Poisoning attempts on Multimodal Entity Linking model

Sam Maley
smaley@ufl.edu

Malvika Jadhav
jadhav.m@ufl.edu

December 11, 2024

Abstract

Multimodal Entity Linking (MEL) tasks aim to match entities with the right types of data across different formats, such as text, images, and videos, in large datasets. The goal is to connect mentions of these entities to the correct entries in a structured knowledge base. However, data poisoning attacks, where attackers deliberately inject misleading or malicious data into the training set, can significantly compromise the performance of these systems. By manipulating the data, these attacks can cause the algorithms to make incorrect associations, misinterpret entities, or fail to link them properly. By intentionally introducing corrupted data into the training datasets, we aim to assess the resilience and robustness of MEL models in our project. Our study includes the replication of baseline models, such as Generative Multimodal Entity Linking (GEMEL), and the evaluation of their performance under various poisoning strategies. Our findings suggest that contextual infilling can negatively impact the performance of a Multimodal Entity Linking model that uses text, even when images are present in the knowledge base to enhance entity linking.

Keywords: *Multimodal Entity Linking; Adversarial Attacks; Data Poisoning.*

1 Introduction

Multimodal Entity Linking (MEL) task involves associating entities mentioned in documents with the correct entries in a structured knowledge base, using multiple data formats such as text, images, and videos. By integrating these modalities, MEL systems aim to enhance the accuracy and comprehensiveness of entity linking, enabling more robust content understanding. This capability is crucial for various applications, including information retrieval, question answering, and knowledge graph construction, where precise and contextually relevant entity identification is essential.

Despite the significant potential of MEL systems, they encounter a variety of challenges that can hinder their performance. These challenges include handling ambiguous entity references, ensuring accurate cross-modal alignment like different images pointing to the same entity, polysemous textual entities and more. One of the key areas we address in our project is the impact of contextual ambiguity. Although techniques like contextual infilling or paraphrasing are generally used as aids in the task of entity linking by providing additional context, its effect on MEL models has not been extensively studied. In our experiments, we focus on understanding how these techniques influence the performance of MEL systems, particularly in the context of text-based models that also utilize other modalities, such as images.

Data poisoning is an adversarial technique in which malicious actors intentionally inject misleading or corrupted data into the training set of a machine learning model. In the context of MEL, this can involve altering text, images, or other modalities within the training data to alter the model's learning process. The goal of data poisoning is to degrade the performance of the model, causing it to make incorrect associations or misinterpret entity relationships. Such attacks can be particularly damaging to MEL systems, as the integrity of both the textual and visual data is essential to accurately link entities across modalities. By exploring how data poisoning affects MEL models, our aim is to highlight potential vulnerabilities in the models and understand how they can be mitigated to ensure more reliable and resilient performance.

The aim of our project is to evaluate the robustness of MEL models under various conditions. To achieve this, we conducted a series of experiments that evaluated the impact of different factors, such as the ablation of specific modules of baseline architecture, the inclusion of entity mentions outside the knowledge base, and the introduction of data poisoning attacks. These experiments are designed to test the resilience of MEL models, offering insight into their ability to maintain accuracy and reliability in challenging scenarios.

2 Methodology

2.1 Datasets

We conducted our experiments on two MEL datasets, WikiDiverse Wang et al. (2022) and WikiMEL Luo et al. (2023). **WikiDiverse** is a high-quality, human-annotated Multimodal Entity Linking (MEL) dataset that focuses on diversified contextual topics and entity types. It is derived from Wikinews, with Wikipedia serving as the corresponding knowledge base. It consists of 7,824 image-text pairs and 16,327 mentions, with an average text length of 10.2 words and an average of 2.1 mentions per instance. For our experiments we have split the dataset into training, validation, and test sets in an 8:1:1 ratio. **WikiMEL** is a large, human-verified Multimodal Entity Linking (MEL) dataset extracted from Wikidata and Wikipedia. WikiMEL contains 22,136 image-text pairs and 25,846 mentions, with an average text length of 8.2 words and an average of 1.2 mentions per instance. We split the dataset into training, validation, and test sets in a 7:1:2 ratio.

