. 2023. [Generative multimodal entity linking](#).
- Lin Sun. 2017. Research on product attribute extraction and classification method for online review. In *2017 International Conference on Industrial Informatics-Computing Technology, Intelligent Technology, Industrial Information Integration (ICIICII)*, pages 117–121. IEEE.
- Wenxiang Sun, Yixing Fan, Jiafeng Guo, Ruqing Zhang, and Xueqi Cheng. 2022a. [Visual Named Entity Linking: A New Dataset and A Baseline](#). *arXiv*.
- Wenxiang Sun, Yixing Fan, Jiafeng Guo, Ruqing Zhang, and Xueqi Cheng. 2022b. Visual named entity linking: A new dataset and a baseline. *arXiv preprint arXiv:2211.04872*.
- Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.
- Hongyin Tang, Xingwu Sun, Beihong Jin, and Fuzheng Zhang. 2021. A bidirectional multi-paragraph reading model for zero-shot entity linking. *35th AAAI Conference on Artificial Intelligence, AAAI 2021*, 15:13889–13897.
- Aaron Van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv e-prints*, pages arXiv–1807.
- Aparna Nurani Venkatasubramanian, Tinne Tuytelaars, and Marie Francine Moens. 2017. [Entity linking across vision and language](#). *Multimedia Tools and Applications*, 76:22599–22622.
- Qifan Wang, Li Yang, Bhargav Kanagal, Sumit Sanghai, D Sivakumar, Bin Shu, Zac Yu, and Jon Elsas. 2020. Learning to extract attribute value from product via question answering: A multi-task approach. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 47–55.
- Qifan Wang, Li Yang, Jingang Wang, Jitin Krishnan, Bo Dai, Sinong Wang, Zenglin Xu, Madian Khabsa, and Hao Ma. 2022a. Smartave: Structured multimodal transformer for product attribute value extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 263–276.
- Sijia Wang, Alexander Hanbo Li, Henry Zhu, Sheng Zhang, Chung-Wei Hang, Pramuditha Perera, Jie Ma, William Wang, Zhiguo Wang, Vittorio Castelli, Bing Xiang, and Patrick Ng. 2023. [Benchmarking diverse-modal entity linking with generative models](#).
- Xuwu Wang, Junfeng Tian, Min Gui, Zhixu Li, Rui Wang, Ming Yan, Lihan Chen, and Yanghua Xiao. 2022b. [WikiDiverse: A multimodal entity linking dataset with diversified contextual topics and entity types](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4785–4797, Dublin, Ireland. Association for Computational Linguistics.
- Xuwu Wang, Junfeng Tian, Min Gui, Zhixu Li, Rui Wang, Ming Yan, Lihan Chen, and Yanghua Xiao. 2022c. [WikiDiverse: A Multimodal Entity Linking Dataset with Diversified Contextual Topics and Entity Types](#). *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4785–4797.
- Xiu-Shen Wei, Yang Shen, Xuhao Sun, Han-Jia Ye, and Jian Yang. 2021. A²-net: Learning attribute-aware hash codes for large-scale fine-grained image retrieval. *Advances in Neural Information Processing Systems*, 34:5720–5730.
- Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2019. Scalable zero-shot entity linking with dense entity retrieval. *arXiv preprint arXiv:1911.03814*.
- Huimin Xu, Wenting Wang, Xin Mao, Xinyu Jiang, and Man Lan. 2019. [Scaling up Open Tagging from Tens](#)

to Thousands: Comprehension Empowered Attribute Value Extraction from Product Title. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5214–5223.

Jun Yan, Nasser Zalmout, Yan Liang, Christian Grant, Xiang Ren, and Xin Luna Dong. 2021. *AdaTag: Multi-Attribute Value Extraction from Product Profiles with Adaptive Decoding*. *arXiv*.

Li Yang, Qifan Wang, Zac Yu, Anand Kulkarni, and Sumit Sanghai. 2022. *MAVE: A Product Dataset for Multi-source Attribute Value Extraction*. *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 1256–1265.

Xiyuan Yang, Xiaotao Gu, Sheng Lin, Siliang Tang, Yueting Zhuang, Fei Wu, Zhigang Chen, Guoping Hu, and Xiang Ren. 2019. Learning dynamic context augmentation for global entity linking. *arXiv preprint arXiv:1909.02117*.

