Multimodal Entity Linking for Tweets

Omar Adjali[1](✉), Romaric Besançon[1], Olivier Ferret[1], Hervé Le Borgne[1], and Brigitte Grau[2]

[1] CEA, LIST, Laboratoire Analyse Sémantique Texte et Image,
91191 Gif-sur-Yvette, France
{omar.adjali,romaric.besancon,olivier.ferret,herve.le-borgne}@cea.fr
[2] Université Paris-Saclay, CNRS, LIMSI, 91400 Orsay, France
brigitte.grau@limsi.fr

Abstract. In many information extraction applications, entity linking (EL) has emerged as a crucial task that allows leveraging information about named entities from a knowledge base. In this paper, we address the task of multimodal entity linking (MEL), an emerging research field in which textual and visual information is used to map an ambiguous mention to an entity in a knowledge base (KB). First, we propose a method for building a fully annotated Twitter dataset for MEL, where entities are defined in a Twitter KB. Then, we propose a model for jointly learning a representation of both mentions and entities from their textual and visual contexts. We demonstrate the effectiveness of the proposed model by evaluating it on the proposed dataset and highlight the importance of leveraging visual information when it is available.

Keywords: Information extraction · Entity linking · Multimodality

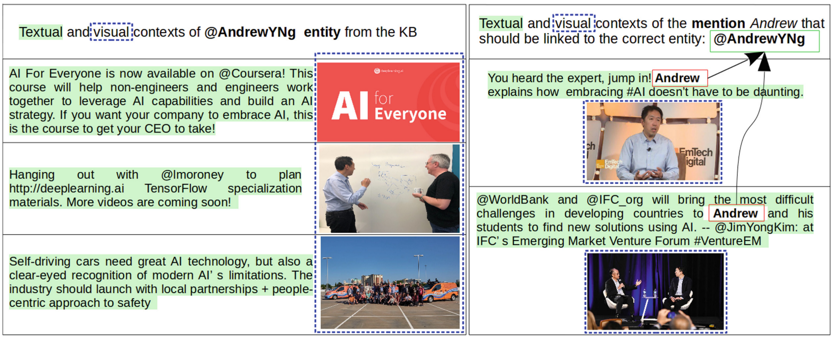
1 Introduction

Entity linking (EL) is a crucial task for natural language processing applications that require disambiguating named mentions within textual documents. It consists in mapping ambiguous named mentions to entities defined in a knowledge base (KB). To address the EL problem, most of the state-of-the-art approaches use some form of textual representations associated with the target mention and its corresponding entity. [3] first proposed to link a named mention to an entity in a knowledge base, followed by several other local approaches [9,10,14,15,35] where mentions are individually disambiguated using lexical mention-entity measures and context features extracted from their surrounding text. In contrast, global approaches [19,21,39,42] use global features at the document-level to disambiguate all the mentions within the document and can also exploit semantic relationships between entities in the KB. These approaches have shown their

This research was partially supported by Labex DigiCosme (project ANR11LABEX0045DIGICOSME) operated by ANR as part of the program "Investissements d'Avenir" Idex Paris Saclay (ANR11IDEX000302).

J. M. Jose et al. (Eds.): ECIR 2020, LNCS 12035, pp. 463–478, 2020.
https://doi.org/10.1007/978-3-030-45439-5_31

effectiveness on standard EL datasets such as TAC KBP [27], CoNLL-YAGO [23], and ACE [2] which contain structured documents that provide a rich context for disambiguation. However, EL becomes more challenging when a limited, informal, textual context is available, such as in social media posts. On the other hand, social media posts often include images to illustrate the text that could be exploited to improve the disambiguation by adding a visual context.


Textual and visual contexts of **@AndrewYNg entity** from the KB
Textual and visual contexts of the **mention** *Andrew* that should be linked to the correct entity: **@AndrewYNg**
AI for Everyone

AI For Everyone is now available on @Coursera! This course will help non-engineers and engineers work together to leverage AI capabilities and build an AI strategy. If you want your company to embrace AI, this is the course to get your CEO to take!

