# Multimodal Entity Linking: A New Dataset and A Baseline

Jingru Gan[1,2,3], Jinchang Luo[3], Haiwei Wang[3], Shuhui Wang[1]*, Wei He[3], and Qingming Huang[2,1]
[1]Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences
[2]School of Computer Science and Technology, University of Chinese Academy of Sciences [3]Baidu Inc.
jingru.gan@vipl.ict.ac.cn, wangshuhui@ict.ac.cn, {luojinchang,wanghaiwei,hewei06}@baidu.com, qmhuang@ucas.ac.cn

## ABSTRACT

In this paper, we introduce a new Multimodal Entity Linking (MEL) task on the multimodal data. The MEL task discovers entities in multiple modalities and various forms within large-scale multimodal data and maps multimodal mentions in a document to entities in a structured knowledge base such as Wikipedia. Different from the conventional Neural Entity Linking (NEL) task that focuses on textual information solely, MEL aims at achieving human-level disambiguation among entities in images, texts, and knowledge bases. Due to the lack of sufficient labeled data for the MEL task, we release a large-scale multimodal entity linking dataset M3EL (abbreviated for MultiModal Movie Entity Linking). Specifically, we collect reviews and images of 1,100 movies, extract textual and visual mentions, and label them with entities registered in Wikipedia. In addition, we construct a new baseline method to solve the MEL problem, which models the alignment of textual and visual mentions as a bipartite graph matching problem and solves it with an optimal-transportation-based linking method. Extensive experiments on the M3EL dataset verify the quality of the dataset and the effectiveness of the proposed method. We envision this work to be helpful for soliciting more research effort and applications regarding multimodal computing and inference in the future. We make the dataset and the baseline algorithm publicly available at https://jingrug.github.io/research/M3EL.

## CCS CONCEPTS

• **Computing methodologies** → **Information extraction**.

## KEYWORDS

Dataset; Multimodal alignment; multimodal entity linking; optimal transportation

**ACM Reference Format:**
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 Int'l Conference on Multimedia (MM '21), Oct. 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3474085.3475400

*Corresponding author.


*MM '21, October 20–24, 2021, Virtual Event, China.*

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475400

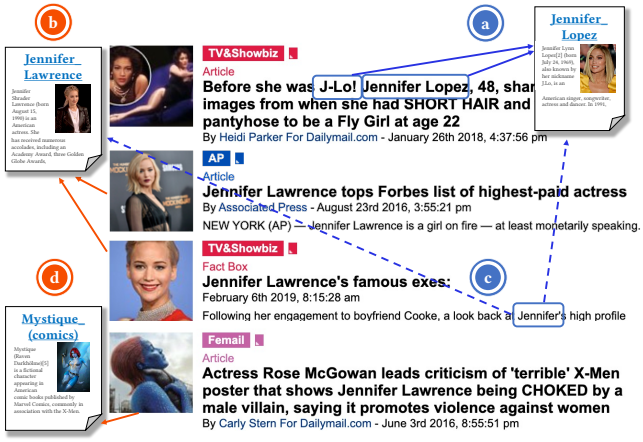


**Figure 1: Illustration of the MEL task. It links textual and visual mentions with entities in a knowledge base (e.g., Wikipedia). There are four challenges regarding the MEL problem: (a) textual mention ambiguity, (b) image variety, (c) polysemous textual entity, and (d) facial ambiguity.**

## 1 INTRODUCTION

Named entities in various modalities and forms are ubiquitous in our daily life. The nature of information extraction requires us to link possible mentions to the source entity that is known by human. We have words, pictures, audios and videos from various sources to depict an entity. However, it is quite challenging for computers to connect these entities expressed by multi-granularity information in different modalities like human. This bottleneck also hinders the construction of large scale multimodal knowledge base and poses great challenges on the development of more advance computing techniques for comprehensive understanding of real world multi-modal data.

The aforementioned situation leads to the advent of **Multimodal Entity Linking**, where we disambiguate mentions in multiple modalities by linking them to the corresponding named entities in a large knowledge base such as Wikipedia.

**Definition 1** ***Multimodal Entity Linking (MEL).*** *Given a set of textual mentions* $M^t = \{m_1^t, \cdots, m_n^t\}$ *and a set of visual mentions* $M^v = \{m_1^v, \cdots, m_m^v\}$ *in a multimodal document* $D$*, and a knowledge base* $KB = \{E, L\}$*, the MEL problem is to find the assignment* $\Gamma : M^t, M^v \rightarrow E$.

Considering the complexity of real world multimodal data, the MEL problem poses critical challenges as depicted in Figure 1.

- **Textual mention ambiguity**, which indicates the textual mention diversity different from the standard form of the

factual entity. For example, "J-Lo" and "Jennifer Lopez" point to the same Wikipedia entry "Jennifer_Lopez".
- **Visual diversity**, which means that each entity is displayed as dissimilar image mentions in different perspectives, light conditions and scenarios.
- **Polysemous textual entity**, which stands for possible ambiguity of a textual mention referring to multiple entities. Taking an ambiguous mention of "Jennifer" for example, it may refer to some entity with the same first name.
- **Facial ambiguity**, which suggests a common case that several human entities share the same facial information, e.g., movie stars and the characters they played.

Unfortunately, the lack of fine-grained annotated data becomes a main obstacle for further attempts to solve this emerging multimodal problem. Although diversified types of multimodal data are available thanks to the thriving multimedia applications, the labeled linkage between entities in different data modalities is still rare within these datasets. Moreover, considering diversity in multimodal contents, it is hard to have entity alignment information between modalities. For example, the presented entities even in the aligned image-text pair may still be uncorrelated, while the entities in some other unaligned image-text pairs may be actually linked. To address the above-mentioned concerns and facilitate the development of the fine-grained multi-modal entity linkage techniques, we collect and label a new benchmark dataset for the proposed MEL task.

