OpenMEL: Unsupervised Multimodal Entity Linking Using Noise-Free Expanded Queries and Global Coherence

Xinyi Zhu
The Hong Kong University of Science and Technology (Guangzhou), Data Science and Analytics
xzhu683@connect.hkust-gz.edu.cn

Yongqi Zhang[*]
The Hong Kong University of Science and Technology (Guangzhou), Data Science and Analytics
yongqizhang@hkust-gz.edu.cn

Lei Chen
The Hong Kong University of Science and Technology (Guangzhou), Data Science and Analytics
leichen@cse.ust.hk

ABSTRACT

Multimodal Entity Linking (MEL), which involves disambiguating a mention composed of multimodal inputs to a multimodal knowledge base (KB), has gained increasing attention. Although existing MEL approaches using supervised learning show promising performance, they depend heavily on large-scale labeled training data, which is expensive to obtain for each new scenario. Unsupervised learning MEL methods, on the other hand, typically consist of two main steps. In the first multimodal data encoding step, these methods either assume that the multimodal data inputs are of high quality or attempt to filter out the noisy modality. In the second entity ranking step, they employ a bipartite graph to model the relationships only between mentions and entities. However, unsupervised methods face challenges in both steps. In the first step, data quality issues arise, including limited context in textual inputs and noise in the corresponding images. Moreover, in the second step, the bipartite graph fails to capture coherence between highly correlated entities within the KB, which offers clues on shared domains among entities. This limitation hinders effective retrieval of the target entity. To address these issues, we propose a novel unsupervised learning framework, OpenMEL, for solving the MEL task. We enhance the textual modality contextual information by incorporating full context comprehension and general knowledge, and generates three levels of visual inputs for further adaptive selection to handle noise. To capture global entity coherence, we construct a tree cover structure, defining it as a maximum spanning tree with bounded nodes to meet the MEL objective. We then introduce a greedy algorithm with theoretical guarantees to solve this problem. Experimental results on three public benchmark datasets show that OpenMEL outperforms various state-of-the-art baselines.

PVLDB Reference Format:
Xinyi Zhu, Yongqi Zhang, and Lei Chen. OpenMEL: Unsupervised Multimodal Entity Linking Using Noise-Free Expanded Queries and Global Coherence. PVLDB, 18(8): 2454 - 2467, 2025.
doi:10.14778/3742728.3742740

PVLDB Artifact Availability:
The source code, data, and/or other artifacts have been made available at https://github.com/DAISYzxy/OpenMEL.

1 INTRODUCTION

Over the decades, the task of text-based entity linking has been widely studied [7, 8, 12, 15, 18, 24] to organize knowledge from natural language documents into the structural format employed by Knowledge Bases (KBs). This task supports a broad range of applications, including question answering [12, 32, 42, 49], KB population [10, 22], and information retrieval [6, 19]. Recently, with the surge of multimodal information such as images along with texts [1, 31, 53], studies have increasingly focused on enhancing text-based entity linking with visual information, leading to the development of Multimodal Entity Linking (MEL) [1, 27, 31, 51–53, 56, 58, 59]. The MEL task is an entity linking problem that aims to link a mention (consisting of textual context along with a related image) to the most similar entity (represented by its corresponding image and textual context) in a given KB [1, 27, 31, 51–53, 56, 58, 59]. By leveraging multimodal information, MEL makes it easier and more accurate to resolve ambiguous cases that are challenging in traditional text-based entity linking. For example, the individual profile is helpful in distinguishing whether "Michael Jordan" refers to the computer scientist or the basketball player.

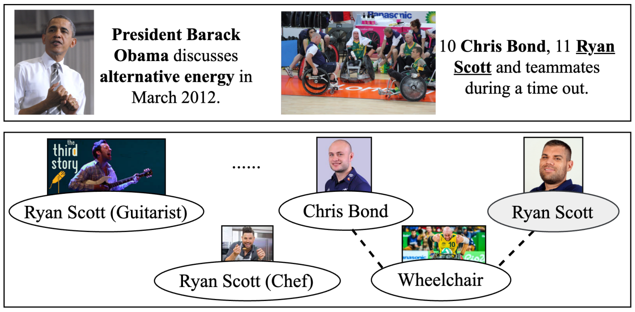

President Barack Obama discusses alternative energy in March 2012.
10 Chris Bond, 11 Ryan Scott and teammates during a time out.
......
Ryan Scott (Guitarist)
Chris Bond
Ryan Scott
Ryan Scott (Chef)
Wheelchair

Figure 1: Examples of multimodal entity linking. Upper: Two multimodal samples with four mentions, highlighted in bold, pending for linking. Lower: Link an example mention "Ryan Scott" to the corresponding entity in KB.

Most existing works [1, 27, 31, 51, 52, 56] approach the MEL task through supervised learning, which necessitates a large amount of labeled data—an average of 15, 978 labeled instances across three benchmark datasets [51, 53]. However, preparing such a large labeled dataset is resource-intensive. Consequently, some studies evaluate existing methods in low-resource settings, using 10% and 20% of the training data. In these scenarios, the accuracy on the WikiDiverse dataset [53] drops by an average of 24.8% and 13.9%,

[*] Corresponding Author

This work is licensed under the Creative Commons BY-NC-ND 4.0 International License. Visit https://creativecommons.org/licenses/by-nc-nd/4.0/ to view a copy of this license. For any use beyond those covered by this license, obtain permission by emailing info@vldb.org. Copyright is held by the owner/author(s). Publication rights licensed to the VLDB Endowment.

Proceedings of the VLDB Endowment, Vol. 18, No. 8 ISSN 2150-8097.
doi:10.14778/3742728.3742740

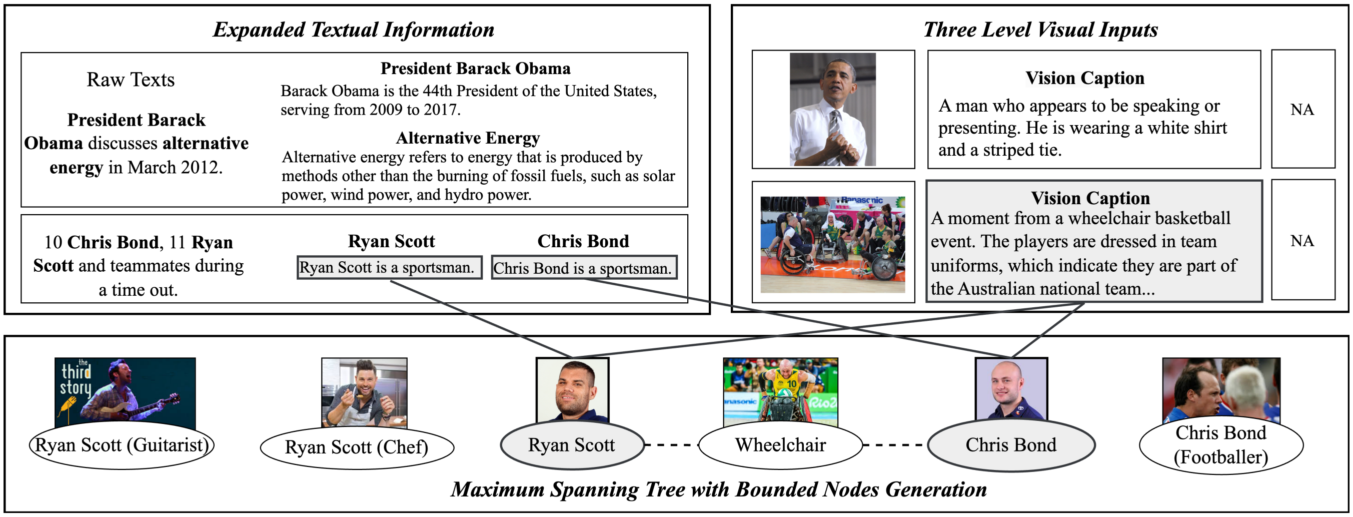

Expanded Textual Information
Raw Texts
President Barack Obama discusses alternative energy in March 2012.
President Barack Obama
Barack Obama is the 44th President of the United States, serving from 2009 to 2017.
Alternative Energy
Alternative energy refers to energy that is produced by methods other than the burning of fossil fuels, such as solar power, wind power, and hydro power.
10 Chris Bond, 11 Ryan Scott and teammates during a time out.
Ryan Scott
Ryan Scott is a sportsman.
Chris Bond
Chris Bond is a sportsman.
Three Level Visual Inputs
Vision Caption
A man who appears to be speaking or presenting. He is wearing a white shirt and a striped tie.
NA
Vision Caption
A moment from a wheelchair basketball event. The players are dressed in team uniforms, which indicate they are part of the Australian national team...
NA
Ryan Scott (Guitarist)
Ryan Scott (Chef)
Ryan Scott
Wheelchair
Chris Bond
Chris Bond (Footballer)
Maximum Spanning Tree with Bounded Nodes Generation

Figure 2: The OpenMEL framework operates as follows: given a multimodal mention context, OpenMEL first optimizes the multimodal inputs by expanding the contextual information in the textual data and generating three levels of visual information from the visual data. It then constructs a maximum spanning tree with bounded nodes for each mention, linking them to entities in the KB while ensuring global coherence throughout the unsupervised entity ranking process.

respectively, across four commonly compared supervised MEL models [27]. This significant decline highlights the heavy reliance of supervised models on large-scale training data, while preparing such datasets for every new scenario is costly. On the other hand, unsupervised learning typically reduce annotation costs by encoding multimodal inputs into feature representations and ranking entities by their semantic similarity between the mention's features and those of the entities, selecting the entity with the highest similarity as the linking target [27, 45, 53, 56]. However, there are still challenges within the two main steps of unsupervised learning—multimodal data encoding and entity ranking—that need to be addressed to develop a high-performing unsupervised MEL method.