Li Zhang, Zhixu Li, and Qiang Yang. 2021a. Attention-based multimodal entity linking with high-quality images. In *Database Systems for Advanced Applications: 26th International Conference, DASFAA 2021, Taipei, Taiwan, April 11–14, 2021, Proceedings, Part II 26*, pages 533–548. Springer.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. 2022. Tip-adaptor: Training-free adaption of clip for few-shot classification. In *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pages 493–510. Springer.

Wenzheng Zhang, Wenyue Hua, and Karl Stratos. 2021b. Entqa: Entity linking as question answering. *arXiv preprint arXiv:2110.02369*.

Qiushuo Zheng, Hao Wen, Meng Wang, Guilin Qi, and Chaoyu Bai. 2022. Faster zero-shot multi-modal entity linking via visual-linguistic representation. *Data Intelligence*, 4(3):493–508.

Xingchen Zhou, Peng Wang, Guozheng Li, Jiafeng Xie, and Jiangheng Wu. 2021. Weibo-mel, wikidata-mel and richpedia-mel: Multimodal entity linking benchmark datasets. In *Knowledge Graph and Semantic Computing: Knowledge Graph Empowers New Infrastructure Construction: 6th China Conference, CCKS 2021, Guangzhou, China, November 4-7, 2021, Proceedings 6*, pages 315–320. Springer.

Tiangang Zhu, Yue Wang, Haoran Li, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. *Multimodal Joint Attribute Prediction and Value Extraction for E-commerce Product*. *arXiv*.

A Filtering of Uninformative Reviews

For each review and its corresponding product, we extract four features, including *# of mentioned attributes* (i.e., the number of product attributes

mentioned in the review based on string match), *image-based similarity* (i.e., the maximum similarity between review images and gold product images based on CLIP (Radford et al., 2021) image embeddings), *description-based similarity* (i.e., the similarity between gold product description and review text based on SBERT (Reimers and Gurevych, 2019)), *title-based similarity* (i.e., the similarity between the gold product title and review text using SBERT (Reimers and Gurevych, 2019)). We further manually annotate 500 pairs of reviews and products while each pair is assigned with a label: *positive* if the review is informative enough to correctly link the mention to the target product, otherwise, *negative*, and use them to evaluate a threshold-based approach which predicts the reviews as uninformative reviews if the four extracted feature scores, $\{\# \text{ of mentioned attributes, image-based similarity, description-based similarity, title-based similarity}\}$, do not overpass the four corresponding thresholds, which are hyperparameters searched on these examples. The threshold-based method reaches 85% of precision and 82% of recall in predicting informative reviews on these 500 examples¹⁸. We further apply it to clean the dataset by removing the reviews predicted as uninformative.

B Human Annotation

We recruited 12 student volunteers as annotators. 8 of them are from China, and 4 volunteers are from India. For human annotation, we provide the annotation tool as shown in Figure 4 and the following instructions to annotators:

- Open one of your annotation web pages
- Click on product 1 to expand its text, images, and attributes information. Compare the review with product 1. Are there any specific product attributes, e.g., memory size, color, can be recognized in the review text/images? Do review images and product images share the same color, shape, or subtle pattern?
- If you want to zoom in on any image, you can click on the image, and it will be shown on full screen.

¹⁸We compared the threshold-based method with a series of classifiers, like SVM, by training these classifiers on 385 examples and testing on 165 examples. Threshold-based Method reaches the highest accuracy.

- If it seems that product 1 is not the target product, you can fold it by double-clicking the dark product header, and begin to check product 2, product 3, and so on.
- Finally, you find the target product. Now you can record the index (1-10) on the provided answer sheet.

Table 5 shows the annotation accuracy for each annotator.

Annotator ID	#Correct	#Finished	Accuracy (%)
1	244	330	73.94
2	256	330	77.58
3	274	330	83.03
4	240	330	72.73
5	270	324	83.33
6	253	330	76.67
7	124	170	72.94
8	290	330	87.88
9	222	330	67.27
10	295	330	89.39
11	272	330	82.42
12	285	330	86.36
Overall	3025	3794	79.73

Table 5: The annotation result. #Finished stands for the reviews annotated by the corresponding annotator, while #Correct stands for the correct prediction.

C Category Distribution

Table 6 shows the category distribution.