If we investigate the interpretation of multimodal entity linking, several tasks that have connections with MEL and related public datasets have been proposed in previous study. For the Neural Entity Linking (NEL) task, a number of unimodal entity linking datasets are available. For instance, AIDA datasets [12] are constructed for training and testing, and other smaller datasets UIUC [25] and WNED [9] are only used for testing. Another task similar to MEL is Multimodal Named Entity Recognition (MNER) that performs named entity recognition (NER) with extra multimodal information. For this task, two preprocessed collections of multimodal Twitter posts are made public [19, 34, 35], where each tweet contains a related image. Moreover, another task [1, 21], also called MEL, focuses on linking word-image pairs to entities. In their work, two datasets collected from social media are proposed for this task. Adjali et al. [1] collect tweets for multimodal entity linking which links social media posts with users. Moon et al. [21] propose to find entities in short sentences with images from Snapchat data. It is worth noting that these multimodal entity linking tasks perform NEL with the assistance of multimodal data. For instance, the mentioned users with tweeted picture and users with profile picture are linked in Adjali et al. [1]. Compared to the tasks mentioned above, our proposed MEL requires joint disambiguation of both modalities, and we will discuss the difference between our task and others in consequent sections.

Different from existing datasets that are collected from social media platforms, we explore the MEL problem on movie related multimodal data corpus, which involve textual documents (e.g., movie reviews) and rich resources of photos from movies and behind the scenes. To create the M3EL dataset, we conduct multimodal entity linking among movie reviews, movie-related images, and Wikipedia entries of movie stars and movie characters. Specifically, we collect a total of 1,100 movies including their reviews and related images. Each named entity and images of person are manually labeled with the name of its linked entity. The resulting dataset is divided into training set, validation set, and testing set consisting of 1,000, 50, and 50 movies without overlap, respectively.

Based on the specific setting, the standard pipeline for our task is clearly different from existing ones such as Adjali et al. [1], Lu et al. [19], Moon et al. [21], Yu et al. [34], Zhang et al. [35]. Considering that both the entity to be linked and entity in the knowledge base may be expressed as either image or text, the entity linking in either modality can be naturally formulated as a bipartite graph matching problem given a batch of training data. Therefore, we develop an optimal-transportation-based deep learning baseline model to perform multimodal entity linking on this dataset.

To summarize, the main contribution of this paper are as follows.

- We present a novel Multimodal Entity Linking (MEL) task, which disambiguates mentions in multiple modalities with corresponding named entities.
- We release a finely-labeled Multimodal Movie Entity Linking dataset M3EL that focuses on disambiguation of movie characters given textual documents and pictures.
- We further provide a baseline method that models the MEL task as a bipartite graph matching problem, which can be efficiently solved by optimal transportation techniques.
- Extensive experiments demonstrate the high quality of the dataset and the effectiveness of our baseline algorithm.

To foster future research, we release the dataset as well as the baseline algorithm at https://jingrug.github.io/research/M3EL.

## 2 RELATED WORK

### 2.1 Entity Linking in NLP

The problem of Neural Entity Linking (NEL) lies in identifying an entity mention in unstructured text and establishing a link to an entry in a structures Knowledge Base (KB), e.g., Wikipedia and Freebase [7, 26]. As an essential component of information extraction, NEL has functioned as a basis of knowledge-based applications, including knowledge-based question answering, content analysis, information retrieval and knowledge base population.

In the language domain, two obvious challenges of conducting such task are mention variations and entity ambiguity. Mention variation indicates that an entity could be mentioned in various forms. For example, "Jennifer Lopez" may be referred to as "J.Lo", "JL", and "Lopez" in texts. On the contrary, entity ambiguity means that the surface form of a mention could relate to different entities under different contexts. Take the same example, "Jennifer" may stand for "Jennifer Lopez" and "Jennifer Lawrence". Apart from these two key problems, NEL methods also suffer from low resources and data noises and gradually outdating from knowledge bases.

Many NEL methods tackle these challenges with a two-stage disambiguation process. The general architecture includes candidate entity generation and entity ranking. The former stage aims at reducing the large decision space by means of surface form matching, dictionary lookup and calculating prior probabilities such as linkcount statistics. While the latter stage generates a final entity

matching score ranking result using normally the mentions with contexts to compare with candidate entities.

During entity ranking stage, the key to disambiguating mentions is to encode information from mentions and context. This process is first conducted with convolutional networks [29] with attention modeling the coherence between candidate entities [6, 9, 16, 18, 23], and then with recurrent architectures like LSTMs [11, 17, 36]. Gupta et al. [11] take two LSTMs to encode the left and right side words of the mention. Le and Titov [17] incorporate the embedding of a bi-LSTM and word position embeddings. Latter, pre-trained models, such as pre-trained ELMo [32] and BERT [2], are used to generate more semantic consistent embeddings.

## 2.2 NEL Systems and Applications

In order to train an entity linking model, human annotated pairs of mentions and entities are required. The most prominent training data of entity linking task is the AIDA (Accurate Online Disambiguation of Named Entities) dataset proposed by Hoffart et al. [12]. AIDA is created by annotating the CoNLL-2003 data, which contains 1,393 documents with mention-entity labels. As the largest one so far, the AIDA-CoNLL dataset is used for training NEL models using dataset splits referred to as AIDA-training, AIDA-A and AIDA-B. The technology of NEL has already been ubiquitous in a wide range of commercial systems in our daily lives. As a basic means of recommendation, NEL can be seen in almost every application on our smart phones. The construction and updating of every knowledge database behind the screen ought to thank NEL models for the disambiguation among words and entities.

## 2.3 Entity Linking in Multimodal Data

Many applications often require detection, classification, or extracting information from a single type of data such as audio, text, image, etc. With regards to multimodal data like social media contents, the task of Multimodal Named Entity Recognition (MNER) is introduced. The task of MNER aims at discovering named entities in unstructured text and documents in other modalities and entity types including person (PER), location (LOC), organization (ORG), and miscellaneous (MISC). Previous attempts made on MNER [3, 34] has lead to significant consequence. All these works focus on social media contents such as multimodal tweets from Twitter. Yu et al. [34] propose a multimodal transformer model to obtain word-object level attention. While Asgari-Chenaghlu et al. [3] incorporate visual information to empower the proposed Transformer-based NER model.