The first challenge is the **multimodal data quality** problem. Without labeled data, unsupervised MEL methods either make the strong assumption that multimodal data is of high quality for direct encoding and comparison [27, 45, 53, 56], or filter out noisy modal data based on the similarity of multimodal data features [58, 59], which risks losing valuable information from the discarded modality [56]. In real-world scenario, the multimodal data inputs has the following quality problem. For the textual input, mentions often lack sufficient context, making it difficult to resolve ambiguities, such as polysemy (the same name can refer to different entities) [17, 31, 53] and mention diversity ("J-Lo" and "Jennifer Lopez" refer to the same person) [17, 27, 31, 53]. For instance, there exists polysemy problem (several distinct "Ryan Scott" entities in the KB) in the second sample as shown in Figure 1. For the visual inputs, there is varying relevance between the text and its accompanying image, ranging from strongly correlated (e.g., President Barack Obama with the image) to moderately correlated (e.g., Chris Bond and Ryan Scott with the image) or even irrelevant (e.g., alternative energy with the image), as illustrated in Figure 1. Thus, effectively addressing the limited context problem and managing noise from moderately correlated or irrelevant images remains a significant challenge.

The second challenge is the **entity global coherence** problem. Existing unsupervised methods typically use a bipartite graph [17, 34] in the entity ranking step, comprising a mention set and an entity set, with mention-to-entity relationships represented as edges between the two sets. The entity with the edge of highest similarity to the mention is then selected as the target [27, 45, 53, 56]. However, this structure fails to capture global coherence among entities, as it excludes entity-entity edges within the entity set. Global coherence within the entities is crucial for effective entity linking. As illustrated in Figure 1, the co-occurrence of "Ryan Scott" and "Chris Bond" associating with "Wheelchair" in the KB implies a shared domain between them, providing a clue that mentions of "Ryan Scott" and "Chris Bond" in the second sample are likely linked to their athlete identities rather than other potential matches, such as the chef or guitarist. Thus, it is essential for us to design an appropriate data structure to capture both mention-to-entity relationships and entity coherence within the KB [8, 24]. Moreover, to effectively achieve the objectives of the MEL task, framing a problem within the proposed data structure remains an ongoing challenge.

In this paper, we address the aforementioned challenges by proposing an unsupervised MEL framework on open domains named OpenMEL (illustrated in Figure 2). Leveraging multimodal data for a given mention, we employ human-in-the-loop contextual prompts with instructions for large language models (LLMs) to understand the full context and enhance queries with general knowledge. To address potential visual noise, we process it into three levels of visual inputs. Subsequently, we construct a tree cover structure rooted at each given mention and its entities in the KB to capture entity global coherence. To meet the MEL problem objective, we define a maximum spanning tree problem with bounded nodes. To summarize, in this paper, we make the following contributions.

- We propose an unsupervised OpenMEL framework to effectively tackle the MEL task by improving multimodal data quality during the multimodal data encoding step, and resolving entity global coherence in the entity ranking step.
- To improve the multimodal data quality, we design a human-in-the-loop contextual prompts module for LLMs to enhance textual inputs and process visual inputs into three levels of information for adaptive selection.
- To resolve entity global coherence, we construct a tree cover structure capturing both mention-entity and entity-entity edges. We then formulate it as a maximum spanning tree with bounded nodes problem, which is NP-hard but can be solved by our proposed greedy algorithm under theoretical guarantees, to meet the MEL problem objective.
- Our extensive experiments demonstrate that the proposed OpenMEL framework outperforms many state-of-the-art methods, including both supervised and unsupervised approaches, on three real-world benchmark datasets. Additionally, the ablation study, parameter sensitivity analysis, prompting comparison, and scalability evaluation further validate the robustness of OpenMEL.

2 PRELIMINARY

In this part, we introduce some important definitions and the problem setup used in this paper.

Definition 1 (Knowledge Base (KB)). *A KB is a collection of facts representing real-world concepts and events stored as a triple (e_1, r, e_2). e_1, e_2 are two entities and r represents the relation between them. Specifically, we denote the set of all entities in the KB as E.*

Definition 2 (Multimodal Input). *Each entity e_i is characterized by the corresponding visual context (correlated image) V_{e_i} and textual context (textual spans around the entity) T_{e_i}. Each mention m is also represented by its associated textual context T_m and visual context V_m.*

Definition 3 (Mention set). *A mention m refers to a span of text, typically a noun phrase, in the given text T_m that can be linked to an entity in the KB. A dataset $\mathcal{D}$ contains a set of mentions, $M = \{m_1, m_2, \dots\}$, which is called the mention set.*

For example, in Figure 2, two samples contain mentions of “President Barack Obama”, “alternative energy”, “Ryan Scott” and “Chris Bond” in the mention set.

Definition 4 (Candidate Entity Set). *For each mention m_i in the mention set, there are multiple candidate entities $\{e_i^1, e_i^2, \dots\} \subset E$ for that m_i. We then define the union of all candidate entities for every mention in the mention set as the candidate entity set $\mathcal{E}$ where $\mathcal{E} \subset E$.*

As shown in Figure 2, the candidate entities for the mention “Chris Bond” are {Chris Bond (footballer), Chris Bond (Wheelchair), ...}, and for the mention “Ryan Scott”, the candidates are {Ryan Scott (Wheelchair), Ryan Scott (Chef), ...}. The candidate set $\mathcal{E}$ is the union of candidate entities for all mentions in the mention set.

We then formally present the problem that we aim to solve.

Problem 1 (Open Multimodal Entity Linking). *Let $\mathcal{E}$ represents the candidate entity set from an existing KB, where each entity $e_i \in \mathcal{E}$ is characterized by multimodal input $e_i(V_{e_i}, T_{e_i})$. Given a mention with multimodal input, our objective is to effectively link m to the most similar entity in $\mathcal{E}$ directly, without any training, as $e^*(m) = \arg\max_{m\sim e_i \in L} \Phi(m(T_m, V_m, \sigma); e_i(T_{e_i}, V_{e_i}))$, where $\Phi(\cdot)$ denotes the similarity score. $m(V_m, T_m, \sigma)$ indicates the mention's multimodal inputs, which are initially affected by limited context in T_m and noise σ in V_m. The set L denotes the edges formed between the mention and some entities in the candidate entity set $\mathcal{E}$, with each edge represented by $m \sim e_i$.*

To effectively address Problem 1, we face two technical challenges as discussed in Section 1: (1) the multimodal data quality issue, where mention's textual input T_m suffers from limited context and visual input V_m contains noise σ, and (2) the need to resolve entity global coherence within the linked edge set L and retrieve the most similar entity in L. Accordingly, OpenMEL is structured into two distinct modules: one module aims to optimize multimodal data quality, transforming the original low-quality multimodal inputs, denoted as $m(T_m, V_m, \sigma)$, into $m(T'_m, \{V_m, V'_m, \emptyset\})$, where T'_m represents the enhanced texts and $\{V_m, V'_m, \emptyset\}$ represents the three-level visual inputs for adaptive selection, as detailed in Section 3. The other module addresses entity global coherence when constructing the linked edge set L and further retrieves the entity with highest similarity via the function $\arg\max \Phi(\cdot)$, as discussed in Section 4.

3 NOISE-FREE EXPANDED QUERIES GENERATION

In this section, we aim to address the multimodal data quality problem in the given low-quality multimodal inputs $m(T_m, V_m, \sigma)$ by handling limited context in T_m and noise σ in V_m. Specifically, the original visual input V_m is processed into three levels of information $\{V_m, V'_m, \emptyset\}$ for further adaptive selection. Finally, we propose a human-in-the-loop greedy algorithm to help mitigate potential hallucinations.

3.1 Principles

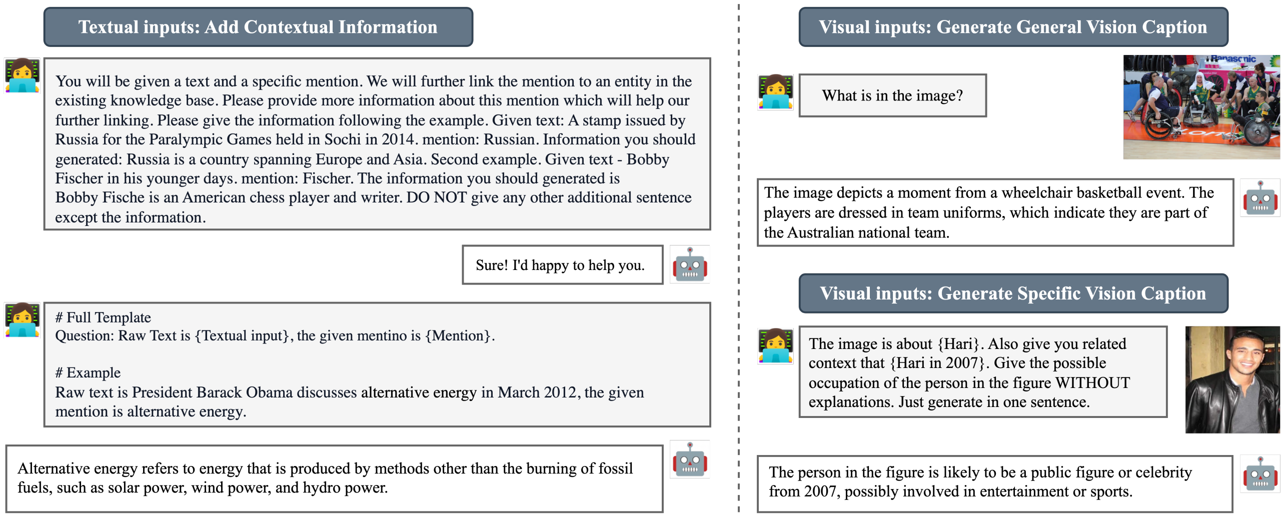
As discussed in Section 1, mention's multimodal input lacks sufficient context thus suffering from polysemy and mention diversity. On the other hand, there is noise within visual inputs such as images that are either irrelevant or relevant but misleading. We give detailed analysis from the perspective of each modality.

- Polysemy and mention diversity issues become more prominent when the textual content is limited. In particular, incorporating noise-free visual information from accompanying images and enriching the textual context with general knowledge can aid in further disambiguation.
- As shown in Figure 1, images can display varying levels of relevance to the mention being linked, ranging from highly correlated to moderately correlated, or even completely irrelevant.

Consequently, our goal is to address the aforementioned multimodal data quality issues to ultimately optimize the mention's multimodal inputs based on our analysis.