Category	# Product	Percentage %
All Refrigerators	847	2.44
Action Figures (Toys)	730	2.10
Dash Installation Kits	682	1.97
Wall Mount Range Hoods	680	1.96
Nintendo Switch Games	628	1.81
Gas Ranges	603	1.74
Building Sets & Blocks (Toys)	576	1.66
Nintendo Switch Game Downloads	574	1.65
PC Laptops	554	1.60
Cooktops	547	1.58

Table 6: Category Distribution of 10 most frequent categories. # **Product** indicates the number of products in the corresponding category while **Precentage** indicates how many percentages of all products are in this category.

D Preliminary Experiments

Table 7 shows the preliminary experiments on candidate retrieval.

E Prompt Templates for GPT-2, ChatGPT, Vicuna, and LLaVA

We show the applied prompt templates in Figure 5 and Figure 6.

F Baseline Approaches

We compare our approach with several baselines on the entity disambiguation task:

- a *Random Baseline* which chooses the target product randomly;
- *V2VEL* (Sun et al., 2022b), which is a visual entity linking model with entity image and mention image as the input, Resnet150 (He et al., 2015) as the image encoder, and one adapter layer to adapt the representation to the task representation space;
- *V2TEL* (Sun et al., 2022b), which incorporates CLIP to encode entity text and mention image for prediction;
- *V2VTEL* (Sun et al., 2022b), which combines *V2VEL* and *V2TEL* in a two-step retrieval-then-rerank pipeline. We first apply the trained *V2VEL* model to select top- L entities from top- K candidate entities ($K>L$), then use the trained *V2TEL* model to predict the gold entity from top- L entities. We set $K=10$ and $L=5$ in our experiments.
- *GHMFC* (Amigo et al., 2022), which applies textual-guided visual attention and visual-guided textual attention to extract multimodal features, followed by a gated fusion and contrastive training;
- *Wikidiverse* (Wang et al., 2022b)¹⁹, which concatenates patch-level image representation and token-level text representation and feeds them into a self-attention transformer for multimodal fusion;
- *GHMFC** and *Wikidiverse**, where we improve *GHMFC* (Amigo et al., 2022) and *Wikidiverse* (Wang et al., 2022b) with a post-process “Attribute Filter”, which leverages a straightforward elimination of candidate entities whose entity attributes do not align with the predicted review attributes.

¹⁹*V2VEL*, *V2TEL*, *V2VTEL*, *GHMFC*, and *Wikidiverse* are all fine-tuned on our dataset. For a fair comparison, they are used to predict the gold entity from top- K candidate entities, the same setting as our method.

Review ID: 29734

Review Text: Excellent wireless keyboard!. I purchased this wireless keyboard because of the price and small form-factor. The size is perfect and the keys are very responsive. It has media keys for playback and volume, also a complete number pad. Overall the keyboard is perfect for anyone looking for a wireless keyboard and modern look.

Review Image:



1.

Expand All

1. Logitech - K360 Full-size Wireless Scissor Keyboard - Black. (Click to unfold)
2. Logitech - MK360 Full-size Wireless Scissor Keyboard and Mouse - Black. (Click to unfold)
3. Microsoft - Designer Compact Wireless Keyboard - Matte Black. (Click to unfold)
4. Logitech - K400 Plus TKL Wireless Membrane Keyboard for PC/TV/Laptop/Tablet with Built-in Touchpad - Black. (Click to unfold)
5. Logitech - K380 TKL Wireless Bluetooth Scissor Keyboard for PC, Laptop, Windows, Mac, Android, iPad OS, Apple TV - Gray. (Click to unfold)
6. Microsoft - All-In-One Media Wireless Keyboard with Track Pad - Black. (Click to unfold)
7. Logitech - K580 Multi-Device Chrome OS Edition Full-size Wireless Membrane Keyboard - Graphite. (Click to unfold)
8. Logitech - MX Keys Mini TKL Wireless Bluetooth Scissor Keyboard with Backlit Keys - Black. (Click to unfold)
9. Logitech - MX Mechanical Mini Compact Wireless Mechanical Clicky Switch Keyboard for Windows/macOS with Backlit Keys - Graphite. (Click to unfold)
10. Logitech - MK470 Full-size Wireless Scissor Keyboard and Mouse Bundle with Plug and Play - Black/Gray. (Click to unfold)

Figure 4: Screenshot of human annotation tool.

- *LLaVA* (Liu et al., 2023a), where we conduct an experiment of employing a SOTA multimodal large model, i.e., *LLaVA*, directly for attribute-aware multimodal entity linking in a few-shot manner. Specifically, given a particular review and an entity mention, we first employ the same candidate retrieval approach to obtain the top- K ($K=10$) candidate entities, then we ask the *LLaVA* model to directly choose the most plausible candidate entity title from the 10 candidate entity titles based on the multiple-choice QA prompt template in Figure 5.