Some researches [1, 21] also contribute to entity linking among multimodal data that is Multimodal NEL. Moon et al. [21] proposed a multimodal named entity disambiguation task with a dataset constructed on Snapchat data. They use paired image to disambiguate a textual mention in a short caption less than 5 words. Adjali et al. [1] built a multimodal entity linking dataset with tweets, to perform linking from mentions in tweets to Twitter users.

The similarities of the two works and ours are that all tasks aim at correctly linking mentions to entities in knowledge bases with multimodal information. However, we believe that both tasks are boosting textual NEL with visual clues while we treat both modalities equally important. What most clearly differentiates our proposed task from theirs is the topical comprehension and linking on both textual and visual mentions.

In general, our proposed multimodal entity linking task can be part of complicated multimodal tasks. For instance, in a visual question answering (VQA) task, the linked entities of image regions and texts provided by MEL models could help machine better comprehend questions and answers. Similarly, for video captioning, MEL models could provide rich information of entities in frames and subtitles. We believe that the problem of MEL would be an essential but challenging task and a fascinating research direction as well.

## 2.4 MEL Methods

In this section, we briefly explain the MEL problem, from uni-modal entity linking methods to a naïve multimodal method and to our joint disambiguation methods.

First, to perform the multimodal entity linking on the given multimodal documents, we reduce this linking problem to a unimodal neural entity linking task, a long studied fundamental problem in natural language processing, specialized in linking textual mentions in a document to textual entities in knowledge bases.

Most previous work of NEL in textual domain follows a three-step paradigm:

(1) First, to locate a referred entity in the ocean of entities, many researches [12, 14, 17, 22, 24, 37] perform a preliminary filtering process called candidate entity generation, which generates a list of possibly related entities given a quested ambiguous mention using statistics.
(2) Second, to rank among the selected candidates, NEL models use recurrent networks [8, 11, 14, 17, 20, 27] or self-attentions [30] to leverage the information of mention contexts and entity descriptions.
(3) To further improve linking accuracy, some works [9, 15, 17, 32] find it crucial to use the topical coherence of entities in the same document as well as relations among mentions.

With regard to the three aforementioned procedures, the unimodal entity linking is not quite consistent to our multimodal setting, given the reasons below. First, the lack of prior statistical information like linkcount (mention-entity hyperlink count statistics) data makes it improbable to filter and produce a candidate matching list that contains possible matches between a given image and a massive number of visual entities. Second, the multimodal mentions describing one target entity are much more diverse than plain texts; the accuracy of directly comparing mentions with entity information lowers as the variety becomes large. Third, the naturally existed many-to-many matching among multimodal mentions are more complex than the unimodal setting.

To date, whereas NEL has been studied extensively, we argue that it is nontrivial to adapt existing NEL to the MEL setting. In what follows, a systematic solution framework for MEL remains to be considered. A straightforward solution would be a two-stage process to consider the linking of textual features and visual features independently, which overcomes the first drawback that needs to rank among all entities. To be specific, it firstly links similar texts and then directly compares images to that of the entities selected in the previous textual entity linking stage. Such a naïve method is reasonable only when the mentioned images and images of entities

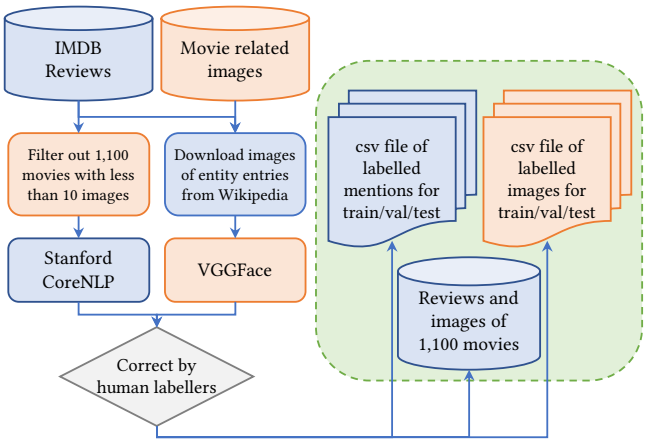


**Figure 2: An overview of the data collection process. The data collection process is composed of automatic labelling stage and human annotating stage. Textual information and visual information are displayed in blue and orange respectively.**

come from the same domain, such as linking customers' comments with pictures to products in online store or disambiguating citizens with registered voters. However, it is hard to generalize to other real-life data, involving social media posts and movie comments for example. When the mentioned images and entity images are from various domains, and the diversity of mentions will make the model performance significantly degrade.

Considering the drawbacks of directly using single modal NEL and the naïve solution, we believe the key to improving the multimodal entity linking accuracy is to explicitly model a *many-to-many relation* of multimodal mentions within a document. Motivated by this observation, we propose a simple but effective baseline for the MEL task. In particular, for each multimodal document, we construct textual mentions and visual mentions as two graphs. The many-to-many relationships can then be naturally modeled as a bipartite matching problem that can be solved by optimal transport [5, 31, 33]. We measure the distance of mentions from two modalities with a Gromov-Wasserstein distance (GWD) [33]. Then we jointly train the GWD learning and visual linking process mentioned in heuristic linking method above. The performance evaluation on our proposed method indicates that it can be used a strong baseline for MEL task.

## 3 DATASET CONSTRUCTION

In this section, we describe in detail the construction process of our dataset. Figure 2 shows the overview of the data collection process. The whole dataset is originated from three main sources of data: Internet Movie Data Base (IMDb)[1], The Movie Database (TMDb) [2], and Wikipedia. We extract full reviews of 1,100 random selected movies from IMDB dataset and correlating images from IMDB and TMDB which are organized by MovieNet [13]. To label multimodal mentions with related entities, we perform a two-stage labelling process with automatic labelling first and human labelling later. For textual and visual mentions, we use the Stanford CoreNLP package and VGGFace2 respectively to automatically link mentions to entities in a Wikipedia dump. Human annotators are introduced next to correct any false labels.