3.2 In-context Learning with Instructions

Large models, particularly transformer-based architectures have been shown to effectively capture the contextual nuances of language, which is beneficial in cases where traditional methods may



You will be given a text and a specific mention. We will further link the mention to an entity in the existing knowledge base. Please provide more information about this mention which will help our further linking. Please give the information following the example. Given text: A stamp issued by Russia for the Paralympic Games held in Sochi in 2014. mention: Russian. Information you should generated: Russia is a country spanning Europe and Asia. Second example. Given text - Bobby Fischer in his younger days. mention: Fischer. The information you should generated is Bobby Fische is an American chess player and writer. DO NOT give any other additional sentence except the information.

Sure! I'd happy to help you.

Full Template
Question: Raw Text is {Textual input}, the given mentino is {Mention}.

Example
Raw text is President Barack Obama discusses alternative energy in March 2012, the given mention is alternative energy.

Alternative energy refers to energy that is produced by methods other than the burning of fossil fuels, such as solar power, wind power, and hydro power.

What is in the image?

The image depicts a moment from a wheelchair basketball event. The players are dressed in team uniforms, which indicate they are part of the Australian national team.

The image is about {Hari}. Also give you related context that {Hari in 2007}. Give the possible occupation of the person in the figure WITHOUT explanations. Just generate in one sentence.

The person in the figure is likely to be a public figure or celebrity from 2007, possibly involved in entertainment or sports.

Figure 3: In-depth in-context learning prompts, accompanied by instructions and additional demonstration examples, are utilized to expand textual content and generate the second level visual information to address multimodal data quality problem.

struggle due to a lack of deeper understanding of full context [3, 21]. In this subsection, we use contextual prompts with instructions to guide LLMs in understanding the task and full textual contents, enabling them to generate enhanced text T'_m and process three-level visual information $\{V_m, V'_m, \emptyset\}$ as previously analyzed.

3.2.1 Textual Context Expansion. In the textual modality, we intuitively harness the general knowledge of LLMs and guide them in understanding the MEL task, providing additional information on the mention to be linked.

We illustrate the use of contextual prompts with instructions, further supported by demonstration examples as depicted in Figure 3. Specifically, we craft a targeted question (*i.e.*, Please provide more information about this mention which will help our further linking), alongside contextual instructions (*i.e.*, You will be given a text and a specific mention. We will further link the mention to an entity in the existing knowledge base) for the MEL task. Given the mention's raw text and the name to be linked, LLMs process the input and generate more detailed explanations following our demonstration examples (*i.e.*, Given text: A stamp issued by Russia for the Paralympic Games held in Sochi in 2014. Mention: Russian. Information you should generate: Russia is a country spanning Europe and Asia) using general knowledge. In this way, we enrich the textual inputs T_m with contextual information, resulting in T'_m.

3.2.2 Three Levels of Visual Inputs. Similar to how humans process visual information, we divide and analyze it into three levels, as illustrated in Figure 2. The first level is when the image directly depicts the mention itself, such as a person's profile, which allows humans to compare the two profiles and determine if they refer to the same mention. The second level is when the image is related to the mention, and humans extract keywords from the image to assist in further disambiguation, for example, identifying Ryan Scott as a "wheelchair" athlete based on the image. Finally, the third level is when the image is irrelevant to the mention, such as a profile of "President Barack Obama" for the mention "alternative energy", in which case humans disregard the visual information.

Following this, each image can be processed into three levels of visual information: the raw image V_m, which includes complete visual information as input; the vision caption V'_m, which only describes the main contents of the image; and no visual input $\emptyset$ when the image is irrelevant to the mention. The three levels of visual information further support adaptive selection, helping to filter out noise while retaining as much relevant information as possible. Specifically, the vision caption V'_m is processed using the in-context learning capability of multimodal LLMs, as depicted in Figure 3. The initial prompt is "What's in the image?", to provide an objective description of the image. For specific topics, such as PERSON, we further refine the prompt to help LLMs generate more information about the individual, particularly regarding their occupation. To further address noise—defined as moderate relevance or irrelevance between the accompanying image and the mention—we discuss the adaptive selection of three levels of information in Section 4.1.

3.3 Hallucinations Mitigation

In this subsection, we implement a human-in-the-loop greedy algorithm designed to identify potential hallucinations within the texts generated by LLMs as an option module for the first step.

We face two primary challenges in this task. First, given the substantial volume of data, it is impractical for humans to review all the information. Second, there is a need to establish an evaluation metric and a ground truth to assess the likelihood that a given instance is a hallucination. Compared to raw visual inputs might containing noise, textual inputs have a stronger correlation with the mentions they contain. Therefore, raw texts are considered the ground truth. The evaluation metric we use is normalized cosine similarity ranging from 0 to 1 measuring the semantic similarity, denoted as $\phi(x, y) = (cosine\_similarity(x, y) + 1)/2$, where x, y is any two vectors with the same dimension. We measure the similarity of T_m' and V_m' generated by LLMs with ground truth T_m. Given the constraints on human involvement in the system, we greedily rank instances based on their similarity. We then select the data items with the lowest similarity for manual review and re-annotation.

In conclusion, we optimize the multimodal inputs to enhance contextual information and eliminate noise, transforming them into $m(T_m', \{V_m, V_m', \emptyset\})$ for adaptive selection and tree cover construction as follows.

4 GLOBAL COHERENCE RESOLUTION

In this section, we firstly construct a tree cover to establish mention-to-entity and entity-to-entity edges in L, as defined in Problem 1, to resolve the global coherence of entities. Then, we conduct the entity ranking to realize the MEL task objective by further defining the tree cover structure as the maximum spanning tree with bounded nodes problem and propose an approximation algorithm to solve it.

4.1 Tree Cover Construction

In this subsection,we introduce a tree cover structure rooted at each mention to model our goal of capturing the global coherence of entities within the KB. Specifically, we define it as follows.

Definition 5 (Tree Cover). *The tree cover $(V, L) = ((M, \mathcal{E}), L)$ is a weighted undirected heterogeneous graph, where V represents the set of nodes and L is the set of linked edges. Each node belongs to either the mention set M or the candidate entity set $\mathcal{E}$. The edge between any two nodes is assigned a weight that measures the semantic distance between their multimodal data representations. The root of each tree corresponds to a specific mention awaiting linking.*

The objective of tree cover construction is to discover all entities $e_i \in \mathcal{E}$ that are related to the mention m, while simultaneously considering the global coherence of the entities. For example, as shown in Figure 4a, for each tree rooted at the mention m_1, m_2, we establish weighted edges between nodes to capture their co-occurrence. Specifically, two types of edges are defined as follows.

Definition 6 (Tree Cover Edge Setup). *In the tree cover, edges are structured to satisfy the following conditions:*

- *A mention-entity edge is initially constructed by linking the mention to its K highest weighted (most similar) neighbors in the candidate entity set, using the similarity $\Psi(m(V_m', T_m'), e_i(V_{e_i}, T_{e_i}))$, where $e_i \in \mathcal{E}$.*
- *An entity-entity edge is then connected to their K highest weighted neighbors within the candidate entity set using the similarity function on only textual inputs $\phi(e_i(T_{e_i}), e_j(T_{e_j}))$ since textual inputs have a stronger correlation with the entities they contain.*

The similarity function Ψ is defined as the maximum value from all multimodal similarity scores, as shown in Equation (1), to prevent certain lightweight edges from dominating others. This approach ensures that a lightweight edge, which reflects only a small semantic distance from a local perspective, does not overshadow the broader multimodal correlations captured from a global perspective.

$$\Psi(m(V_1, T_1), e(V_2, T_2)) = \max\{\phi(V_1, V_2), \phi(V_1', T_2), \phi(T_1', V_2), \phi(T_1', T_2)\} \qquad (1)$$

V_1, V_2 denotes the raw visual image, V_1', T_1' are the generated vision caption and expanded textual query through the first step. This approach addresses the challenge of varying relevance between vision and text inputs through adaptive selection. For example, linking the mention "Ryan Scott" to the entity "Ryan Scott (Wheelchair)" in the KB is achieved through this adaptive selection of the maximum similarity, which corresponds to the edge between the mention's vision caption and the entity's text, as shown in Figure 2.

Algorithm 1: Tree Cover Edge Setup

```
Input: Mention set $M$, Candidate entity set $\mathcal{E}$, Number of
       Neighbors $K$
Output: Mention-entity edge index set $I_{me}$ and weight set
        $D_{me}$; Entity-entity edge index set $I_{ee}$ and weight
        set $D_{ee}$
1 foreach $m_t \in M$ do
2     foreach $e_i \in \mathcal{E}$ do
3         Calculate mention-entity edge weight using
            similarity function:
                $\Psi(m(V_{m_t}', T_{m_t}'), e_i(V_{e_i}, T_{e_i}))$
4     Link $m$ to its $K$ highest weighted entities, record their
        indices and weights as $I_{m_t e}$ and $D_{m_t e}$;
5     foreach $e_i \in I_{m_t e}$ do
6         foreach $e_j \in \mathcal{E}$ do
7             Calculate entity-entity edge weight using
                similarity function:
                    $\phi(e_i(T_{e_i}), e_j(T_{e_j}))$
8         Link $e_i$ to its $K$ highest weighted entities, record
            their indices and weights as $I_{e_i e}$ and $D_{e_i e}$;
9 return $I_{me}, D_{me}, I_{ee}, D_{ee}$
```

In conclusion, the tree cover edge setup algorithm is detailed in Algorithm 1. The algorithm first constructs the mention-entity edges (lines 1–4), then establishes the entity-entity edges (lines 5–8). Finally, it returns the edge index set along with the corresponding weight set (line 9).

Time Complexity. The time complexity for the Tree Cover Edge Setup is divided into two components. For each mention, the mention-entity edge setup has a complexity of $O(|\mathcal{E}|)$ for similarity calculation and $O(|\mathcal{E}| \cdot \log|\mathcal{E}|)$ for ranking. Similarly, for entity-entity edge setup, the complexity is $O(K \cdot (|\mathcal{E}| + |\mathcal{E}| \cdot \log|\mathcal{E}|))$. Consequently, the dominating term for the time complexity of Algorithm 1 simplifies to $O(|M| \cdot |\mathcal{E}| \cdot \log|\mathcal{E}|)$.

4.2 Maximum Spanning Tree with Bounded Nodes

To realize the MEL objective in Problem 1 within the constructed tree cover, we define the maximum spanning tree with bounded nodes problem as follows.