2.2 Baseline Architecture: Generative Multimodal Entity Linking (GEMEL)

GEMEL framework handles the task of multimodal entity linking by leveraging both visual and textual modalities. GEMEL takes multimodal mention contexts and utilizes several in-context demonstrations to directly generate the target entity names.

Given the inherent incapacity of LLMs to directly process multimodal information, feature alignment is crucial. Initially, image features are extracted using a pre-trained vision encoder. These image features are then projected into the textual embedding space via a lightweight feature mapper. The resultant features are introduced into the LLM as a visual prefix, enabling the LLM to process visual information effectively. The process involves keeping the vision encoder's weights frozen and training the feature mapper to convert these visual embeddings into a sequence of embeddings that share the same hidden di-

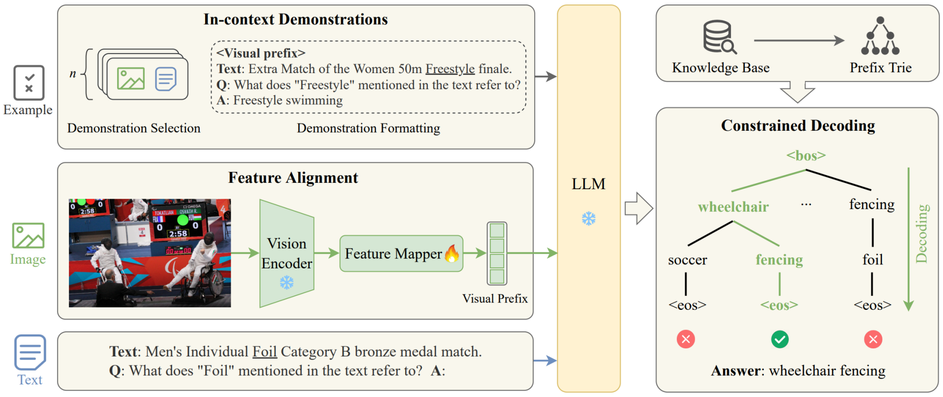
In-context Demonstrations
Example
n
<Visual prefix>
Text: Extra Match of the Women 50m Freestyle finale.
Q: What does "Freestyle" mentioned in the text refer to?
A: Freestyle swimming
Demonstration Selection
Demonstration Formatting
Feature Alignment
Image
Vision Encoder
Feature Mapper
Visual Prefix
Text
Text: Men's Individual Foil Category B bronze medal match.
Q: What does "Foil" mentioned in the text refer to? A:
LLM
Knowledge Base
Prefix Trie
Constrained Decoding
<bos>
wheelchair
...
fencing
soccer
fencing
foil
<eos>
<eos>
<eos>
Decoding
Answer: wheelchair fencing

Figure 1: Baseline Architecture: GEMEL

mensionality as the LLM's text embeddings.

The visual prefix obtained from the feature alignment module is concatenated with text embeddings and fed into the LLM. To enhance the LLM's comprehension of the MEL task, in-context learning (ICL) is employed. Demonstrations are constructed which include the image and textual context for each mention, alongside a question and the corresponding entity name as the answer. Several methods are considered for selecting these demonstrations, including random selection, BM25 for sparse retrieval, and SimCSE for dense retrieval based on semantic matching.

At test time, constrained decoding is employed to ensure the generated entity names are valid. This strategy uses Constrained Beam Search, which confines the generation process to valid identifiers by employing a prefix trie. The trie is built by tokenizing all entity names in the KB, ensuring that the output from the LLM is constrained to valid entity names, matching specific entities in the knowledge base. Through the combination of feature alignment and in-context learning, GEMEL provides a robust framework for multimodal entity linking, ensuring the generation of precise and valid entity names while maintaining the integrity of visual and textual data integration.

2.3 Experiments

Infrastructure for experiments. For all experiments, we used the University of Florida's computing cluster, HiperGator. A single A100 GPU was used during training, and after poisoning each dataset, new sentence embeddings were generated using Princeton SimCSE, Gao, Yao, and Chen (2021).