[1]http://www.imdb.com
[2]https://www.themoviedb.org

**Table 1: Statistics of the proposed dataset M3EL.**

| | |
|---|---|
| # Total movies | 1,100 |
| # Total images (visual entities) | 45,297 |
| # Total textual mentions | 181,240 |
| # Average textual mentions per movie review | 165 |
| # Average images per movie | 41 |
| # Average words per movie review | 17,319 |
| # Average ground truth entities per movie | 18 |

### 3.1 Dataset Statistics

Table 1 summarizes the statistics of M3EL. We collect summarized reviews of 1,100 movies, 181,240 labelled textual mentions, and 45,297 images in total. For each movie review, 165 mentions are extracted out of a 17,319-word-length document in average. 41 images on average are collected for each movie, each of which contains the facial image of the main movie character. Table 2 shows the recall rate of each dataset, which indicates the percentage of mentions containing the ground truth entity in the candidate sets that are generated using the technique proposed by [12].

### 3.2 Dataset Properties

In order to perform high quality multimodal entity linking on the proposed dataset, we set up a few requirements for the data collection.

- *Each movie has 41 images in average, and each contains a main character.* The number of images for each movie is between 10 to 50, which ensures the entity linking task to be challenging due to the diversity of visual appearance.
- *Almost all visual entities refer to actors and characters.* Currently, we mainly focus on the persons mentioned in movie reviews and photographs.
- *Most movies (recent movies especially) have images under different scenes including still scene from the movie scene and other related scenarios.* The images including still frames from the movie, photos of actors taken at related social events, posters and other artificial products.
- *The disambiguation of textual mentions have various levels of difficulty.* About 70% of the mentions are only part (e.g., first name or surname) of the ground truth entities mostly from the movie title, e.g., "Tom" for "Tom Cruise". Only 12% of the mentions match the entities exactly.
- *Textual mentions have overlapping surface forms.* For example, Confusions between the film "Capote" and the character "Truman Capote" are frequently occurred and quite challenging.

**Table 2: Recall rates on different datasets.**

| Dataset | Subset | #Docs | Recall |
|---|---|---|---|
| AIDA | AIDA-train | 946 | — |
| | AIDA-A (val) | 218 | 97.3% |
| | AIDA-B (test) | 232 | 98.3% |
| UIUC | MSNBC | 20 | 98.5% |
| | ACE2004 | 35 | 90.7% |
| | AQUAINT | 50 | 94.4% |
| WNED | CWEB | 320 | 91.7% |
| | Wiki | 318 | 92.3% |
| M3EL | M3EL-train | 1,000 | — |
| | M3EL-val | 50 | 89.5% |
| | M3EL-test | 50 | 89.3% |

### 3.3 Discussions

We compare our proposed M3EL dataset to the other two multimodal social media entity linking datasets [1, 21].

- *The construction.* We construct each document with multiple textual and visual mentions concerning the same movie. Moon et al. [21] build a dataset with Snapchat data that each pair of an image and a short caption links to an entity. And Adjali et al. [1] link paired name and image in a tweet to twitter users.
- *Data domains.* Our dataset contains long movie reviews with various entities and images related to the movie. In comparison, both of the datasets [1, 21] are built from social media data, where the images do not necessarily have semantic objects.
- *Data scale.* Moon et al. [21] collect 12K image-caption pairs and Adjali et al. [1] collect 20k mentions from tweets, while we have 181,240 textual mentions and 45,297 images in total.

We discuss the significance of the proposed M3EL dataset. The purpose of our proposed multimodal entity linking dataset is to provide labelled data for training models for possible applications. We discuss a number of potential application scenarios of this dataset as follows.

- **Visual Question Answering (VQA)** is a promising application that requires the ability of multimodal aligning and reasoning. MEL could provide it with extra linking information on entities mentioned in images and questions.
- **Content analysis** that involves multimodal documents can benefit form MEL as well, e.g., news analysis and social media content analysis.
- **Information retrieval** from large databases using multimodal data can also use MEL as a component.
- **Knowledge base population** indicates the updating and evolution of large-scale multimodal knowledge base which requires multimodal disambiguation among old and new entities by using the MEL model.

## 4 THE PROPOSED FRAMEWORK

In this section, we describe in detail the proposed framework of multimodal entity linking. As depicted in Figure 3, our proposed framework is composed of three modules: neural entity linking, visual entity linking and multimodal disambiguation. Given a multimodal document $D$, which contains textual mentions $M^t = \{m_1^t, \cdots, m_n^t\}$ and visual mentions $M^v = \{m_1^v, \cdots, m_m^v\}$, a multimodal entity linking method predicts the linking entity $e_i^t, e_j^v \in E$ for mentions $m_i^t$ and $m_j^v$ respectively.

### 4.1 Textual Entity Linking

Given the document $D$, the textual entity linking module will perform the linking process individually for textual mentions as the first stage of the framework. Sticking to traditions[9, 10, 12, 15, 32], the textual entity linking process is conducted by two phases, i.e., candidate entity generation and entity ranking.

*Candidate Entity Generation.* Following [9, 15, 32], we employ the aforementioned linkcount statistics of large corpuses, which indicates the probability of the existence of linkage between given mentions and entities in the wild without considering any context. Similarly, we calculate this mention-entity prior score $\hat{p}(e \mid m)$ by averaging probabilities of mention-entity pair with hyperlink statistics from Wikipedia and YAGO dictionary [12, 28]. We first take the top 30 candidates with the highest prior probabilities as [9]. Then we further reduce the number to 7 by selecting the top 4 with highest prior probabilities and another top 3 out of 30 with highest local context-entity similarities $e^{t\top}(\sum_{w \in d_i} w)$. After the candidate entity generation process, a candidate list $C_i$ containing 7 possible candidate entities for each textual mention are generated for entity ranking.