Problem 2 (The Maximum Spanning Tree with Bounded Nodes Problem). *Given the pre-constructed tree cover, the target* *mention m for linking, and the number of entities to be retrieved as the selection bound B, our goal is to identify the maximum spanning tree rooted at m. This maximum spanning tree is a weighted, connected, acyclic graph S consisting of B leaf nodes (entities), rooted at the target mention m, that maximizes G(S), as defined in Equation* (2).

$$G(S) = \sum_{e_i \in S} v(e_i), |S| \leq B \tag{2}$$

Specifically, $v(e_i)$ represents the edge weights in the tree rooted at m. As shown in Equation (3), we would like to find the set of entities S^* which can maximize the target value $G(S)$ denoted as Equation (2). For instance, for the tree rooted at the mention "Ryan Scott," as shown in Figure 2, we first add the entity "Ryan Scott" in the KB to S if the weight of the edge between it and the root mention is the highest.

$$S^* = \arg\max_{S} G(S) \tag{3}$$

Notably, we explain the intuition to formulate the MEL problem on the constructed tree cover as a maximum spanning tree with bounded nodes problem as follows.

- In each step, each tree is rooted with a mention m for linking. This ensures that no mention is excluded.
- Each tree, rooted at a specific mention m, is connected to the most similar K entities in the KB. Additionally, each entity is connected to its correlated neighbors in the KB, acting as leaf nodes. This structure ensures the co-occurrence between the mention and all potentially correlated entities within each tree, thereby resolving the global coherence.
- By defining the cost of a tree as the sum of the weights of its included leaf nodes and limiting the total number of leaf nodes to B, where B is the number of entities to be linked, maximizing the cost of the tree equates to linking the mention to the B most similar entities.

Specifically, $B = 1$ denotes linking the root m to the most similar (Top-1) leaf node (entity) in the tree. Then, the solution to Problem 2 satisfies the objective function of OpenMEL denoted as $e^*(m) = \arg\max_{m \sim e_i \in L} \Phi(m(T'_m, V'_m); e_i(T_{e_i}, V_{e_i}))$ in Problem 1. Notably, compared to existing works which commonly employs a bipartite graph structure [17, 34], the coherence between entities is not captured because the bipartite graph property precludes edges within the set of candidate entities. On the other hand, our constructed tree cover captures the relationships between mentions and entities, facilitating co-occurrence for further entity linking as described in Problem 2. For instance, as shown in Figure 2, a bipartite graph structure only establishes similarity between the mention "Chris Bond" and candidate entities such as "Chris Bond (Wheelchair)" or "Chris Bond (Footballer)". However, our tree cover additionally considers the similarity between entities "Chris Bond (Wheelchair)" and related terms "Wheelchair" and "Ryan Scott (Wheelchair)", increasing the likelihood of correctly linking "Chris Bond (Wheelchair)" to the mention.

Theorem 1 (Problem 2 is NP-hard). *The maximum spanning tree with the bounded node problem can be proven to be NP-Hard by a reduction to the Knapsack problem.*

Proof. Given an instance I of the Knapsack problem as follows: with a set of items $\{i_1, i_2, \ldots, i_n\}$, each with a weight w_i and a value v_i, given the knapsack maximum weight capacity W, the goal is to find a subset of items $T \subseteq \{i_1, i_2, \ldots, i_n\}$ such that the total weight of S does not exceed W and the total value is maximized.

We transform it to an instance J of our maximum spanning tree with bounded nodes problem by the following steps:

- Each entity in the candidate entity set $\mathcal{E}$ represents an item e_i from the knapsack problem.
- We set the value of each entity $v(e_i)$ as its edge weight linked to its linked node n in the distance set $D_{(e_i,n)}$.
- The weight capacity W is pre-defined, and we set the weight for each entity to 1.

□

We then introduce how we address Problem 2 as follows.

4.3 The Approximation Algorithm

To solve Problem 2, we propose an approximation algorithm inspired by the methods in [23, 37]. In particular, the approximation algorithm includes an edge pruning strategy and a greedy approach.

The input to the approximation algorithm is the tree cover constructed in Section 4.1, consisting of a set of mention nodes, each rooted to a tree, along with a bound on the number of retrieved entities, B. The output is a maximum spanning tree containing B entities for each mention $m \in M$.

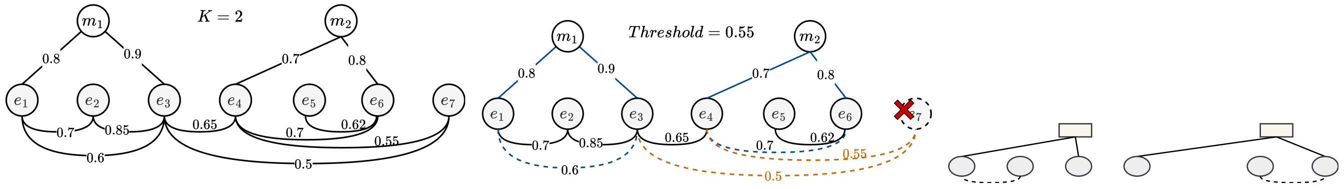
4.3.1 Edge Pruning. Our goal is to prune redundant edges in the tree to reduce the complexity of generating the maximum spanning tree. There are two types of edges that can be pruned. The first is any edge with a weight below a pre-defined threshold after edge weight normalization. For example, as shown in Figure 4b, if the threshold is set to 0.55, edges such as the one between e_3 and e_7 with weights equal to or below this threshold should be pruned.

The second step involves removing cycles associated with the tree root. This means that if there is an edge between e_1 and e_2 while the mention m is already linked to both e_1 and e_2, the edge between e_1 and e_2 can be pruned. Since e_1 and e_2 are already part of the candidate set, removing this edge helps reduce the complexity in the MST generation. For example, the pairs m_1, e_1, e_3 and m_2, e_4, e_6 form cycles, where we prune the edges between e_1 and e_3, as well as between e_4 and e_6.

4.3.2 Maximum Spanning Tree (MST) Generation. To generate a MST with bounded nodes in polynomial time, we propose a greedy algorithm as shown in Algorithm 2. The *start node* is the root mention m. The basic idea is to firstly rank all edges for each node in descending order then greedily select the entity which will not form a cycle and bring the maximum weighted edge $\Delta G(e|S)$ in Equation (4) to its corresponding node in existing set S until S exceeding the selection bound. For example, as shown in Figure 4c, we start from m_1 and first select e_3, as the edge between m_1 and e_3 has the highest similarity score.

$$\Delta G(e|S) = G(S + \{e\}) - G(S) \tag{4}$$

The details are illustrated in Algorithm 2. We begin by initializing a heap to record the pushed edges, starting with the root



(a) Tree Cover Construction. **(b) Edge Pruning.** **(c) Maximum Spanning Tree Generation.**

Figure 4: Running Examples for Entity Global Coherence Resolution. The blue line represents the existence of a circle. The orange line indicates that the similarity of the edge is below the threshold. Dashed lines represent pruned edges. The red entities are the Top-B entities linked to the corresponding mention.

mention. At each step, we pop the highest-weighted edges and check whether the linked entity forms a cycle, in order to avoid redundant selections. Specifically, to detect the cycles, we utilize the Union-Find data structure [23]. Each node initially serves as its own set (i.e., it is its own parent). For every new edge (u, v), we check if u and v are already connected (i.e., they belong to the same set). If they do not belong to the same set, we merge them using the union function and add this edge to the tree.

We further prove that the proposed greedy Algorithm 2 does have a theoretical guarantee on its approximation ratio.

THEOREM 2 (APPROXIMATION RATIO). *The greedy algorithm on the maximum spanning tree with bounded nodes problem can achieve an approximation ratio of* $1 - \frac{1}{e}$.

PROOF. We first prove the following lemma.

LEMMA 1. *The objective of maximum spanning tree with bounded nodes problem as shown in Equation* (2) *is monotone and submodular.*

PROOF. (1) $v(e_i)$ represents the maximum edge weights among the multimodal similarities, as shown in Equation (1), where each component ϕ is positive. Consequently, $v(e_i)$ is positive, and thus, the summation of $v(e_i)$ is monotonically increasing.

(2) We prove submodularity by showing that $G(S+v(e_i))-G(S) = \Delta G(e_i|S) \geq G(T + v(e_i)) - G(T) = \Delta G(e_i|T)$, where $S \subseteq T$ and node $v(e_i) \notin S, T$. This holds because the order in which entities are added depends on their weights. An entity with a higher weight relative to the nodes in the existing set will be added to the result earlier in S than one with a lower weight. □

Since $G(S)$ as defined in Equation (2) is monotone and submodular, according to Hochbaum [20], the approximation ratio of proposed Algorithm 2 is $1 - \frac{1}{e}$. □

Running Example. We present a comprehensive running example of the global coherence resolution module as depicted in Figure 4, which encompasses tree cover construction, edge pruning, and maximum spanning tree generation. Specifically, we first follow Definition 6 to construct a tree cover for each mention such as m_1, m_2 pending linking. Secondly, we prune edges whose weight is below the threshold, as well as the edge which forms a circle with the root mention. For example, there are circles between m_1 and e_1, e_3 as well as m_2 and e_4, e_6. Then, according to the design in Section 4.3.1, edge between e_1, e_3 and edge between e_4, e_6 should be pruned. Thirdly, we conduct Algorithm 2 to retrieve the target entities. As shown in Figure 4c, we set $B = 2$ which requires Top-2 most similar entities to be retrieved.

Algorithm 2: Maximum Spanning Tree for Mention m

```
Input: Mention $m$, mention-entity indices $I_{me}$, distances
       $D_{me}$, entity-entity indices $I_{ee}$, distances $D_{ee}$,
       selection bound $B$
Output: Maximum spanning tree for mention $m$
Initialization:
start_node ← $m$
max_heap $H$ ← $\emptyset$
total_weight $W$ ← 0
results ← $[m]$
edge_list $E$ ← $\emptyset$
$H$.push($D_{me}, m, I_{me}$)
$e_{me_i}$ ← $H$.pop()
results.add($e_{me_i}$)
foreach $e_j \in I_{me}$ do
    $H$.push($D_{ee}[e_j], e_j, I_{me}[e_j]$)
while $|results| < B$ do
    $e_{ab}$ ← $H$.pop() // Pop the next edge (a, b)
    while $a \notin results$ do
        $e_{ab}$ ← $H$.pop()
    if $E$.add($b$) does not form a cycle then
        results.add(b)
return results
```

Time Complexity. The first module requires $O(logK)$ to push and pop an item in the Heap H which we pre-defined the number of entities recorded in I_{me} is K. And the loop to push further entities requires $O(K \cdot \log K)$. The second module is the main loop which iterates B times. Each iteration takes $O(\log K)$ to pop and $O(\alpha(n)) \approx O(1)$ for the cycle detection. Specifically, $\alpha(n)$ is the inverse Ackermann function, which grows very slowly and is nearly constant for all practical purposes [48]. The overall time complexity for the proposed greedy algorithm is $O(|M| \cdot K \cdot \log K)$ for all mentions in the dataset.