Hanging out with @lmoroney to plan http://deeplearning.ai TensorFlow specialization materials. More videos are coming soon!

Self-driving cars need great AI technology, but also a clear-eyed recognition of modern AI' s limitations. The industry should launch with local partnerships + people-centric approach to safety

You heard the expert, jump in! **Andrew** explains how embracing #AI doesn't have to be daunting.

@WorldBank and @IFC_org will bring the most difficult challenges in developing countries to **Andrew** and his students to find new solutions using AI. -- @JimYongKim: at IFC' s Emerging Market Venture Forum #VentureEM

Fig. 1. An illustrative example of multimodal entity linking on our Twitter dataset. The example depicts a subset of (text, image) pairs representing the entity @AndrewNg (left), and some example of (text, image) pairs related to the mention *Andrew* (right).

In this paper, we address the problem of multimodal entity linking (MEL), by leveraging both semantic textual and visual information extracted from mention and entity contexts. We propose to apply MEL on Twitter posts as they form a prototypical framework where the textual context is generally poor and visual information is available. Indeed, an important mechanism in Twitter communication is the usage of a user's screen name (@UserScreenName) in a tweet text. This helps to explicitly mention that the corresponding user is somehow related to the posted tweet. One observation we made is that some Twitter users tend to mention other users without resorting to screen names but rather use their first name, last name, or acronyms (when the Twitter user is an organization). Consider the following example tweet (see Fig. 1):

***Andrew** explains how embracing AI doesn't have to be daunting*

In this example, the mention *Andrew* refers to the user's screen name @AndrewYNg. Obviously, the mention *Andrew* could have referred to any Twitter user whose name includes Andrew. Such a practice, when it occurs, may lead to ambiguities that standard EL systems may not be able to resolve, as most of Twitter users do not have entries in standard knowledge bases. We thus propose to link ambiguous mentions to a specific Twitter KB, composed of Twitter users.

Our contributions are: **(i)** we investigate the multimodal entity linking task. To the best of our knowledge, this is the first EL work that combines textual and visual contexts on a Twitter dataset; **(ii)** we construct a new large-scale dataset for multimodal EL and above all, define a method for building such datasets at convenience; **(iii)** we present a new EL model-based for learning a multimodal joint representation of tweets and show, on the proposed dataset, the interest of adding visual features for the EL task. Code and data are available at https://github.com/OA256864/MEL_Tweets.

2 Related Work

Entity Linking on Twitter. Recently, several research efforts proposed to meet the challenges posed by the EL task on Twitter media posts. Collective approaches are preferred and leverage global information about tweets in relation with the target mention. [33] collectively resolves a set of mentions by aggregating all their related tweets to compute mention-mention and mention-entity similarities while [25] takes advantage of both semantic relatedness and mention co-referencing. In a more original way, [44] determines the user's topics of interest from all its posted tweets in order to collectively link all its named entity mentions. Similarly [24] considers social (user's interest + popularity) and temporal contexts. Other collective approaches include in their EL model additional non-textual features. For example, [16] and [6] use global information of tweets that are close in space and time to the tweet of the target mention. Finally, [13] proposes a joint cross-document co-reference resolution and disambiguation approach including temporal information associated with their corpus to improve EL performance. While these works yield interesting results using non-textual features, they often depend on the availability of social data and do not exploit visual information.

Multimodal Representation Learning. Joint multimodal representations are used in several multimodal applications such as visual question answering [1,26], text-image retrieval [4,7,48] and image captioning [28,30] and exploits different approaches e.g., Canonical Correlation Analysis [46], linear ranking-based models and non-linear deep learning models to learn projecting image features and text features into a joint space [17]. In our work, we take inspiration from [47] for our joint representation learning model as it showed satisfying results for multimodal learning.