*Entity Ranking.* Local model considers only the local context of mentions while ignoring the correlation among the pairwise linking decisions which is called coherence. Let $\hat{m}_i^t$ be a mention with its local context. Then the final ranking results are produced by ranking the scoring matrix of $(e_i^t, \hat{m}_i^t)$ [9, 10, 25]. Different from local model, many neural entity linking methods take advantage of the global coherence information for disambiguating among candidate entities. The global ranking results of mentions are computed as

$$E^* = \underset{e_i^t, e_j^t \in C_i \times \ldots \times C_n}{\operatorname{argmax}} \sum_{i=1}^{n} \Psi(e_i^t, \hat{m}_i^t) + \sum_{i \neq j} \Phi(e_i^t, e_j^t, D). \quad (1)$$

For our ranking method, we follow the entity ranking method MulRel [15] that jointly models $K$ latent relations among all mentions pairs. Presumably, each pair of mentions $(m_i^t, m_j^t)$ within the same document have certain relations, which can be modeled as $K$ latent relation embeddings. Then, the global-wise pairwise score is computed by summing up the weighted pairwise scores for each relation $k$:

$$\Phi(e_i^t, e_j^t, D) = \sum_{k=1}^{K} \alpha_{ijk} \Phi_k(e_i^t, e_j^t, D), \quad (2)$$

where $\Phi_k(e_i^t, e_j^t, D)$ is computed as $e_i^{t\top} R_k e_j^t$ for each pair, where each of the $k$ relations is parameterized by a matrix $R_k$ to be learned.

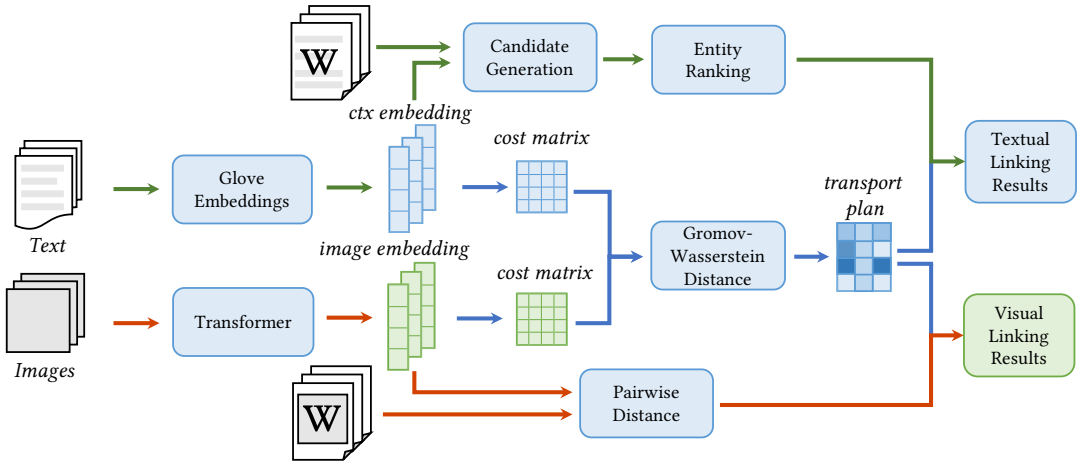


**Figure 3: The Framework of Multimodal Entity Linking. The framework contains three modules: neural entity linking, visual entity linking and multimodal disambiguation. It performs linking process respectively on textual and visual input, and then jointly disambiguates.**

Here, we use the scheme that normalizes over all mentions, which suggests that $\sum_{j=1,j\neq i}^{n} \alpha_{ijk} = 1$, to achieve better performance over the relation-wise normalization accordingly.

$$\alpha_{ijk} = \frac{1}{Z_{ijk}} \exp\left\{\frac{f^\top(\hat{m}_i^t)\mathbf{D}_k f(\hat{m}_j^t)}{\sqrt{d}}\right\}, \tag{3}$$

$$Z_{ijk} = \sum_{\substack{j'=1 \\ j'\neq i}}^{n} \exp\left\{\frac{f^\top(\hat{m}_i^t)\mathbf{D}_k f(\hat{m}_{j'}^t)}{\sqrt{d}}\right\}. \tag{4}$$

A ranking loss of mentions $m^t$ within a document $D$ is then calculated for training:

$$\mathcal{L}_{\text{ranking}}^t = \sum_{m_i^t \in D} \sum_{\hat{e}^t \in C_i} h\left(m_i^t, \hat{e}^t\right), \tag{5}$$

$$h(m_i^t, e^t) = \max(0, \gamma - \rho_i(e^t) + \rho_i(\hat{e}^t)), \tag{6}$$

where $\hat{e}^t \in C_i$ are candidate entities and $e^t$ are the ground truth entities.

## 4.2 Visual Entity Linking

We use a 6-layer transformer encoder with multi-head attention [30] to encode visual information of given images and profile images of entities. By dicing the input image into 64 pieces, the expanded sequence $x$ of pieces with positional encodings forms the input of transformer encoder. Within each layer of the transformer encoder, the input sequence is first processed by a self-attention sublayer for global attention then by a feedforward sublayer. A following average pooling layer then produces compact visual embeddings $m^v$. These computation can be mathematically expressed as:

$$\begin{aligned} o^v &= \text{LayerNorm}(x + (\text{MultiHead}(x))), \\ z_i^v &= \text{LayerNorm}(o^v + (\text{FNN}(o^v))), \\ m^v &= \frac{1}{64} \sum_{i\in\{1,\cdots,64\}} z_i^v. \end{aligned} \tag{7}$$

We assign each pair of visual mention and candidate entity with a similarity score $S_{ij}^v$ by computing a cosine pairwise distance, which will partially decide the visual linking results. A triplet loss is calculated to supervise the training procedure by sampling positive entities $e^v$ and negative entities $\hat{e}^v$:

$$\mathcal{L}_{\text{triplet}}^v = \max(0, \text{margin} - s(e^v, m^v) + s(\hat{e}^v, m^v)). \tag{8}$$

## 4.3 Multimodal Joint Disambiguation

Instead of reasoning through all candidate entities as illustrated in the heuristic model, we assign images to entities by discovering possible relations among mentions from different modalities. Within the same document, intra-modal mentions have subtle relations and inter-modal mentions may have strong relations of pointing to the same entity. Therefore, such relations could be viewed as a many-to-many bipartite graph matching problem.

However, finding the multimodal many-to-many matching is of great difficulty. On the one hand, identifying an entity of person in visual expression is challenging due to the visual appearance diversity. Images tend to have low similarity with the ground truth entity due to high variety of visual contents including images taken in different view angles, light conditions and surrounding environments. Second, due to the speciality of actors, one may appear as different entries of Wikipedia, indicating the actor himself and the role he played, as shown in the Figure 1. In such circumstance, methods like face recognition or person Re-ID are no longer reliable. On the other hand, the situation is just as tough for textual mention disambiguation for variety and diversity as well.