Prefix Tree Ablation. GEMEL was among the first approaches to Multimodal Entity Linking (MEL) that utilized a generative model Shi, Xu, Hu, and Zhang (2024). Since MEL operates within a knowledge base, GEMEL employs constrained decoding to efficiently navigate the space of valid entities Cao, Izacard, Riedel, and Petroni (2021). It leverages a prefix tree containing the valid entity names from the knowledge base. To evaluate how GEMEL handles scenarios outside the knowledge base, we tested the model's accuracy without using constrained decoding.

Out-of-Knowledge Base Entities. We observed that GEMEL performed reasonably well even without using the prefix tree. To evaluate its accuracy on entities outside of its knowledge base, we conducted experiments using the WikiMEL and WikiDiverse datasets. The model was trained on the unique entities from one dataset and then tested on the unique entities from the other. The results were calculated as follows:

$$\text{Unique\_entities}_1 = \text{Entities}_1 - \text{Union}(\text{Entities}_1, \text{Entities}_2)$$
$$\text{Unique\_entities}_2 = \text{Entities}_2 - \text{Union}(\text{Entities}_1, \text{Entities}_2)$$

Popular vs. Unpopular Entities. To further investigate, we conducted a preliminary experiment to analyze the impact of entity frequency on accuracy. The testing dataset was divided into two subsets: one for popular entities (those appearing more than three times in the training set) and another for unpopular entities (those appearing three times or fewer). The unique entities were created as follows:

	Popular	**Unpopular**
WikiDiverse	586	984
WikiMEL	564	4605

Table 1: Popular & Unpopular Entity Count

Data poisoning. We focused our experiments on the changes in textual modality for the scope of this project. In our study we have considered two main approaches to generate adversarial examples using data poisoning. The first approach consists of using augumentation models from Textattack framework that focus on randomized alterations to the entity mentions using different strategies like word replacement or word deletion. The second approach uses the ContextuaLized AdversaRial Example (**CLARE**) generation model to perform contextual infilling that is a process where missing or altered parts of a text are replaced or merged in a way that is context-aware, fluent, and grammatically correct. For the data poisoning experiments we have used the WikiDiverse

1. **Text Attack**:To understand how GEMEL would be affected by data poisoning that keeps the meaning of the data the same, we explored TextAttack Morris et al. (2020). TextAttack is a framework for poisoning NLP data. For our purposes, we explored three different textual poisoning techniques Each one is listed below, and was chosen specifically for sentence paraphrasing. The ”Embedding” attack is based on the sentence embedding, and aims to alter the text so that the sentence embedding stays roughly unchanged. The ”WordNet” augmentation takes a random word and replaces it with a word with a similar embedding. ”RandomSwap” simply swaps and

deletes characters at random. Additionally, a data poisoning rate of 40% was chosen for this experiment.

2. **CLARE** With the aim to study the impact of contextual infilling we used the CLARE augmentation Li et al. (2020) model on our MEL task input dataset to generate adversarial examples. CLARE generates adversarial examples by applying a sequence of contextualized perturbations to the input through a mask-then-infill procedure using a pre-trained masked language model (MLM). The perturbations consist of three actions: Replace, Insert, and Merge. In the Replace action, a token at a specific position is masked and replaced with a candidate token selected based on MLM probability, similarity to the original input, and its ability to confuse the victim model. The Insert action adds a mask after a token, increasing the sequence length by one, while Merge masks a bigram and fills it with a single token, reducing the sequence length by one. For both Insert and Merge, the replacement token is selected in the same manner as Replace, based on MLM scoring and similarity. To generate an adversarial example, CLARE constructs and scores the perturbations for each position in parallel, ranks them, and iteratively applies the highest-scoring action. The process stops once an adversarial example is found or a limit of actions is reached. We have used conducted our experiments with 10%, 20%, 30% and 40% poisoning rates.

Augmentation	**Output**
Original	President Trump holds a Bible in front of [START_ENT] St. John's Episcopal Church [END_ENT].
Embedding	Chairs Trump holds a Bible in front of [START_ENT] St. John's Episcopal Church [END_ENT].
WordNet	President ruff make a Bible in look of [START_ENT] St. John's Episcopal Church [END_ENT].
RandomSwap	Prrsident Truml hopds a Bibke in front of [START_ENT] St. John's Episcopal Church [END_ENT].