To our knowledge, [36] is the only other work that leverages multimodal information for an entity disambiguation task in a social media context. However, they use a dataset of 12K annotated image-caption pairs from Snapchat (which is not made available), whereas our work relies on a much larger base (85K samples for the benchmark and 2M to build the KB) and is more in line with the state-of-the-art which mostly uses Twitter as case study for EL on social media. Furthermore, they use Freebase as KB, which does not contain image information and does not allow a multimodal representation at the KB level, leading to a very different disambiguation model.

3 Problem Formulation

In Twitter terms, each user account has a unique screen name (@AndrewYNg), a user name (Andrew Ng), a user description (Stanford CS adjunct faculty) and a timeline containing all the tweets (text+image) posted by the user (see Fig. 1). On this basis, we consider the following definitions:

- an entity e corresponds to a Twitter user account u (generally associated with a person or an organization);
- a mention m corresponds to an ambiguous textual occurrence of an entity e_j mentioned in a tweet t that does not belong to the timeline of e_j;
- the knowledge base is the set of entities (i.e. Twitter accounts).

Formally, we denote the knowledge base $KB = \{e_j\}$ as a set of entities, each entity being defined as a tuple $e_j = (s_j, u_j, TL_j)$ including their screen name s_j, user name u_j and timeline TL_j (we did not use the user description in the representation of the entity). The timeline contains both texts and images. A mention m_j is defined as a pair (w_i, t_i) composed of the word (or set of words) w_i characterizing the mention and the tweet t_i in which it occurs: t_i contains both text and images.

The objective of the task consists in finding the most similar entity $e^*(m_i)$ to the mention m_i in the KB according to a given similarity. From a practical point of view, we do not compute the similarities between the mention and all entities in the KB. We first select a subset of the entities that are good candidates to disambiguate m_i. It is defined as $Cand(m_i) = \{e_j \in KB | w_i \sim u_j\}$ where $w_i \sim u_j$ indicates that the mention words w_i are close to the entity name u_j. In our case, due to the nature of the dataset, we only use a simple inclusion (i.e. the mention words are present in the entity name) but a more complex lexical distance that takes into account more variations could be considered [36]. The disambiguation of m_i is then formalized as finding the best multimodal similarity measure between the tweet containing the mention and the timeline of the correct entity, both containing text and images:

$$e^*(m_i) = \underset{e_j \in Cand(m_i)}{\operatorname{argmax}} \; sim(t_i, TL_j) \tag{1}$$

4 Twitter-MEL Dataset

One important contribution of our work is a novel dataset for multimodal entity linking on short texts accompanied by images derived from Twitter posts. The process to build this dataset is mostly automatic and can, therefore, be applied to generate a new dataset at convenience.

All the tweets and user's metadata were collected using the Twitter official API[1]. The dataset creation process comprises two phases: the construction of the knowledge base and the generation of ambiguous mentions. As mentioned

[1] https://dev.twitter.com.

in Sect. 1, our multimodal EL task aims at mapping mentions to entities (i.e. twitter users) from our Twitter knowledge base, which are characterized by a timeline, namely the collection of the most recent tweets posted by a given user.

As a first step, we established a non-exhaustive initial list of Twitter user's screen names using Twitter lists in order to have users that are likely to produce a sufficient number of tweets and be referred by a sufficient number of other users. A Twitter list is a curated set of Twitter accounts[2] generally grouped by topic. From this initial list of users, we started building the KB by collecting the tweets of each user's timeline along with its meta-information ensuring that both re-tweets and tweets without images were discarded. Moreover, as explained previously, users tend to create ambiguous mentions in tweets when they employ any expression (for example first or last name) other than Twitter screen names to mention other users in their post (see Sect. 1). Consequently, we have drawn inspiration from this usage to elaborate a simple process for both candidate entity and ambiguous mention generation.