Therefore, we propose to jointly disambiguate both textual and visual mentions. By doing so, the method have two significant advantages. First, for each entity, among the several corresponding textual mentions and images, one image may respond to certain mentions and contexts but not all of them. Similarly, considering two images belonging to the actor himself and the role he played, they will respond to different mentions and contexts, *i.e.*, relating photos to ground truth entities under different circumstances.

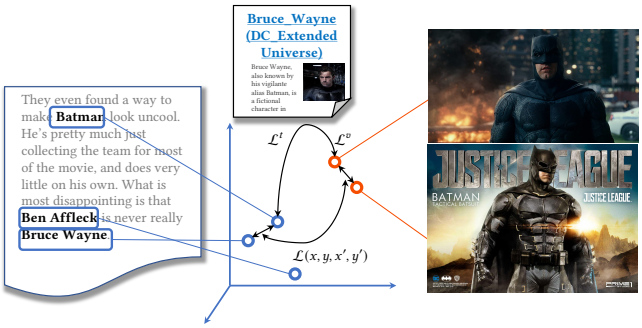


**Figure 4: Loss functions for multimodal joint disambiguation.**

*Graph Matching.* Given the analysis above, the matching between multimodal mentions can be formulated as a graph matching problem, and optimal transportation can be used to find a near optimal solution. Following many former practices [5, 31, 33], the Gromov-Wasserstein Distance is taken into consideration on such a problem setting. The distance measures the similarity between pairs of nodes across domains, which models the interdependencies between multimodal mentions. For example, in Figure 4, the mention pair ("Batman", "Bruce Wayne") and a pair of scene of Batman in the movie may have similar cosine similarity.

To perform optimal transportation, we use $\mu_t$ and $\nu_v$ to represent textual and visual mention distribution individually. We define a transportation plan $T$ that $T(M^t) = M^v$ maps the embeddings of all textual mentions in a document into visual mention embeddings. The distance $\mathcal{D}(\mu_t, \nu_v)$ indicates the minimum cost of transmitting from $M^t$ to $M^v$.

The GWD is calculated as

$$\mathcal{D}(\mu_t, \nu_v) = \min_{T \in \Pi(\mu_t, \nu_v)} \sum_{i,i',j,j'} T_{ij} T_{i'j'} \mathcal{L}(x_i, y_j, x'_i, y'_j), \tag{9}$$

where $(x_i, x'_i)$, $(y_i, y'_i)$ are two pairs of nodes from textual and visual mentions each relating to one entity, $\mathcal{L}(x_i, y_j, x'_i, y'_j) = \|c_1(x_i, x'_i) - c_2(y_i, y'_i)\|$ and $c(a, b)$ indicates Wasserstein Distance between two nodes.

Sinkhorn algorithm is applied to solve the GWD problem with an entropy regularizer:

$$\min_{T \in \Pi(\mu_t, \nu_v)} \sum_{i=1}^{n} \sum_{j=1}^{m} T_{ij} c\left(x_i, y_j\right) + \beta H(T), \tag{10}$$

where hyperparameters $H(T) = \sum_{i,j} T_{ij} \log T_{ij}$ and $\beta$ controls the weight of entropy regularizer.

The joint cost function of multimodal learning is defined as:

$$\mathcal{L}_{\text{joint}} = \mathcal{L}^t_{\text{ranking}} + \mathcal{L}^v_{\text{triplet}} + \mathcal{L}(x, y, x', y'). \tag{11}$$

Finally, the transport plan $T$ from joint disambiguation will involve both the textual entity linking and visual entity linking. For the similarity matrix of visual mentions and entities $S^v$, we rank the textual mention with minimum cost and add the weighted cost to $S^v$. Similar process is conducted for textual correlation matrix $S^t$ of mentions and candidate entities as well. The final linking results are produced by ranking among score matrices.

**Algorithm 1:** The multimodal entity linking algorithm

**Input :** Textual mentions $M^t$ and visual mentions $M^v$ in a document.
**Output :** Assigned entity to each mention.

**Procedure** `MM-nel`$(M^t, M^v)$:
- *Textual Entity Linking*:
  - Candidate entities generation
  - Compute global similarity scores $S^t$
- *Visual Entity Linking*:
  - Generate visual embeddings via Equation 7
  - Compute cosine similarity scores $S^v$
- *Joint Disambiguation*:
  - Compute GWD
  - Update mention-entity similarity matrices by $T$
- **return** assigned entities

## 5 EXPERIMENTS

**Table 3: Multimodal entity linking performance in terms of Micro-F1.**

| | Val | | Test | |
|---|---|---|---|---|
| | Textual | Visual | Textual | Visual |
| Naïve | 80.2 | 20.5 | 80.3 | 23.6 |
| MM-nel (ours) | 84.2 | 29.0 | 83.3 | 30.7 |

### 5.1 Experimental Setup

We perform experiments on our M3EL dataset. Several textual datasets are used for evaluation on the unimodal entity linking. These datasets include an established textual entity linking system AIDA for training and testing [12], UIUC datasets (ACE, MSNBC, and AQUAINT) [25] and WNED datasets (CWEB and Wiki) [9]. The performance on all datasets are measured by F1 score which incorporates precision and recall of mention-entity linking. While collecting our M3EL dataset, Stanford CoreNLP tool and VGGFace2[4] are used to carry out the first round of labeling. We perform all the linking on top of the English Wikipedia dump (edition of 2021-03-01). Wikipedia and YAGO corpus are used to compute the prior probability of mention-entity linkage.

We report experimental results of our proposed method for the MEL task, which uses M3EL dataset for training, validation and testing and serves as a baseline. To evaluate the quality of our released data and to test the performance of multimodal linking, the results of several linguistic entity linking methods trained on AIDA dataset are reported for comparison. We illustrate the baseline models of language entity linking [12], Guo and Barbosa [10], local and global model [9], as well as recent models including MulRel [15], the model we incorporated in our framework and DGCN model [32].