G Error Examples

We show several error examples of Attribute-aware Multimodal Entity Linking in Figure 7.

H Attribute Extraction Performance

We have conducted experiments to analyze the performance of each individual attribute extractor and obtained 54.61%, 53.41%, 27.28%, and 22.24% F-scores on attribute value extraction, corresponding to *String Match*, *zero-shot GPT-2*, *ChatGPT*, and *OCR*, as shown in Table 8. Combining these four extractors leads to a significantly higher F1 score of 76.39%. To shed light on future research, we further conduct experiments of applying *GPT-2*, and an open-source LLM, *Vicuna* (Chiang et al., 2023), for few-shot attribute extraction, and obtained attribute extraction F1 scores of 59.47% and 58.28% on the Test set, respectively. Due to the computation cost, we set *Vicuna*'s max token length to 64, which may hurt the performance. In this study, we concentrate on establishing the baseline per-

Text Field	Method	Recall@1	Recall@10	Recall@20	Recall@50	Recall@100
Title	Pre-trained SBERT	12.29	38.12	48.92	63.81	75.88
Desc	Pre-trained SBERT	14.85	40.42	50.31	63.81	74.17
Attri	Pre-trained SBERT	12.81	36.88	47.46	63.77	77.67
Title+Attri	Pre-trained SBERT	16.42	43.56	54.10	67.68	79.53
Title+Desc	Pre-trained SBERT	19.08	47.21	57.10	69.21	79.82
Attri+Desc	Pre-trained SBERT	17.62	45.06	55.45	69.06	80.48
Title+Desc+Attri	Pre-trained SBERT	19.52	46.63	57.06	71.18	82.52

Table 7: Preliminary performance of entity candidate retrieval based on the cosine similarity between the review text and the corresponding entity text field. “Title”, “Desc”, and “Attri” stand for entity title, entity description, and entity attributes, respectively. “Desc+Title” stands for the concatenation of entity description and entity title.

Multiple Choice Question Answering Prompt Template

{few-shot demonstrations}
Review: Splatoon 2. We love Splatoon 2, the only downfall is that it is one player, we would absolutely love if you could play multiplayer on one console. Very fun, very colorful, we are loving this game({review_text})
{review_image}
Question: What is the product title ({attribute_category}) of the game ({mention}) based on this review?
A. Splatoon 3 - Nintendo Switch (OLED Model), Nintendo Switch, Nintendo Switch Lite ({attribute candidate 1})
B. Belkin - USB-C 11-in-1 Multiport Dock - Gray ({attribute candidate 2})
C. Splatoon 2 Standard Edition - Nintendo Switch ({attribute candidate 3})
D. {another 7 candidates}
Answer:

Figure 5: The multiple-choice QA prompt template is applied in ChatGPT-based and Vicuna-based Attribute Extraction and LLaVA-based Entity Disambiguation.

Attribute Extractor	Precision (%)	Recall (%)	F1 (%)
String Match	97.82	37.88	54.61
Zero-shot GPT-2	92.38	37.57	53.41
Zero-shot ChatGPT	64.57	17.29	27.28
OCR	98.46	12.54	22.24
Match+GPT2+ChatGPT+OCR	94.33	64.18	76.39
Few-shot GPT-2	90.04	44.39	59.47
Few-shot Vicuna	74.05	48.04	58.28

Table 8: Performance of attribute value extraction. The term "Match+GPT2+ChatGPT+OCR" signifies the combination of the String Match, zero-shot GPT-2, ChatGPT, and OCR extractor. Due to the computation cost, ChatGPT is only applied for a subset of attribute categories and Vicuna’s max token length is set to 64

formance for our attribute-aware multimodal entity linking task, and we encourage subsequent research to investigate more advanced methods for extracting and utilizing attribute information.