4.1 Selection of Possibly Ambiguous Entities

To ensure that the KB contains sufficiently ambiguous entities, i.e. entities with possible ambiguous mentions, to make the EL task challenging, we expanded the KB from the initial lists with ambiguous entities. More precisely, we first extracted the last name from each Twitter account name of the initial list of users. Then, we used these names as search queries in the Twitter API user search engine to collect data about similar users[3]. Users that have been inactive for a long period, non-English users and non-verified user accounts were filtered. Furthermore, to ensure more diversity in the entities and not only use person names, we manually collected data about organization accounts. We relied on Wikipedia acronym disambiguation pages to form groups of ambiguous (organization) entities that share the same acronym.

4.2 Generation of Ambiguous Mentions

After building the KB, we used the collected entities to search for tweets that mention them. The Twitter Search API[4] returns a collection of relevant tweets matching the specified query. Thus, for each entity in the KB, **(i)** we set its screen name (@user) as the query search; **(ii)** we collect all the retrieved tweets; **(iii)** we filtered out tweets without images. Given the resulting collections of tweets mentioning the different entities of the KB, we systematically replaced the screen name mentioned in the tweet with its corresponding ambiguous mention: last names for *person* entities and acronyms for *organization* entities. Finally, we kept track of the ground truths of each tweet in the dataset reducing the cost of a manual annotation task and resulting in a dataset composed of annotated pairs of

[2] https://help.twitter.com/en/using-twitter/twitter-lists.

[3] Only the first 1,000 matching results are available with the Twitter API.

[4] Twitter API searches within a sampling of tweets published in the past 7 days.

text and image. Although ambiguous mentions are synthetically generated, they are comparable to some extent with real-world ambiguous mentions in tweets. We applied a named entity recognition (NER) system [12] on the tweets of the dataset which achieved a 77% accuracy score on the generated mentions. This suggests that these latter are somehow close to real-world named mentions.

4.3 Dataset Statistics and Analysis

Altogether, we collected and processed 14M tweets, 10M timeline tweets for entity characterization and 4M tweets with ambiguous mentions (mention tweets) covering 20k entities in the KB. Filtering these tweets drastically reduced the size of our data set. Regarding mention tweets, a key factor in the reduction is the elimination of noisy tweets where mentions have no syntactic role in the tweet text (e.g. where the mention is included in a list of recipients of the message). Discarding these irrelevant tweets as well as tweets without image left a dataset of 2M timeline tweets and 85k mention tweets. After 3 months of data collection, we found that only 10% of tweet posts are accompanied by images. In the end, we randomly split the set of mention tweets into training (40%), validation (20%) and test (40%) while ensuring that 50% of mention tweets in the test set correspond to entities unseen during the training.

Table 1. Statistics on timeline and entity distributions in MEL dataset.

	Mean	Median	Max	Min	StdDev
nb Tweets/timeline (text+image)	127.9	52	3,117	1	222.2
nb ambiguous entities/mention	16.5	16	67	2	12

Table 1 shows the timeline tweet distribution of all entities in our KB. As noted by [24], this distribution reveals that most Twitter users are information seekers, i.e. they rarely tweet, in contrast to users that are content generators who tweet frequently. Along with user's popularity, this has an influence on the number of mention tweets we can collect. We necessarily gathered more mention tweets from content generator entities, as they are more likely to be mentioned by others than information seeker entities.

5 Proposed MEL Approach

Visual and textual representations of mentions and entities are extracted with pre-trained networks. We then learn a two-branch feed-forward neural network to minimize a triplet-loss defining an implicit joint feature space. This network provides a similarity score between a given mention and an entity that is combined with other external features (such as popularity or other similarity measures) as the input of a multi-layer perceptron (MLP) which performs a binary classification on a (mention, entity) pair.