OpenMEL Pipeline. The overview is as follows.

- We optimize the multimodal data quality, enhancing textual inputs as T'_m and three-level information as raw image embedding (V_m), image caption (V'_m), and $\emptyset$ by prompts in Figure 3.
- We then construct the edge set L with adaptive selection of visual inputs associated with the target mention, while also considering the entity's global coherence, as outlined in Algorithm 1.
- Finally, we frame Problem 1 of the MEL task as a maximum spanning tree with bounded nodes problem, as defined in Problem 2, which can be solved through Algorithm 2.

Time Complexity. After enhancing data quality through LLMs, OpenMEL requires time complexity of $O(|\mathcal{E}| \cdot \log|\mathcal{E}| + |M| \cdot K \cdot \log K) = O(|\mathcal{E}| \cdot \log|\mathcal{E}|)$ (where $|M|$ and K are constants) for the entity ranking step, which includes constructing the tree cover using Algorithm 1, followed by generating the maximum spanning tree with bounded nodes via Algorithm 2.

In conclusion, to solve Problem 1, the OpenMEL framework follows the procedure of unsupervised learning methods while addressing two key technical challenges during processing multimodal inputs $m(T_m, V_m, \sigma)$ and constructing edges L between mention and entities. The first is improving the quality of multimodal input data during the data encoding step, and the second is resolving entity global coherence within the KB.

5 EXPERIMENTS

In this section, we compare the performance of OpenMEL on the MEL task against the state-of-the-art methods including both supervised and unsupervised ones, on three benchmark datasets. We also conduct ablation studies, analyze parameter sensitivity, evaluate prompting strategies, assess scalability, and conduct data analysis for OpenMEL. Specifically, we intend to investigate the following research questions (RQ):

- **RQ1.** How does the proposed OpenMEL perform compared with various baselines?
- **RQ2**. How do the two proposed modules affect performance?
- **RQ3**. How does OpenMEL's performance change with the parameters?
- **RQ4.** How does the quality of the LLM results influence the outcome?
- **RQ5.** What is the data analysis of OpenMEL in terms of scalability and image relevance?

5.1 Experimental Settings

In this subsection, we present the experimental settings of our work, covering the implementation details of OpenMEL, datasets, baseline comparisons, and evaluation metrics. All the experiments are conducted on a server with 256GB RAM, 2.60GHz 18cores CPU, with CentOS Linux release 7.9.2009 (Core) installed.

5.1.1 Implementation Details. We list the detailed techniques in OpenMEL as follows.

- In the noise-free query expansion generation module, the LLMs employed by OpenMEL are Llama3 [11] for the textual modality and MiniCPM-Llama3-V-2_5 [57] for the visual modality. After multimodal data enhancement, the multimodal data is further processed into vector embeddings through the Vision-and-Language Pre-training (VLP) model CLIP [39].
- For the WikiDiverse, which cover multiple mention topics, we use the general visual prompt "What's in the image?". For datasets focused on a specific topic, such as PERSON, we modify the prompt to guide LLMs to provide more detailed descriptions by leveraging in-context learning related to the individual's occupation, as detailed in Section 3.
- A KB containing large-scale entities makes mention linking highly time-consuming, particularly when images are incorporated. For a fair comparison, the candidate entity set $\mathcal{E}$ is constructed according to prior works [27, 51, 53]. We use Wikidata as the KB and exclude any mentions for which no corresponding entity could be found in Wikidata [27].

Table 1: The statistics of multimodal entity linking datasets. * stands for the average number.

Statistics	WikiMEL	RichpediaMEL	WikiDiverse
# samples	22,136	17,806	7,969
# test	5,169	3,562	2,078
# mention*	1.2	1.8	2.0
# words*	8.2	13.6	10.1
# topic	1	1	7

5.1.2 Datasets. We employ three MEL datasets: WikiMEL, RichpediaMEL [51], and WikiDiverse [53]. 1. **WikiMEL** [51] is sourced from Wikipedia entity pages and comprises over 22,000 multimodal sentences. The topic in WikiMEL is only PERSON. 2. **RichpediaMEL** [51] is derived from a multimodal knowledge base Richpedia [50]. The authors first extract entities from Richpedia, and then enrich this information by obtaining additional multimodal data from Wikidata. Like WikiMEL, RichpediaMEL also features a single mention topic type: PERSON. 3. **WikiDiverse** [53] is a dataset constructed from Wikinews, designed for multimodal tasks such as entity linking or classification. WikiDiverse encompasses 7 topics, including sports, technology, economy, and more, offering diverse content that mirrors real-world scenarios.

We further illustrate the detailed statistics of each multimodal entity linking dataset in Table 1. Specifically, each sample contains multiple mentions to link in the textual input.

5.1.3 Baselines. We compare our method with various competitive baselines, considering two aspects: the learning approach (supervised or unsupervised) and the different modality inputs, including both text and vision.

Supervised learning methods include: 1. **ARNN** [13]: Takes text as inputs and utilizes Attention-RNN to predict associations. 2. **DZMNED** [31]: Employs a concatenated multimodal attention mechanism to fuse visual, textual, and character features together. 3. **JMEL** [1]: Projects visual and textual features into an implicit joint space via fully connected layers. 4. **MEL-HI** [59]: Tries to remove noisy images and mine implicit cues via multiple attention mechanisms. 5. **GHMFC** [51]: Takes gated multimodal fusion and a novel attention mechanism to link multimodal entities. 6. **DWE** [46]: Serves as an advanced baseline enriching entity semantics and obtaining a more comprehensive textual representation. Since not all

Table 2: Performance comparison. C, S, U, M separately stand for Classification, Supervised, Unsupervised Learning methods, and Modality. T and V refer to Textual and Visual modality, respectively. The best result in S and U is highlighted in bold. Baseline ended with * is implemented by us. Others are provided by corresponding papers. "Baseline (Aligned)" represents using the same LLMs enhanced inputs aligned with OpenMEL. "w/o HITL" refers to evaluation conducted without any human-in-the-loop for hallucination mitigation, while other inputs enhanced by LLMs are all tested with 10-shots annotations.

C	M	Methods	WikiMEL			RichpediaMEL			WikiDiverse		
			HIT@1	HIT@5	HIT@10	HIT@1	HIT@5	HIT@10	HIT@1	HIT@5	HIT@10
S	T	ARNN	32.0%	45.8%	56.6%	31.2%	39.3%	45.9%	22.4%	50.5%	68.4%
	T+V	DZMNED	34.7%	53.9%	58.1%	32.4%	43.7%	48.2%	-	39.1%	-
	T+V	JMEL	31.3%	49.4%	57.9%	29.6%	42.3%	46.6%	21.9%	54.5%	69.9%
	T+V	MEL-HI	38.6%	55.1%	65.2%	34.9%	43.1%	50.6%	45.7%	76.5%	88.6%
	T+V	GHMFC	43.6%	64.0%	74.4%	38.7%	50.9%	58.5%	46.0%	77.5%	88.9%
	T+V	DWE	**44.7%**	**65.9%**	**80.8%**	**67.6%**	**97.1%**	**98.6%**	**47.5%**	**81.3%**	**92.0%**
U	T	BERT	31.7%	48.8%	57.8%	31.6%	42.0%	47.6%	22.2%	53.8%	59.8%
	T	BERT (Aligned)*	40.5%	54.1%	61.5%	37.6%	48.2%	50.4%	30.1%	59.9%	63.4%
	T+V	CLIP*	40.7%	56.0%	64.6%	38.1%	54.5%	62.4%	34.4%	59.7%	62.2%
	T+V	CLIP (Aligned)*	45.7%	64.2%	71.3%	41.4%	57.8%	65.3%	41.6%	64.8%	68.5%
	T+V	BM*	33.2%	50.7%	57.5%	45.1%	62.3%	69.9%	28.8%	48.8%	58.1%
	T+V	BM (Aligned)*	57.8%	72.1%	78.9%	54.4%	69.7%	76.8%	52.4%	68.2%	71.7%
	T	OpenMEL	61.9%	71.8%	74.6%	57.8%	66.5%	68.2%	63.0%	73.3%	75.6%
	T+V	OpenMEL (w/o HITL)	69.8%	81.0%	83.4%	65.5%	77.4%	80.4%	67.1%	82.2%	85.2%
	T+V	OpenMEL	**69.8%**	**81.0%**	**83.4%**	**65.6%**	**77.4%**	**80.4%**	**67.1%**	**82.2%**	**85.2%**
	T+V	OpenMEL (GPT-4o)	75.1%	84.3%	85.9%	72.6%	81.1%	83.2%	72.4%	87.2%	90.3%

Table 3: Experimental results of ablation studies on multimodal inputs and the two main modules of OpenMEL. All tests are performed under same testing parameters ($K = 10, f = 0.5$). The best scores are highlighted in bold.

Model	WikiMEL			RichpediaMEL			WikiDiverse		
	HIT@1	HIT@5	HIT@10	HIT@1	HIT@5	HIT@10	HIT@1	HIT@5	HIT@10
OpenMEL	**69.8%**	**81.0%**	**83.4%**	**65.6%**	**77.4%**	**80.4%**	**64.5%**	**76.8%**	**78.1%**
m (T+V) + GCR	63.2%	79.9%	83.1%	55.5%	67.4%	71.4%	40.8%	63.8%	67.1%
NFEQ (T+V) + Top-K	49.7%	65.9%	73.6%	43.7%	59.3%	66.7%	44.6%	67.1%	70.8%
NFEQ (T) + GCR	61.9%	71.8%	74.6%	57.8%	66.5%	68.2%	63.0%	73.3%	75.6%
m (T) + GCR	54.9%	69.9%	73.4%	48.9%	61.0%	64.1%	45.8%	62.2%	66.4%
m (T+V) + Top-K	40.7%	56.0%	64.6%	38.1%	54.5%	62.4%	34.4%	59.7%	62.2%

methods and datasets support textual Wikipedia demonstrations, DWE is also not equipped with it to ensure a fair comparison.