I Application Scenarios

To demonstrate the broad application scenario of our proposed attribute-aware entity linking task and approach, we employ both the *String Match* attribute extractor and *Vicuna* attribute extractor on the popularly used entity linking dataset. As illustrated in Table 9, on the Richpedia (MEL-Bench) (Zhou et al., 2021) dataset, a public benchmark dataset for multimodal entity linking, the

average number of attributes extracted from each mention context is 2.06. Note that the number is only based on the textual descriptions while the images in our multimodal entity linking task may contain more visual attributes. In addition, two other studies (Hu and Liu, 2004; Sun, 2017) have also reported the extraction of 2.20 and 1.11 attributes, respectively, from each mention context within their datasets. Finally, within our dataset, our attribute extractors reveal an average of 1.65 attributes per review, while human annotation yields an average of 3.54 attributes per review. This discrepancy highlights the potential for uncovering more attributes with advanced attribute extractors. Based on these statistics, we respectfully assert that

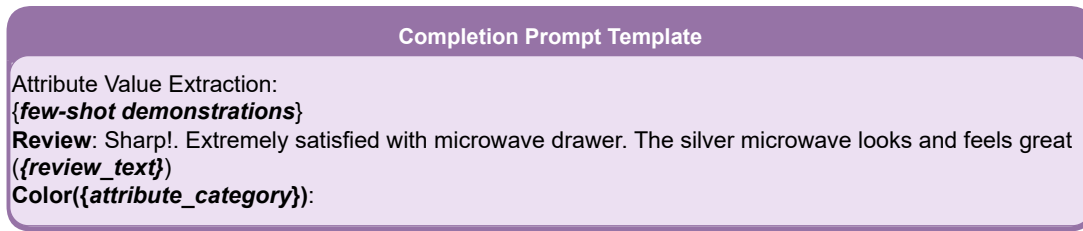


Figure 6: The text completion prompt template is applied in GPT-2 based Attribute Extraction.

Dataset	Attribute Extractor	#Attribute	#Mention Context	Attribute/Mention Ratio
Ours	System	27533	16735	1.65
Ours - Test Set	Human	9716	2741	3.54
MELBench-Richpedia (Zhou et al., 2021)	System	36705	17800	2.06
(Sun, 2017)	System*	2198	1000	2.20
(Hu and Liu, 2004)	System*	348	314	1.11

Table 9: Statistics of attributes within mention context in several datasets. The term "System*" signifies that the attributes have been extracted and documented in the respective work, rather than by our system.

within the Entity Linking (EL) scenario, entity attributes are frequently either explicitly mentioned or implicitly implied within the mention context, and thus, our proposed attribute-aware entity linking task and approach have broad application scenarios.

J Experiment Details

One training for the candidate retrieval model can be done with 1 NVIDIA A40 for 10 hours. One training for the entity disambiguation model can be done with 4 NVIDIA A40 for 7 hours. The search space of hyperparameters for the entity disambiguation model is as follows: the learning rate $\in \{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 5 \times 10^{-3}, 5 \times 10^{-4}\}$ and batch size $\in \{12, 16, 20, 24, 32\}$.

K Data Statement

K.1 Licensing

Our dataset is licensed under the CC BY 4.0²⁰. The associated codes to AMELI for data crawler and baseline are licensed under Apache License 2.0²¹.

K.2 Intended Use

Our dataset contains products and user reviews in English from E-commerce domains.

The dataset can be used for attribute-aware multimodal entity linking task. A model trained on this task can also be used to link user posts to some

products or general entities, a.k.a. detecting user interests from social media.

The dataset can also be used in the unimodal setting, like text-only entity linking.

K.3 Dataset Format

Our dataset encompasses a new multimodal entity linking benchmark dataset that contains 16,735 mentions described in text and associated with 30,472 images and a multimodal knowledge base that covers 34,690 entities along with 177,873 entity images and 798,216 attributes.

1. Multimodal knowledge base

- (a) Image folder “product_images”, which contains all entity images.
- (b) Entity information JSON file named “best-buy_products.json”, which contains entity text, image name, and attributes.
 - i. product_category: Category of the product, e.g., “Video Games -> Nintendo Switch -> Nintendo Switch Games”
 - ii. product_name: Name for the product
 - iii. overview_section:
 - A. description: Description of the product
 - iv. image_path: filename of the corresponding image
 - v. image_url: The link to the corresponding BestBuy image
 - vi. Spec: Attribute category and attribute value pairs for the product
 - vii. id: Unique ID for the product