5.1 Features

Textual Context Representation. We used the unsupervised Sent2Vec [37] model to learn tweet representations, pre-trained on large Twitter corpus. We adopted this model as training on the same type of data (short noisy texts) turns out to be essential for performing well on the MEL task (see Sect. 6.2). The Sent2Vec model extends the CBOW model proposed by [34] to learn a vector representation of words. More precisely, it learns a vector representation of a sentence S by calculating the average of the embeddings of the words (unigrams) making up the sentence S and the n-grams present in S:

$$V_S = \frac{1}{|R(S)|} \sum_{w \in R(S)} v_w \tag{2}$$

where R(S) is the set of n-grams (including unigrams) in the sentence S, and v_w is the embedding vector of the word w. Therefore, the textual context of a mention m_i within the tweet t_i is represented by the sentence embedding vector of t_i. We produce then for each mention two continuous vector representations (D=700), a sentence embedding $U_m^{(i)}$ inferred using only tweet unigrams and a sentence embedding $B_m^{(i)}$ inferred using tweet unigrams and bigrams. Combining their vectors is generally beneficial [37]. An entity context being represented by a set of tweets (see Sect. 3), given an entity e_i, we average the unigram and bigram embeddings of all e_i's timeline tweets yielding two average embedding vectors $U_e^{(i)}$, $B_e^{(i)}$ representing the entity textual context used as features.

BM25 Features

Given that the disambiguation task aims at finding the correct entity for a tweet, it can be viewed as an IR problem, where we try to associate a given tweet with the most relevant timeline in the KB. We, therefore, consider, as a baseline, a *tf-idf*-like model to match the mention with the entity: in our case, both the tweet and the timeline are represented as bag-of-words vectors and we used the standard BM25 weighting scheme to perform the comparison.

Popularity Features. Given an entity e representing a Twitter user u, we consider 3 popularity features represented by: N_{fo} the number of followers, N_{fr} the number of friends and N_t the number of tweets posted by u.

Visual Context Features. The visual features are extracted with the Inception_v3 model [45], pre-learned on the 1.2M images of the ILSVRC challenge [43]. We use its last layer (D = 1,000), which encodes high-level information that may help discriminating between entities. For an entity e_i, we retain a unique feature vector that is the average of the feature vectors of all the images within its timeline, similarly to the process for the textual context. The visual feature vector of a mention is extracted from the image of the tweet that contains the mention.

5.2 Joint Multimodal Representation Learning

The proposed model measures the multimodal context similarity between a mention and its candidate entities. Figure 2 shows the architecture of the proposed joint representation learning model. It follows a triplet loss structure with two sub-models as in [22,41], one processing the mention contexts and the other processing the entity contexts. Each sub-model has a structure resembling the similarity model proposed by [47], i.e., it comprises three branches (unigram embedding, bigram embedding, image feature), each in turn including 2 fully connected (FC) layers with a Rectified Linear Unit (ReLU) activation followed by a normalization layer [32]. Then, the output vectors of the three branches are merged using concatenation, followed by a final FC layer. Other merging approaches exist such as element-wise product and compact bilinear pooling [17,47]. However, in our work, simple concatenation showed satisfying results as a first step. Moreover, the mention and entity inputs differ in our task, a mention being characterized by one (text, image) pair and an entity by a large set of (text, image) pairs. Thus, we investigated the performance of our model when the parameters of the two sub-models are partially or fully shared. We found that shared parameters yielded better accuracy results on the validation set, thus the weight of the FC layers of both branches are shared.

In summary, using the extracted visual and textual features $\{U_m, B_m, I_m\}$, $\{U_e, B_e, I_e\}$ respectively of a mention m_i and an entity e_i, our model is trained to project each modality through the branches into an implicit joint space. The resulting representations are concatenated into C_m and C_e and passed through the final FC layer yielding two multimodal vectors J_m and J_e. The objective is to minimize the following triplet loss function:

$$\min_{W} \sum_{e^- \neq e^+} \max(0, 1 - \|f_m(m), f_e(e^+)\| - \|f_m(m), f_m(e^-)\|)$$