Unsupervised learning methods include: 1. **BERT** [9]: Stacks several layers of transformers to encode each token in the text for further similarity comparison. 2. **CLIP** [39]: Considers both textual and visual features, concatenating multimodal features and further calculating the similarity. 3. **BM**: is a multimodal Bipartite graph Matching (BM) algorithm implemented from M3EL [17].

5.1.4 Evaluation Metrics. During evaluation, we set the parameter B to 1, 5, 10 following the previous works [46, 51, 56, 59] to check whether the Top-B entities, generated by OpenMEL, contain the ground truth. We use HIT@B, as defined in Equation (5), to represent the hit rate of the ground truth entity when only considering the top-B ranked entities produced by our method.

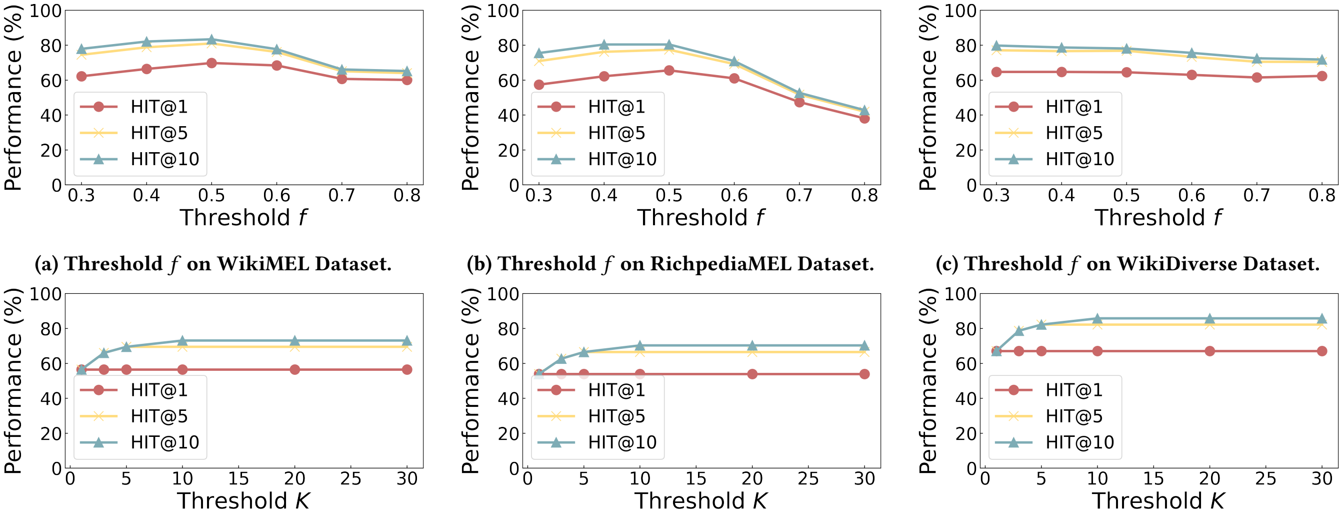
$$HIT@B = \frac{1}{N} \sum_{i}^{N} I(rank(i) < B) \tag{5}$$

N represents the total number of instances in test set, $rank(i)$ denotes the rank of the i-th ground truth entity in the ranking list of KB entities, and $I(\cdot)$ refers to the indicator function, which equals 1 if the subsequent condition is satisfied and 0 otherwise [27].

5.2 Overall Comparison (RQ1)

We compare our proposed OpenMEL, using both textual inputs and combined textual-visual inputs, against nine baseline models across three benchmark datasets. To ensure a fairer comparison, we also equip unsupervised baselines with the same LLM-enhanced inputs, aligned with OpenMEL. Overall, OpenMEL achieves the best performance in three benchmark datasets when compared to all supervised and unsupervised learning methods in HIT@1 and HIT@5 on average. Based on the experimental results in Table 2, we present detailed observations and analysis as follows.

- According to Problem 1, the MEL task is to link the mention to the most similar entity in the KB, measured by HIT@1 as the



(a) Threshold f on WikiMEL Dataset. **(b) Threshold f on RichpediaMEL Dataset.** **(c) Threshold f on WikiDiverse Dataset.**

(d) # Edges K on WikiMEL Dataset. **(e) # Edges K on RichpediaMEL Dataset.** **(f) # Edges K on WikiDiverse Dataset.**

Figure 5: Parameter sensitivity experiments of filtering threshold f and # edges K on three datasets.

precision, while HIT@5 and HIT@10 measure the recall. Based on the experimental results in Table 2, different methods have their respective pros and cons. Our OpenMEL demonstrates competitive performance in the precision metric, while supervised learning methods, such as DWE, perform better in the recall metrics, especially HIT@10.

- According to results of "Baseline (Aligned)", integrating contextual information via LLMs enhances the performance of all baselines across benchmark datasets and evaluation metrics, which verifies that contextual enhancement by LLMs in MEL task is effective. However, there is still a substantial performance gap remaining between "Baseline (Aligned)" and OpenMEL. This difference is primarily due to two key technical contributions unique to OpenMEL: (1) a three-level adaptive selection of visual inputs, and (2) the construction of a maximum spanning tree to ensure global coherence. Together, these advancements significantly contribute to OpenMEL's superior performance.
- When comparing models using text-only inputs and those using textual-and-visual inputs, models with textual inputs demonstrate satisfying results. OpenMEL (T) achieves a promising result compared to other baselines with multimodal data inputs. Notably, MEL-HI [59] and JMEL [1], which incorporate both visual and textual inputs, exhibits similar or even worse performance compared to BERT [9], which uses only textual inputs. This highlights the presence of noise in multimodal data. If a method fails to properly manage such noise, its performance may degrade when utilizing multimodal inputs. Additionally, textual information remains a fundamental and crucial component for entity linking, even within the MEL task [27].

5.3 Ablation Study (RQ2)

To examine the impact of the two main proposed modules as well as multimodal inputs, we design four groups of experiments. In the first group, we replace our first Noise-Free Expanded Queries (NFEQ) generation module with the concatenated multimodal features used in unsupervised baselines, along with our second Global Coherence Resolution module (GCR), denoted as $m([T, V])$ + GCR. In the second group, we keep the first NFEQ module, and replace the second GCR module with Top-K retrieval from unsupervised baselines, denoted as NFEQ(T+V) + Top-K. The third group employs only textual inputs as NFEQ(T) + GCR and $m(T)$ + GCR. The final group removes both modules in OpenMEL, resulting in CLIP.

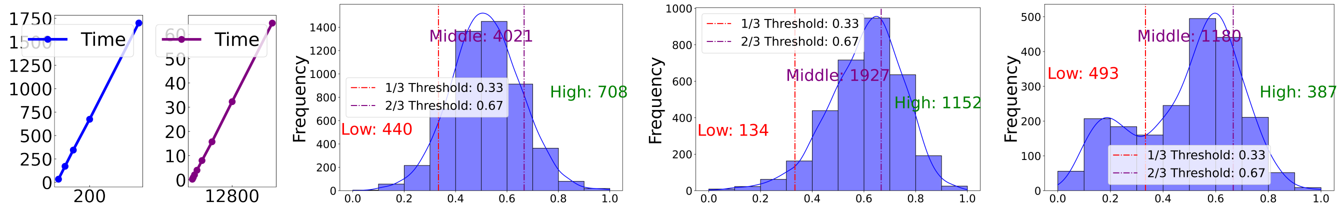
As shown in Table 3, replacing any module in OpenMEL with the common unsupervised learning function as $m(T + V)$ + GCR and NFEQ(T+V) + Top-K compared with original OpenMEL leads to a noticeable decline in all metrics across all datasets, demonstrating the effectiveness of our designed NFEQ and GCR modules. Specifically, the declines are more significant on the WikiDiverse dataset, as it covers a wider range of mention topics and contains the highest average number of mentions per sample, making it the most challenging benchmark dataset for the MEL task. This dataset exacerbates the problem of limited context, the noise tradeoff, and the entity global coherence challenge, all of which require robust solutions to address. Additionally, removing visual inputs from the multimodal inputs also results in a performance drop. This demonstrates that visual inputs, which contain valuable contextual information, can contribute to the improvement of the MEL task.

5.4 Parameter Sensitivity (RQ3)

In this subsection, we investigate the sensitivity of parameters in OpenMEL on three metrics across three datasets. The experimental results are shown in Figure 5. First, we examine the filtering threshold f, which varies from 0.3 to 0.8, and filter out entities with low similarity based on the mention's name. Specifically, the WikiMEL and RichpediaMEL datasets, which focus on person-related topics, show optimal performance when f is around 0.5. For person entities, filtering by name helps narrow down the entity candidates,

Table 4: Various LLM Backbones and Prompt Design Evaluation on OpenMEL Framework.