²⁰<https://creativecommons.org/licenses/by/4.0/>

²¹<https://www.apache.org/licenses/LICENSE-2.0>






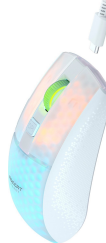












Review	Predicted Product	Gold Product
<p>#1. I love the set up of Icue. Plus their are 10 programmable buttons and rated to 50 million clicks with omron swiches now you can't beat that</p> 	<p>Name: CORSAIR - Scimitar RGB Elite Wired Optical Gaming Mouse with 17 Programmable Buttons - Black</p> <p>Attribute: Number of Buttons (Total) : 17</p> 	<p>Name: CORSAIR - IRONCLAW RGB Wireless Optical Gaming Mouse - Black</p> <p>Attribute: Number of Buttons (Total) : 10</p> 
<p>#2. Lightweight, easy to use mouse. A++ Gaming Mouse. Tldr: Quality, lightweight wireless mouse that has stellar battery life per charge.</p> <p>OCR: Burst</p> 	<p>Name: ROCCAT - Kone Pro Air Lightweight Wireless Bluetooth Optical Gaming Mouse</p> <p>Attribute: Model: Kone Pro Air</p> 	<p>Name: ROCCAT - Burst Pro Air Lightweight Wireless Optical Gaming Ambidextrous Mouse</p> <p>Attribute: Model: Burst Pro Air</p> 
<p>#3. i replaced it for my old logitech g533. i bought this because you can use it on your phone too</p> 	<p>Name: SteelSeries - Arctis 9 Wireless Gaming Headset for PC, PS5, and PS4 - Black</p> <p>Attribute: Compatible Platform(s) : Windows, Mac, PlayStation 4, PlayStation 5</p> 	<p>Name: SteelSeries - Arctis 1 Wireless Stereo Gaming Headset for PC - Black</p> <p>Attribute: Compatible Platform(s) : Windows, PlayStation 4, PlayStation 5, Xbox Series S, Android</p> 
<p>#4. The removable water reservoir is convenient and easy to fill and clean. The capacity fits a large 70 ounces, or about 14 cups</p> <p>Attribute: Carafe Capacity : 14 cups</p> 	<p>Name: Ninja - Coffee 12-Cup Coffee Brewer - Silver</p> <p>Attribute: Carafe Capacity : 12 cups</p> 	<p>Name: Ninja - Programmable XL 14-Cup Coffee Maker PRO, Glass Caraf</p> <p>Attribute: Carafe Capacity : 14 cups</p> 
<p>#5. Corsair review. I gotta say this case from Corsair is well thought out</p> 	<p>Name: CORSAIR - 4000D AIRFLOW MidTower Case</p> <p>Attribute: Brand : CORSAIR</p> 	<p>Name: CORSAIR - iCUE 220T RGB Airflow ATX Mid-Tower Smart Case</p> <p>Attribute: Brand : CORSAIR</p> 
<p>#6. Excellent product. Lots of space of the right proportions in both fridge & freezer</p> 	<p>Name: Whirlpool - 25.2 Cu. Ft. French Door Refrigerator with Internal Water Dispenser</p> <p>Attribute: Ice and Water Dispenser Location : Internal</p> 	<p>Name: Whirlpool - 24.7 Cu. Ft. French Door Refrigerator</p> <p>Attribute: Ice and Water Dispenser Location : External</p> 

Figure 7: Examples of Attribute-aware Multimodal Entity Linking.

- viii. url: The link to the corresponding BestBuy webpage
2. Multimodal entity linking dataset, which is split into *Train*, *Dev*, *Test* subsets.
- (a) Image folder “review_images”, which contains all review images.
- (b) Image folder “cleaned_review_images”. As explained in Section 3, review images can also contain irrelevant objects or infor-

mation. So we apply the object detection model to detect the corresponding object and save the detected image patch as the cleaned review image.

- (c) Review information JSON file named “bestbuy_reviews.json”, which contains review text, review image name and review attributes.
 - i. header: Each review text contains one header and one body
 - ii. body: Each review text contains one header and one body
 - iii. mention: The entity mention shown in the review
 - iv. review_image_path: filename of the corresponding review image
 - v. review_image_url: The link to the corresponding BestBuy image
 - vi. predicted_attribute: Review attributes predicted by our attribute extractors
 - vii. gold_attribute: Annotated review attributes for the *Test* Set. For the *Train* and *Dev* sets, we clean the predicted_attribute to obtain gold_attribute by removing the attributes that do not match with the attributes of the gold entity product.
 - viii. review_id: Unique ID for the review
 - ix. fused_candidate_list: Entity IDs for Top-10 candidate entities
 - x. gold_entity_info
 - A. id: Entity ID for the gold entity
 - B. product_name: Entity name for the gold entity
 - C. product_category: Entity category for the gold entity