where m, e^+, e^-, are the mention, positive and negative entities respectively. $f_m(\cdot)$ is the mention sub-network output function, $f_e(\cdot)$ the entity sub-network function and $\|\cdot\|$ is the L_2-norm. The objective aims at minimizing the L_2 distance between the multimodal representation of m and the representation of a positive entity p^+, and maximizing the distance to the negative entity e^- multimodal representation. For MEL, we calculate the cosine similarity between the two vectors of a pair (m_i, e_i) given by $f_m(\cdot)$ and $f_e(\cdot)$ to represent their multimodal context similarity: $sim(m_i, e_i) = cosine(J_m^{(i)}, J_e^{(i)})$.

Finally, we propose to integrate the popularity of the entity to the estimation of the similarity by combining the multimodal context similarity score and popularity features through a MLP. Other external features may also be combined at the MLP level, as in our case the BM25 similarity. This MLP is trained to minimize a binary cross-entropy, with label 0 (resp. 1) for negative (resp. positive) entities w.r.t the mention.

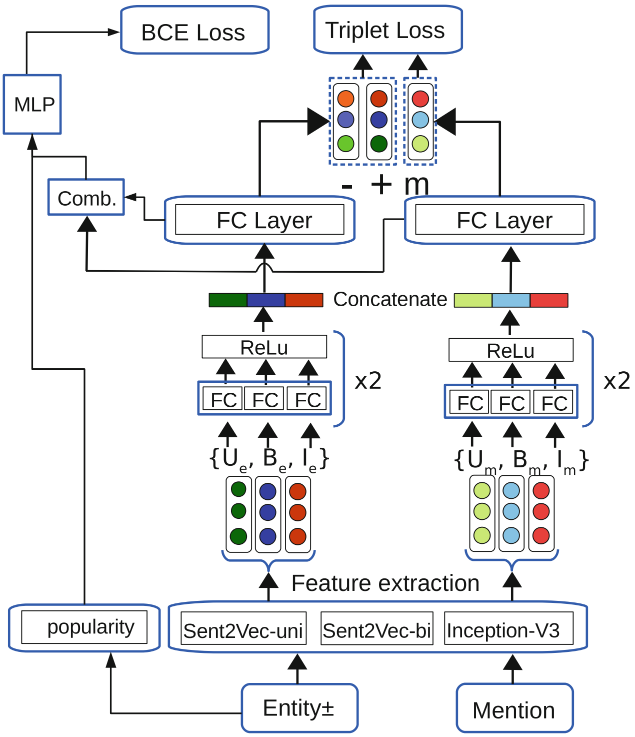


Fig. 2. Triplet ranking model for learning entity and mention joint representations.

6 Experiments Analysis and Evaluation Setup

6.1 Parameter Settings

We initialized the weights of each FC layer using the Xavier initialization [20]. Training is performed over 100 epochs, each covering 600k (mention, entity) pairs of which 35k are positive samples. The triplet loss is minimized by stochastic gradient descent, with a momentum of 0.9, an initial learning rate of 0.1 and a batch size of 256. The learning rate is divided by 10 when the loss does not change by more than $1e^{-4}$ during 6 epochs. After 50 epochs, if the accuracy on the validation set does not increase during 5 successive evaluations, the training is early stopped. Regarding the BCE branch, the MLP has two hidden layers with one more neuron than the number of inputs and *tanh* non-linearities. The BCE loss is minimized using L-BGFS with a learning rate of 10^{-5}. The network is implemented in PyTorch [38].

Since our approach is not directly comparable with previous works [36], we compare the results with different configurations and baselines. As 50% of mentions are unique in each split of our dataset, we do not take the standard most frequent entity prediction as a baseline, which links the entity that was seen most in

Table 2. Features and models used in our experiments.