LLMs	Prompt	WikiMEL			RichpediaMEL			WikiDiverse		
		HIT@1	HIT@5	HIT@10	HIT@1	HIT@5	HIT@10	HIT@1	HIT@5	HIT@10
Llama3-8B (free)	Our Prompt	**69.8%**	**81.0%**	**83.4%**	**65.6%**	**77.4%**	**80.4%**	**67.1%**	**82.2%**	85.2%
	Ours w/o exp	66.4%	76.7%	79.1%	52.6%	65.6%	69.2%	57.8%	81.8%	**87.0%**
	Ours w/o exp & ins	69.7%	79.5%	81.7%	59.4%	71.6%	74.7%	57.0%	80.7%	85.5%
DeepSeek-V3 (costly)	Our Prompt	67.2%	**78.9%**	**81.7%**	**69.3%**	**78.7%**	**81.0%**	**63.8%**	**80.9%**	**84.7%**
	Ours w/o exp	**68.9%**	78.7%	80.9%	58.8%	71.5%	74.7%	60.6%	80.2%	84.6%
	Ours w/o exp & ins	65.2%	75.8%	78.5%	53.7%	67.2%	71.3%	54.8%	72.9%	77.1%
GPT-4o (high-cost)	Our Prompt	75.1%	**84.3%**	**85.9%**	**72.6%**	**81.1%**	**83.2%**	**72.4%**	**87.2%**	**90.3%**
	Ours w/o exp	74.8%	81.8%	83.2%	61.7%	72.0%	74.6%	61.5%	85.5%	89.5%
	Ours w/o exp & ins	**77.7%**	84.0%	85.5%	64.8%	75.1%	77.6%	58.7%	82.1%	87.2%
GPT-4o-mini (costly)	Our Prompt	70.2%	**81.0%**	**83.3%**	**64.5%**	**75.0%**	**78.3%**	**69.5%**	**84.6%**	**89.3%**
	Ours w/o exp	**71.0%**	79.5%	81.5%	57.1%	68.8%	72.4%	59.1%	82.3%	87.1%
	Ours w/o exp & ins	67.6%	77.3%	79.9%	55.1%	68.3%	72.3%	47.8%	70.7%	77.1%



(a) # Mentions (b) # Entities (c) Distribution of WikiMEL (d) Distribution of RichpediaMEL (e) Distribution of WikiDiverse

Figure 6: Data Analysis on Scalability and Normalized Similarity Distribution Between the Accompanying Image and Text.

significantly improving the linking performance. On the other hand, the WikiDiverse dataset, covering a range of topics, is less sensitive to the filtering threshold. Secondly, we explore the impact of the number of edges K in the constructed tree cover. As shown in Figure 5, performance initially increases and then plateaus as K grows for metrics HIT@5 and HIT@10. Moreover, we observe that $K = 1, 5, 10$ is sufficient to achieve the best result for HIT@1, HIT@5 and HIT@10. In conclusion, the parameter K can be set to the corresponding value of B in HIT@B to achieve optimal results.

5.5 LLMs Generation Quality Influences (RQ4)

We investigate the impact of LLM-generated content quality on the final linking results by a series of experiments focusing on two key factors that could lead to variations in LLM outputs: (1) Different LLM Backbones including GPT-4o, GPT-4o-mini, Llama3-8B, and DeepSeek-V3. (2) Prompting Strategies including "Our prompt" as shown in Figure 3, which includes both instructions and demonstration examples. Ours without demonstration examples denoted as "Ours w/o exp". Ours without both demonstration examples and instructions denoted as "Ours w/o exp & ins", using minimal guidance. As shown in Table 4, integrating advanced backbone LLMs such as GPT-4o for contextual enhancement can lead to superior performance in the entity linking task, but incurs higher API costs. GPT-4o-mini, DeepSeek-V3, and Llama3-8B exhibit similar performances. However, Llama3 is a cost-effective choice, achieving satisfactory results without additional expenses, particularly suitable for large-scale datasets. Additionally, the "Our prompt" strategy, which includes instructions and demonstration examples, has proven robust across datasets and metrics on average, making it the recommended approach for practical applications.

5.6 Data Analysis (RQ5)

We conclude with data analysis, as illustrated in Figure 6, to evaluate the scalability of OpenMEL and the relevance of images within the evaluation datasets. Specifically, the enhancement of multimodal data quality, as outlined in Section 3, is implemented during the preprocessing stage. This process requires LLMs for content generation, but since only a single round of generation is needed, the time cost remains reasonable. Furthermore, as shown in Figure 6a, OpenMEL demonstrates linear scalability with respect to the number of mentions awaiting linking, given a fixed number of candidate entities, totaling 132,460. Finally, as outlined in Section 4, the overall time complexity of OpenMEL is $O(|\mathcal{E}| \cdot \log |\mathcal{E}|)$ where $\mathcal{E}$ is the number of candidate entities. In Figure 6b, the results indicate that the running time remains relatively stable, with minimal increase as the number of candidate entities grows.

Additionally, to highlight the importance of addressing visual noise for multimodal inputs, we provide an analysis of the relevance between the accompanying image and the text, as shown in Figure 6. Specifically, accompanying images exhibit varying levels of similarity to the text inputs, with a significant number of instances falling into the low- and mid-similarity ranges. These statistical results highlight the critical need to address noise within accompanying images to enhance MEL performance.

6 RELATED WORK

In this section, we explore the related work on the entity linking task, encompassing both traditional text-based and multimodal entity linking methods.

6.1 Text-based Entity Linking

Mention and entity textual contexts are typically processed and encoded separately [14, 15, 24, 25, 32]. To capture more information from contexts, some methods employ *query expansion* [28, 33, 40, 43] to introduce external information aiding entity disambiguation [24, 25, 47]. However, traditional query expansion methods fail when surrounding context is unclear due to a lack of general knowledge and comprehension of the full context [30, 41]. Then, the enhanced or processed mention and entity text is represented via deep learning models [16, 26, 55]. For entities, some existing works use statistical co-occurrence methods to encode words [29] or train dense embeddings by retrieving sparse entity features through relationships between entities in the KB [5, 38]. Moreover, some works [35, 44] leverage knowledge graph embedding methods [2, 36] to integrate multi-hop information.

The entity ranking step takes the encoded mention and entity contexts and ranks the entities by assigning a score to each one. Existing methods for this step can be broadly categorized into unsupervised and supervised learning approaches. Unsupervised learning methods typically use dot product or cosine similarity to compute the similarity. Some approaches add an additional feed-forward network layer or a softmax layer to generate probabilities [16, 18]. Supervised learning methods, on the other hand, employ models such as reinforcement learning [15] and graph convolution networks (GCNs) [4, 54] to integrate features and explore diverse external cues between entities. Although text-based entity linking methods have made significant progress, they are limited to handling only textual data and are therefore incapable of addressing multimodal entity linking tasks.

6.2 Multimodal Entity Linking

Most existing studies [1, 27, 31, 51, 52, 56] approach the MEL task through supervised learning, which requires a large amount of labeled data for training. In contrast, unsupervised MEL methods reduce annotation costs by encoding multimodal inputs into feature representations and ranking entities based on semantic similarity.

Unlike text-based entity linking, multimodal entity linking faces the challenge of more severe limited context and noise within images during encoding with fusion. Supervised learning methods focus on deeply fusing multimodal inputs to extract implicit cues or attempt to eliminate noisy modalities. For the former, current approaches use models like cross-modal attention [1, 31], inter- and intra-modal attention [27, 53], or gate fusion techniques [51] for multimodal concatenation. On the other hand, in the absence of large training data, unsupervised learning methods typically encode textual and visual information using a large multimodal pre-trained model, such as CLIP [39]. These methods often propose a two-stage approach that examines the correlation between image categories and text semantics, filtering out noisy images with similarity scores below a predefined threshold [59]. However, directly filtering out noisy images exacerbates the limited context problem, potentially further diminishing MEL performance.

In the entity ranking step, methods in MEL can also be divided into two types, unsupervised and supervised learning approach. Supervised learning methods with labeled data employ additional model structures, particularly attention-based mechanisms, in the encoding and fusion steps. These methods also use various techniques to match mention-entity information, including static alignment [1, 27, 31, 51, 53] and dynamic alignment through GCNs [56], each with their own specially designed loss functions. On the other hand, unsupervised learning methods typically employ a bipartite graph and further define the MEL task as the bipartite graph matching problem to solve it [17, 34], where the bipartite graph consists of two sets, the mention set and entity set with edges between mention and entity. However, the bipartite graph structure does not account for entity coherence, as their design inherently precludes edges within the set of candidate entities.

7 CONCLUSIONS

In this paper, we propose a novel unsupervised MEL framework, OpenMEL, designed to effectively tackle the MEL task by utilizing noise-free expanded queries and a tree cover structure that captures global entity coherence within the KB. Our approach specifically addresses the challenge of multimodal data quality in the initial data encoding step of unsupervised learning. We process images into three levels of visual information and enhance textual inputs with comprehensive context and general knowledge expansion. To ensure global entity coherence in the entity ranking step, we introduce a tree cover structure that captures both mention-entity and entity-entity relationships, framing the task as a maximum spanning tree problem with bounded nodes. We further propose a greedy algorithm with theoretical guarantees to solve it. Extensive experiments on five research questions related to the OpenMEL framework demonstrate its effectiveness compared to various supervised and unsupervised baselines.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their insightful reviews. Lei Chen's work is partially supported by National Key Research and Development Program of China Grant No. 2023YFF0725100, National Science Foundation of China (NSFC) under Grant No. U22B2060, Guangdong-Hong Kong Technology Innovation Joint Funding Scheme Project No. 2024A0505040012, the Hong Kong RGC GRF Project 16213620, RIF Project R6020-19, AOE Project AoE/E-603/18, Theme-based project TRS T41-603/20R, CRF Project C2004-21G, Guangdong Province Science and Technology Plan Project 2023A0505030011, Guangzhou municipality big data intelligence key lab, 2023A03J0012, Hong Kong ITC ITF grants MHX/078/21 and PRP/004/22FX, Zhujiang scholar program 2021JC02X170, Microsoft Research Asia Collaborative Research Grant, HKUST-Webank joint research lab and 2023 HKUST Shenzhen-Hong Kong Collaborative Innovation Institute Green Sustainability Special Fund, from Shui On Xintiandi and the InnoSpace GBA. Yongqi Zhang's work is supported by Guangdong Basic and Applied Basic Research Foundation 2025A1515010304, and Guangzhou Science and Technology Planning Project 2025A03J4491.

REFERENCES

[1] Omar Adjali, Romaric Besançon, Olivier Ferret, Hervé Le Borgne, and Brigitte Grau. 2020. Building a multimodal entity linking dataset from tweets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. 4285–4292.

[2] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems* 26 (2013).

[3] Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).

[4] Yixin Cao, Lei Hou, Juanzi Li, and Zhiyuan Liu. 2018. Neural collective entity linking. *arXiv preprint arXiv:1811.08603* (2018).

[5] Yixin Cao, Lifu Huang, Heng Ji, Xu Chen, and Juanzi Li. 2017. Bridge text and knowledge by learning multi-prototype entity mention embedding. Association for Computational Linguistics.

[6] Tao Cheng, Kevin Chen-Chuan Chang, et al. 2007. *Entity Search Engine: Towards Agile Best-Effort Information Integration over the Web.* Ph.D. Dissertation. University of Illinois at Urbana-Champaign.