The dimension $d$ of latent space of both modalities are set to be 300. Following the implementation of MulRel, GloVe embedding

**Table 4: Textual entity linking results in terms of Micro-F1 on different datasets.**

| Train set | Methods | MSNBC | AQUAINT | ACE2004 | CWEB | WIKI | M3EL-Val | M3EL-Test |
|---|---|---|---|---|---|---|---|---|
| AIDA-train | Hoffart et al. [12] | 79 | 56 | 80 | 58.6 | 63 | — | — |
| | Ratinov et al. [25] | 75 | 83 | 82 | 56.2 | 67.2 | — | — |
| | Local $\hat{p}(e \mid m_i)$ [9] | 89.3 | 83.2 | 84.4 | 69.8 | 64.2 | 70.7 | 67.3 |
| | Ganea and Hofmann [9] | 93.7 | 88.5 | 88.5 | 77.9 | 77.5 | — | — |
| | DGCN [32] | 92.5 | 89.4 | 90.6 | 81.2 | 77.6 | — | — |
| | MulRel [15] | 93.9 | 88.3 | 89.9 | 77.5 | 78.0 | 75.1 | 70.2 |
| M3EL | Local $\hat{p}(e \mid m_i)$ | 89.9 | 82.3 | 86.1 | 69.4 | 67.3 | 79.5 | 79.3 |
| | Naïve (ours) | 90.3 | 88.9 | 88.1 | 72.6 | 66.6 | 80.2 | 80.3 |
| M3EL + AIDA-train | Local $\hat{p}(e \mid m_i)$ | 92.7 | 86.7 | 86.9 | 74.3 | 71.9 | 79.9 | 79.8 |
| | Naïve (ours) | 92.9 | 89.7 | 88.1 | 75.8 | 73.8 | 80.2 | 78.8 |

of words trained on 840B tokens are used to encode mentions and contexts, and entity embeddings are from [9]. Hyper-parameters used in the framework are as follows: the $\gamma = 0.01$, the drop-out rate of NEL module is 0.3, the window size of local contexts is 6, the diagonal matrix representing relations is sampled from $\mathcal{N}(0, 0.1)$. Early stopping is applied while training textual neural entity linking module by reducing the learning rate from $10^{-3}$ to $10^{-4}$ when F1 score of validation set hits 85%, which is the same as in [9, 15] During candidate entity generation, the first 30 entities are selected as candidates and the top 7 with highest prior score among them are kept for further ranking.

## 5.2 Results and Ablation Study

The experimental results of proposed MEL method are shown in Table 3. Our model is trained on M3EL training dataset. Compared to the naïve approach that performs entity linking on either textual modality or visual modality independently, our proposed multimodal entity linking baseline MM-nel demonstrate better accuracy due to the joint training on the two modalities. Specifically, the F1 scores of textual entity linking result and visual entity linking result are 83.3 and 30.7 respectively on testing data. The results is reported after 30 epochs, with training time and testing time less than 10 min. The proposed multimodal entity linking method works in an efficient way. Also, the results indicate that the visual entity linking appears to be more challenging than the linguistic entity linking.

Then we perform separate experiments on textual entity linking in order to compare with state-of-the-art language entity linking methods. The results on different datasets are shown in Table 4. Models trained on our proposed M3EL training set achieve comparable results on other test sets. If we use both AIDA and M3EL for training, the performance is further boosted on all test sets. However, it appears that our model trained using M3EL does not outperform traditional language-based approaches on some of the existing language test datasets. The reason can be explained by the domain discrepancy between M3EL collected from movie reviews and the language test datasets from news and online Web pages. How to bridge the domain gap to achieve better linguistic entity linking performance is still a challenging problem for future study.

## 6 CONCLUSION

We present a new multimodal entity linking task and construct a dataset for future study. For this task, we construct a baseline framework that works on multimodal entity linking. The experimental analysis demonstrate the good quality of our proposed dataset and the efficiency and effectivity of the baseline method.

The problem of Multimodal Entity Linking is a new task that maps ambiguous multimodal contents to structured knowledge bases. This task challenges the solver with the ability of lexical comprehension, visual information representation, multimodal alignment and large scale multimodal joint learning. The results of MEL could serve as stepping stones for many multimodal tasks such as VQA and video captioning, and provide current pre-trained models with more fine-grained entity linking information and even explanation. We sincerely believe the multimodal entity linking will be a challenging task that can inspire a wide range of future research directions.

## ACKNOWLEDGMENTS

This work is supported in part by the National Key R&D Program of China under Grant 2018AAA0102003, in part by National Natural Science Foundation of China under Grants 62022083, 61620106009, 61836002, and 61931008, and in part by Key Research Program of Frontier Sciences, Chinese Academy of Sciences, under Grant QYZDJ-SSW-SYS013.

## REFERENCES

[1] Omar Adjali, Romaric Besançon, Olivier Ferret, Hervé Le Borgne, and Brigitte Grau. 2020. Building a Multimodal Entity Linking Dataset From Tweets. *LREC* (2020).
[2] Omar Adjali, Romaric Besançon, Olivier Ferret, Hervé Le Borgne, and Brigitte Grau. 2020. Multimodal Entity Linking for Tweets. *ECIR* 12035, 1 (2020), 463–478.
[3] Meysam Asgari-Chenaghlu, M Reza Feizi-Derakhshi, Leili Farzinvash, M A Balafar, and Cina Motamed. 2020. A multimodal deep learning approach for named entity recognition from social media. *arXiv.org* (Jan. 2020). arXiv:2001.06888v3 [cs.CL]
[4] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. 2018. VGGFace2 - A Dataset for Recognising Faces across Pose and Age. *FG* (2018).
[5] Liqun Chen, Zhe Gan, Yu Cheng 0001, Linjie Li, Lawrence Carin, and Jingjing Liu 0001. 2020. Graph Optimal Transport for Cross-Domain Alignment. *ICML* (2020).
[6] Shuang Chen, Jinpeng Wang, Feng Jiang, and Chin-Yew Lin. 2020. Improving Entity Linking by Modeling Latent Entity Type Information. *arXiv.org* (Jan. 2020), arXiv:2001.01447. arXiv:2001.01447 [cs.CL]
[7] Alexandre Davis, Adriano Veloso, Altigran Soares da Silva, Alberto H F Laender, and Wagner Meira Jr. 2012. Named Entity Disambiguation in Streaming Data.