Features	Description (see Sect. 5.1)
Popularity (Pop)	Baseline feature where the most popular entity is selected
BM25	Standard textual context similarity with BM25 weighting
S2V-uni	Similarity measured between the unigram embeddings extracted using the Sent2Vec language model
S2V-bi	Similarity measured between the bigram embeddings extracted using the Sent2Vec language model
S2V	For easy readability, we use the S2V notation to represent the combination of S2V-uni and S2V-bi
Img	Similarity measured between the image features extracted using the pretrained Inception-V3 model
ET(X)	Combination of features X using an Extra-Trees classifier
JMEL(X)	Combination of features X with our joint multimodal representation model (see Sect. 5.2)

training to its corresponding mention. We rather consider the popularity features and standard textual similarity measures (BM25), presented in Sect. 5.1. Moreover, for comparison sake, we also report in Table 3 the performance of a baseline combination of the features using an Extra-Trees classifier [18], compared with our Joint Multimodal Entity Linking approach (JMEL), which combines the textual and visual features at the representation level. Table 2 summarizes the features and the combination models used in these experiments.

In order to more thoroughly assess the contribution of textual and visual information in the MEL task, we compare the performance of the proposed model when the visual features are combined with sentence embeddings that are derived from different models the following notable models: Skip-Thought [31] (D = 4,800) trained on a large corpus of novels, Sent2Vec [37] unigram and bigram vectors (D = 700) trained respectively on English Wikipedia and on a Twitter corpus, InferSent [8] (D = 2,048) Bi-LSTM+max pooling model trained on the Standford SNLI corpus, BERT [12] trained on BooksCorpus [49] and on English Wikipedia and ELMo [40] trained on the One Billion Word Benchmark [5]. For BERT and ELMo, sentences are represented by the average word embedding.

Table 3. Multimedia entity linking results (accuracy).

	Valid	Test
Single features		
Popularity	0.369	0.590
BM25	0.415	0.433
S2V-uni	0.482	0.513
S2V-bi	0.487	0.523
Img	0.290	0.299
Combination of features with an ExtraTrees Classifier		
ET(S2V)	0.495	0.529
ET(S2V + Img)	0.507	0.542
ET(S2V + Img + Pop)	0.585	0.627
ET(S2V + Img + Pop + BM25)	0.654	0.671
Combination of features with our JMEL model		
JMEL(S2V)	0.628	0.724
JMEL(S2V + Img)	0.639	0.731
JMEL(S2V + Img + Pop)	**0.767**	**0.776**
JMEL(S2V + Img + Pop + BM25)	**0.795**	**0.803**

6.2 Results

Table 3 reports the accuracy results on the validation and test sets for the binary classification task about the correctness of the first entity selected from the KB for a given mention. First, we note that the baseline points out an imbalance in our dataset, as 59% of the mentions in the test set correspond to the most popular entity among the candidate entities, compared to 36.9% in the validation set. Note that for some Entity linking datasets, popularity can achieve up to 82% accuracy score [11]. We also observe that our popularity baseline outperforms the combination of textual and visual features with the Extra-Trees classifier: this indicates that the features extracted from the textual and visual contexts, when naively used and combined, produce poor results. In contrast, our model achieves significant improvements on both the validation and test sets compared with the popularity baseline and the Extra-Trees combination. We see that combining additional features in the JMEL model (popularity and BM25) also provides significant performance gain. Regarding the visual modality, although considering it alone leads to poor results, its integration in a global model always improve the performance, compared to text-only features.

Image and Sentence Representation Impact Analysis. Table 4 reports the MEL accuracy on the validation and test sets with various sentence representation models. It shows that the integration of visual features always improves the performance of our approach, whatever the sentence embedding model we use, even though the level of improvement varies depending on the sentence

Table 4. Impact of sentence embeddings on EL results.