Table 2: TextAttack Augmentations and their corresponding outputs.

3 Results

3.1 Preliminary Experimentations:

3.1.1 Prefix Tree Ablation

There was a small decrease in accuracy from removing the prefix tree. This initial result showed that GEMEL may have the capacity to generalize to out-of-knowledge base entities.

Augmentation	Output
Original	Former US President George W. Bush and Prime Minister Manmohan Singh exchange handshakes on March 2, 2006, at the Hyderabad House in [START_ENT] New Delhi [END_ENT].
CLARE	Former US President George W. Bush and Prime Minister Manmohan Singh awkwardly exchange handshakes on March 2, 2006, at the Hyderabad Institute in [START_ENT] New Delhi [END_ENT].
Original	Al Franken just before addressing the 2008 Olmsted County [START_ENT] Democratic-Farmer-Labor Party [END_ENT] Convention in Rochester, Minnesota.
CLARE	Al Franken just before addressing the Voters in 2008 Olmsted County [START_ENT] Democratic-Farmer-Labor Party [END_ENT] Convention in Rochester, Minnesota.

Table 3: CLARE Augmentations examples.

	Top-1 Accuracy (%)	
	With Prefix Tree	**Without Prefix Tree**
WikiDiverse	82.4	77.2
WikiMEL	75.5	72.6

Table 4: Prefix Tree Ablation

3.1.2 Out-of-Knowledge Base Entities

As shown in Table 5, the model's accuracy dropped significantly when tested on out-of-knowledge base entities. Interestingly, the accuracy improved when the prefix tree was applied to these entities. The reason for this behavior is unclear, and further investigation into the GEMEL implementation by its authors is needed to provide a thorough explanation. Most importantly, however, was that GEMEL demonstrated a capacity to generate the correct answer on out-of-knowledge base entities.

	Top-1 Accuracy (%)		
Checkpoint	**Baseline**	**On Unseen Entities**	
		(without prefix tree)	**(with prefix tree)**
WikiDiverse	82.4	56.1	58.0
WikiMEL	75.5	50.3	60.7

Table 5: Out-of-Knowledge Base Performance

3.1.3 Popular vs. Unpopular Entities

We found that accuracy was significantly higher for popular entities compared to unpopular ones, whose accuracy was closer to the baseline. This discrepancy may be explained by the in-context learning examples drawn from the training set. Since popular entities have more examples available, the model is better equipped to select the correct response for them.

	Top-1 Accuracy (%)		
Dataset	**Baseline**	**Popular Entities (3 appearances in training)**	**Unpopular Entities**
WikiDiverse	82.4	92.4	81.2
WikiMEL	75.5	93.0	74.7

Table 6: Popular & Unpopular Entity Performance

3.2 Data Poisoning Experiments:

3.2.1 TextAttack

TextAttack showed minimal difference in accuracy. The methods chosen here were based on their ability to paraphrase the training data, as such, the embedding of the data stayed roughly the same, allowing for the model to select the correct response.

	Top-1 Accuracy (%)			
	Baseline	**Embedding**	**WordNet**	**RandomSwap**
WikiDiverse	82.4	81.2	80.4	79.2

Table 7: TextAttack (paraphrasing) Results (40% poisoning

3.2.2 CLARE

CLARE augumentation impacted the MEL task accuracy by 7% with only 10% poisoning rate. It is important to note that our generated adversarial examples are visibly similar to original data, even to a human annotator as we have been careful about using a method that preserves the original context of entity mentions.

	Top-1 Accuracy (%)				
	Baseline	**10%**	**20%**	**30%**	**40%**
WikiDiverse	82.4	75.6	75.5	72.2	71.9

Table 8: CLARE Results by Poisoning Percentage

4 Discussion

This study highlights how generative Multimodal Entity Linking (MEL) models are vulnerable to data poisoning, which can reduce their accuracy and reliability. Through a series of experiments, we found that small changes to the input data, such as contextual modifications, can confuse the model and lead to incorrect entity associations. This shows that even subtle changes, while preserving the overall meaning of the data, can still disrupt the model's performance.