*Annual Meeting of the Association for Computational Linguistics* (2012).
[8] Zheng Fang, Yanan Cao, Ren Li, Zhenyu Zhang, Yanbing Liu, and Shi Wang. 2020. High Quality Candidate Generation and Sequential Graph Attention Network for Entity Linking. In *WWW '20: The Web Conference 2020*. ACM, New York, NY, USA, 640–650.
[9] Octavian-Eugen Ganea and Thomas Hofmann. 2017. Deep Joint Entity Disambiguation with Local Neural Attention. (April 2017). arXiv:1704.04920
[10] Zhaochen Guo and Denilson Barbosa. 2018. Robust named entity disambiguation with random walks. *Semantic Web* (2018).
[11] Nitish Gupta, Sameer Singh 0001, and Dan Roth. 2017. Entity Linking via Joint Encoding of Types, Descriptions, and Context. *EMNLP* (2017), 2681–2690.
[12] J Hoffart, M A Yosef, I Bordino, H Fürstenau Proceedings of the, and 2011. [n.d.]. Robust disambiguation of named entities in text. *aclweb.org* ([n. d.]).
[13] Q Huang, Y Xiong, A Rao, J Wang, D Lin Computer Vision ECCV 2020, and 2020. [n.d.]. Movienet: A holistic dataset for movie understanding. *Springer* ([n. d.]).
[14] Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. 2018. End-to-End Neural Entity Linking. *CoNLL* (2018).
[15] P Le, I Titov arXiv preprint arXiv 1804.10637, and 2018. [n.d.]. Improving entity linking by modeling latent relations between mentions. *arxiv.org* ([n. d.]).
[16] Phong Le and Ivan Titov. 2019. Boosting Entity Linking Performance by Leveraging Unlabeled Documents. *Annual Meeting of the Association for Computational Linguistics* (2019), 1935–1945.
[17] Phong Le and Ivan Titov. 2019. Distant Learning for Entity Linking with Automatic Noise Detection. *Annual Meeting of the Association for Computational Linguistics* (2019), 4081–4090.
[18] Pei-Chi Lo and Ee-Peng Lim. 2020. Interactive Entity Linking Using Entity-Word Representations. *SIGIR* (2020).
[19] Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. 2018. Visual Attention Model for Name Tagging in Multimodal Social Media. *Annual Meeting of the Association for Computational Linguistics* (2018).
[20] Pedro Henrique Martins, Zita Marinho, and André F T Martins. 2019. Joint Learning of Named Entity Recognition and Entity Linking. *Annual Meeting of the Association for Computational Linguistics* (2019).
[21] Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018. Multimodal Named Entity Disambiguation for Noisy Social Media Posts. *Annual Meeting of the Association for Computational Linguistics* (2018).
[22] Jose G Moreno, Romaric Besançon, Romain Beaumont, Eva D'hondt, Anne-Laure Ligozat, Sophie Rosset, Xavier Tannier, and Brigitte Grau. 2017. Combining Word and Entity Embeddings for Entity Linking. *ESWC* (2017).
[23] Yasumasa Onoe and Greg Durrett. 2020. Fine-Grained Entity Typing for Domain Independent Entity Linking. *AAAI* (2020).
[24] Maria Pershina, Yifan He, and Ralph Grishman. 2015. Personalized Page Rank for Named Entity Disambiguation. *HLT-NAACL* (2015).
[25] Lev-Arie Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and Global Algorithms for Disambiguation to Wikipedia. *Annual Meeting of the Association for Computational Linguistics* (2011).
[26] Ozge Sevgili, Artem Shelmanov, Mikhail Arkhipov, Alexander Panchenko, and Chris Biemann. 2020. Neural Entity Linking: A Survey of Models Based on Deep Learning. *arXiv.org* (June 2020), arXiv:2006.00575. arXiv:2006.00575 [cs.CL]
[27] Avirup Sil, Gourab Kundu, Radu Florian, and Wael Hamza. 2018. Neural Cross-Lingual Entity Linking. *AAAI* (2018).
[28] Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago - a core of semantic knowledge. *WWW* (2007).
[29] Yaming Sun, Lei Lin 0001, Duyu Tang, Nan Yang 0002, Zhenzhou Ji, and Xiaolong Wang 0001. 2015. Modeling Mention, Context and Entity with Neural Networks for Entity Disambiguation. *IJCAI* (2015).
[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *arXiv.org* (June 2017). arXiv:1706.03762v5 [cs.CL]
[31] Titouan Vayer, Nicolas Courty, Romain Tavenard, Laetitia Chapel, and Rémi Flamary. 2019. Optimal Transport for structured data with application on graphs. *ICML* (2019).
[32] Junshuang Wu, Richong Zhang, Yongyi Mao, Hongyu Guo, Masoumeh Soflaei, and Jinpeng Huai. 2020. Dynamic Graph Convolutional Networks for Entity Linking. In *WWW '20: The Web Conference 2020*. ACM, New York, NY, USA, 1149–1159.
[33] Hongteng Xu, Dixin Luo, and Lawrence Carin. 2019. Scalable Gromov-Wasserstein Learning for Graph Partitioning and Matching. (2019). arXiv:1905.07645
[34] Jianfei Yu, Jing Jiang 0001, Li Yang, and Rui Xia. 2020. Improving Multimodal Named Entity Recognition via Entity Span Detection with Unified Multimodal Transformer. *Annual Meeting of the Association for Computational Linguistics* (2020).
[35] Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. Adaptive Co-attention Network for Named Entity Recognition in Tweets. *AAAI* 32, 1 (April 2018).
[36] Xiaoling Zhou, Yukai Miao, Wei Wang 0011, and Jianbin Qin. 2020. A Recurrent Model for Collective Entity Linking with Adaptive Features. *AAAI* (2020).
[37] Stefan Zwicklbauer, Christin Seifert, and Michael Granitzer. 2016. Robust and Collective Entity Disambiguation through Semantic Embeddings. *SIGIR* (2016).