[7] 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)*. 708–716.

[8] Antonin Delpeuch. 2019. Opentapioca: Lightweight entity linking for wikidata. *arXiv preprint arXiv:1904.09131* (2019).

[9] Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[10] Mark Dredze, Paul McNamee, Delip Rao, Adam Gerber, Tim Finin, et al. 2010. Entity disambiguation for knowledge base population. In *Proceedings of the 23rd international conference on computational linguistics.*

[11] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[12] Mohnish Dubey, Debayan Banerjee, Debanjan Chaudhuri, and Jens Lehmann. 2018. EARL: joint entity and relation linking for question answering over knowledge graphs. In *The Semantic Web–ISWC 2018: 17th International Semantic Web Conference, Monterey, CA, USA, October 8–12, 2018, Proceedings, Part I 17*. Springer, 108–126.

[13] Yotam Eshel, Noam Cohen, Kira Radinsky, Shaul Markovitch, Ikuya Yamada, and Omer Levy. 2017. Named entity disambiguation for noisy text. *arXiv preprint arXiv:1706.09147* (2017).

[14] Wenfei Fan, Liang Geng, Ruochun Jin, Ping Lu, Resul Tugay, and Wenyuan Yu. 2022. Linking Entities across Relations and Graphs. In *38th IEEE International Conference on Data Engineering, ICDE 2022, Kuala Lumpur, Malaysia, May 9-12, 2022*. IEEE, 634–647. https://doi.org/10.1109/ICDE53745.2022.00052

[15] Zheng Fang, Yanan Cao, Qian Li, Dongjie Zhang, Zhenyu Zhang, and Yanbing Liu. 2019. Joint entity linking with deep reinforcement learning. In *The world wide web conference*. 438–447.

[16] Matthew Francis-Landau, Greg Durrett, and Dan Klein. 2016. Capturing semantic similarity for entity linking with convolutional neural networks. *arXiv preprint arXiv:1604.00734* (2016).

[17] 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*. 993–1001.

[18] Octavian-Eugen Ganea and Thomas Hofmann. 2017. Deep joint entity disambiguation with local neural attention. *arXiv preprint arXiv:1704.04920* (2017).

[19] Emma J Gerritse, Faegheh Hasibi, and Arjen P de Vries. 2022. Entity-aware transformers for entity search. In *Proceedings of the 45th international acm sigir conference on research and development in information retrieval*. 1455–1465.

[20] Dorit S Hochbaum. 1996. Approximating covering and packing problems: set cover, vertex cover, independent set, and related problems. In *Approximation algorithms for NP-hard problems*. 94–143.

[21] Sai Muralidhar Jayanthi, Varsha Embar, and Karthik Raghunathan. 2021. Evaluating pretrained transformer models for entity linking in task-oriented dialog. *arXiv preprint arXiv:2112.08327* (2021).

[22] Heng Ji and Ralph Grishman. 2011. Knowledge base population: Successful approaches and challenges. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*. 1148–1158.

[23] Joseph B Kruskal. 1956. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical society* 7, 1 (1956), 48–50.

[24] Xueling Lin, Lei Chen, and Chaorui Zhang. 2021. Tenet: Joint entity and relation linking with coherence relaxation. In *Proceedings of the 2021 International Conference on Management of Data*. 1142–1155.

[25] Xueling Lin, Haoyang Li, Hao Xin, Zijian Li, and Lei Chen. 2020. KBPearl: a knowledge base population system supported by joint entity and relation linking. *Proceedings of the VLDB Endowment* 13, 7 (2020), 1035–1049.

[26] 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* (2019).

[27] Pengfei Luo, Tong Xu, Shiwei Wu, Chen Zhu, Linli Xu, and Enhong Chen. 2023. Multi-grained multimodal interaction network for entity linking. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1583–1594.

[28] Rila Mandala, Takenobu Tokunaga, and Hozumi Tanaka. 1999. Combining multiple evidence from different types of thesaurus for query expansion. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. 191–197.

[29] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems* 26 (2013).

[30] Bhaskar Mitra, Nick Craswell, et al. 2018. An introduction to neural information retrieval. *Foundations and Trends® in Information Retrieval* 13, 1 (2018), 1–126.

[31] 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)*. 2000–2008.

[32] Dat Ba Nguyen, Abdalghani Abujabal, Khanh Tran, Martin Theobald, and Gerhard Weikum. 2017. Query-driven on-the-fly knowledge base construction. *Proceedings of the VLDB Endowment* 11, 1 (2017).

[33] Yusuke Ohura, Katsumi Takahashi, Iko Pramudiono, and Masaru Kitsuregawa. 2002. Experiments on query expansion for internet yellow page services using web log mining. In *VLDB'02: Proceedings of the 28th International Conference on Very Large Databases*. Elsevier, 1008–1018.

[34] George Papadakis, Vasilis Efthymiou, Emanouil Thanos, and Oktie Hassanzadeh. 2021. Bipartite graph matching algorithms for clean-clean entity resolution: an empirical evaluation. *arXiv preprint arXiv:2112.14030* (2021).

[35] Alberto Parravicini, Rhicheek Patra, Davide B Bartolini, and Marco D Santambrogio. 2019. Fast and accurate entity linking via graph embedding. In *Proceedings of the 2nd Joint International Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA)*. 1–9.

[36] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 701–710.

[37] Robert Clay Prim. 1957. Shortest connection networks and some generalizations. *The Bell System Technical Journal* 36, 6 (1957), 1389–1401.

[38] Radhakrishnan Priya, Talukdar Partha, Varma Vasudeva, et al. 2018. ELDEN: Improved entity linking using densified knowledge graphs. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, LA, USA*. 1–6.

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[40] Muhammad Ahsan Raza, Rahmah Mokhtar, and Noraziah Ahmad. 2019. A survey of statistical approaches for query expansion. *Knowledge and information systems* 61 (2019), 1–25.

[41] Ridho Reinanda, Edgar Meij, Maarten de Rijke, et al. 2020. Knowledge graphs: An information retrieval perspective. *Foundations and Trends® in Information Retrieval* 14, 4 (2020), 289–444.

[42] Ahmad Sakor, Isaiah Onando Mulang, Kuldeep Singh, Saeedeh Shekarpour, Maria Esther Vidal, Jens Lehmann, and Sören Auer. 2019. Old is gold: linguistic driven approach for entity and relation linking of short text. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2336–2346.

[43] Nikos Sarkas, Nilesh Bansal, Gautam Das, and Nick Koudas. 2009. Measure-driven keyword-query expansion. *Proceedings of the VLDB Endowment* 2, 1 (2009), 121–132.

[44] Özge Sevgili, Alexander Panchenko, and Chris Biemann. 2019. Improving neural entity disambiguation with graph embeddings. In *Proceedings of the 57th annual meeting of the association for computational linguistics: student research workshop*. 315–322.

[45] Senbao Shi, Zhenran Xu, Baotian Hu, and Min Zhang. 2023. Generative multimodal entity linking. *arXiv preprint arXiv:2306.12725* (2023).

[46] Shezheng Song, Shan Zhao, Chengyu Wang, Tianwei Yan, Shasha Li, Xiaoguang Mao, and Meng Wang. 2024. A dual-way enhanced framework from text matching point of view for multimodal entity linking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 19008–19016.

[47] Zequn Sun, Qingheng Zhang, Wei Hu, Chengming Wang, Muhao Chen, Farahnaz Akrami, and Chengkai Li. 2020. A Benchmarking Study of Embedding-based Entity Alignment for Knowledge Graphs. *Proc. VLDB Endow.* 13, 11 (2020), 2326–2340. http://www.vldb.org/pvldb/vol13/p2326-sun.pdf

[48] Robert Endre Tarjan. 1975. Efficiency of a good but not linear set union algorithm. *Journal of the ACM (JACM)* 22, 2 (1975), 215–225.

[49] Fang Wang, Wei Wu, Zhoujun Li, and Ming Zhou. 2017. Named entity disambiguation for questions in community question answering. *Knowledge-Based Systems* 126 (2017), 68–77.

[50] Meng Wang, Haofen Wang, Guilin Qi, and Qiushuo Zheng. 2020. Richpedia: a large-scale, comprehensive multi-modal knowledge graph. *Big Data Research* 22 (2020), 100159.

[51] Peng Wang, Jiangheng Wu, and Xiaohang Chen. 2022. Multimodal entity linking with gated hierarchical fusion and contrastive training. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 938–948.

[52] Sijia Wang, Alexander Hanbo Li, Henry Zhu, Sheng Zhang, Chung-Wei Hang, Pramuditha Perera, Jie Ma, William Wang, Zhiguo Wang, Vittorio Castelli, et al. 2023. Benchmarking diverse-modal entity linking with generative models. *arXiv preprint arXiv:2305.17337* (2023).

[53] Xuwu Wang, Junfeng Tian, Min Gui, Zhixu Li, Rui Wang, Ming Yan, Lihan Chen, and Yanghua Xiao. 2022. WikiDiverse: a multimodal entity linking dataset with diversified contextual topics and entity types. *arXiv preprint arXiv:2204.06347* (2022).

[54] Junshuang Wu, Richong Zhang, Yongyi Mao, Hongyu Guo, Masoumeh Soflaei, and Jinpeng Huai. 2020. Dynamic graph convolutional networks for entity linking. In *Proceedings of The Web Conference 2020.* 1149–1159.

[55] 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* (2019).

[56] Shangyu Xing, Fei Zhao, Zhen Wu, Chunhui Li, Jianbing Zhang, and Xinyu Dai. 2023. DRIN: Dynamic Relation Interactive Network for Multimodal Entity Linking. In *Proceedings of the 31st ACM International Conference on Multimedia.* 3599–3608.

[57] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. MiniCPM-V: A GPT-4V Level MLLM on Your Phone. *arXiv preprint arXiv:2408.01800* (2024).

[58] Gongrui Zhang, Chenghuan Jiang, Zhongheng Guan, and Peng Wang. 2023. Multimodal entity linking with mixed fusion mechanism. In *International Conference on Database Systems for Advanced Applications*. Springer, 607–622.

[59] Li Zhang, Zhixu Li, and Qiang Yang. 2021. 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.* Springer, 533–548.