One of the key findings is that MEL models are fairly resistant to poisoning techniques that only slightly alter the text, such as paraphrasing. These changes had minimal impact on the model's accuracy because the sentence embeddings remained largely unchanged. However, when the data was poisoned more aggressively, especially in ways that involved manipulating entities or causing misalignments between different data types (text and images), the model's performance dropped significantly.

We also observed that the model struggles with entities that are less common or not included in the training data. Although GEMEL can generalize to some degree and still make reasonable predictions for unseen entities, accuracy drops when the model encounters unfamiliar or rare entities. This highlights a limitation in MEL systems: they rely on having sufficient context and references in the training data to accurately identify entities.

A particularly concerning result was the impact of contextual infilling, which refers to the process of replacing or merging parts of a text in a way that is context-aware, fluent, and grammatically correct. When done maliciously, this technique can introduce plausible but incorrect information, leading the model to make wrong associations. For example, addition of an adjective to change "Manmohan Singh exchange handshakes on March 2" to "Manmohan Singh awkwardly exchange handshakes on March 2" still seems contextually valid, but the model might link it to the wrong entity. This type of attack shows how small changes to the context can confuse MEL models.

To improve MEL models' resistance to such attacks, we suggest incorporating Multi-View Consistency Training that is incorporating multiple mentions of an entity in various contexts and perspectives in the training process. Another defense could be Context Mapping which refers to including mention labels like context types, domains and sentiment during the entity linking process. This analysis will detect contextually incongruent changes introduced by poisoning attacks on a knowledge base. This approach might be most suited for popular entities.. These techniques would help the model detect when an entity has been altered or misrepresented, making it less likely to link incorrect entities. Additionally, training the model with adversarial data or using data augmentation during training could help it learn to handle poisoned inputs more effectively.

5 Conclusion

In this study, we demonstrated the vulnerability of generative Multimodal Entity Linking (MEL) models to data poisoning attacks, which can lead to significant performance issues.

While minor changes that preserve the meaning of the text have a limited impact, more disruptive attacks, like contextual infilling and cross-modal manipulations, can severely affect the accuracy of these models.

These findings highlight the need for more robust MEL systems that can withstand adversarial conditions. Future research should focus on improving the model's ability to handle poisoned data, such as enhancing generalization capabilities, incorporating adversarial training, and adding stronger validation techniques. Additionally, addressing challenges related to rare or unseen entities, and integrating advanced defenses like Multi-View Consistency Training and Context Mapping, will be important for building more reliable MEL systems.

By understanding these vulnerabilities and developing better defenses, we can make MEL systems more resilient and trustworthy for practical applications, such as in knowledge graph construction, information retrieval, and other areas where accurate entity linking is critical.

References

Cao, N. D., Izacard, G., Riedel, S., & Petroni, F. (2021). Autoregressive entity retrieval. In *International conference on learning representations.* Retrieved from `https://openreview.net/forum?id=5k8F6UU39V`

Gao, T., Yao, X., & Chen, D. (2021). SimCSE: Simple contrastive learning of sentence embeddings. In *Empirical methods in natural language processing (emnlp).*

Li, D., Zhang, Y., Peng, H., Chen, L., Brockett, C., Sun, M.-T., & Dolan, B. (2020). Contextualized perturbation for textual adversarial attack. *arXiv preprint arXiv:2009.07502.*

Luo, P., Xu, T., Wu, S., Zhu, C., Xu, L., & Chen, E. (2023). Multi-grained multimodal interaction network for entity linking. In *Proceedings of the 29th acm sigkdd conference on knowledge discovery and data mining* (pp. 1583–1594).

Morris, J., Lifland, E., Yoo, J. Y., Grigsby, J., Jin, D., & Qi, Y. (2020). Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations* (pp. 119–126).

Shi, S., Xu, Z., Hu, B., & Zhang, M. (2024). *Generative multimodal entity linking.* Retrieved from `https://arxiv.org/abs/2306.12725`

Wang, X., Tian, J., Gui, M., Li, Z., Wang, R., Yan, M., ... Xiao, Y. (2022). Wikidiverse: a multimodal entity linking dataset with diversified contextual topics and entity types. *arXiv preprint arXiv:2204.06347.*