Sent. Embedding	Valid		Test	
	Txt	Txt+Img	Txt	Txt+Img
S2V-uni(Twitter)	0.592	**0.611**	0.698	**0.708**
S2V-uni(Wiki)	0.499	**0.538**	0.625	**0.654**
S2V-bi(Twitter)	0.616	**0.637**	0.709	**0.716**
S2V-bi(Wiki)	0.511	**0.547**	0.639	**0.663**
InferSent(GloVe)	0.559	**0.579**	0.666	**0.683**
InferSent(fastText [29])	0.551	**0.570**	0.671	**0.689**
Avg. BERT	0.580	**0.594**	0.641	**0.687**
Avg. ELMo	0.524	**0.563**	0.605	**0.655**
Skip-Thought	0.464	**0.511**	0.575	**0.605**
S2V(Twitter)	0.628	**0.639**	0.724	**0.731**
S2V(Wiki)	0.524	**0.551**	0.652	**0.666**

model. For example, while the results of averaging BERT word embeddings and InferSent are comparable using images on the test set, InferSent performs significantly better than BERT using text only. This emphasizes the role of the visual context representation of mentions and entities to help in the EL task. If we look at the performance of the various sentence embeddings models, we can see that the sent2vec model trained on a Twitter corpus outperforms all other embeddings: this reveals the importance of training data in a transfer learning setting. Indeed, we observe that the sent2vec model trained on English Wikipedia produces worse results than the model trained on the target task data (Twitter). Hence, we can assume that other models that achieve good results when trained on a general corpus (such as InferSent or BERT) would get better results if trained on a Twitter collection.

Error Analysis. We identified several potential sources of errors that may be addressed in future work. First, in our approach we characterize entity contexts with a collection of text/image pairs and by taking their mean, we consider all these pairs equally important to represent an entity. However, by manually checking some entities, we note that each timeline may contain a subset of outlier pairs and more specifically, images that are not representative of an entity. Sampling strategies may be employed to select the most relevant images and to discard the misleading outliers. Moreover, our model fails on some difficult cases where the visual and textual contexts of entity candidates are indistinguishable. For example, it fails on the mention "*post*", by linking it to the entity *@nationalpost* instead of *@nypost*, two entities representing news organizations whose posts cover various topics. One additional bias is that we restricted our dataset and KB to have only tweets with images. Meanwhile, we observed that tweets without images tend to have more textual context. Thus, it would be interesting to include tweets without images for further experiments.

7 Conclusion

We explore a novel approach that makes use of text and image information for the entity linking task applied to tweets. Specifically, we emphasize the benefit of leveraging visual features to help in the disambiguation. We propose a model that first extracts textual and visual contexts for both mentions and entities and learns a joint representation combining textual and visual features. This representation is used to compute a multimodal context similarity between a mention and an entity. Preliminary experiments on a dedicated dataset demonstrated the effectiveness of the proposed model and revealed the importance of leveraging visual contexts in the EL task. Furthermore, our work is all the more relevant with the emergence of social media, which offer abundant textual and visual information. In that perspective, we propose a new multimodal EL dataset based on Twitter posts and a process for collecting and constructing a fully annotated multimodal EL dataset, where entities are defined in a Twitter KB.

Further exploration is still needed concerning certain points in our model: in particular, our future work includes exploring attention mechanisms for text and images and experimenting different sampling strategies for the triplet loss.

References

1. Agrawal, A., et al.: VQA: visual question answering. Int. J. Comput. Vis. **123**(1), 4–31 (2017)
2. Bentivogli, L., Forner, P., Giuliano, C., Marchetti, A., Pianta, E., Tymoshenko, K.: Extending English ACE 2005 corpus annotation with ground-truth links to Wikipedia. In: Proceedings of the 2nd Workshop on The Peoples Web Meets NLP: Collaboratively Constructed Semantic Resources, pp. 19–27 (2010)
3. Bunescu, R., Paşca, M.: Using encyclopedic knowledge for named entity disambiguation. In: 11th Conference of the European Chapter of the Association for Computational Linguistics (2006)
4. Chami, I., Tamaazousti, Y., Le Borgne, H.: AMECON: abstract meta-concept features for text-illustration. In: International Conference on Multimedia Retrieval (ICMR), Bucharest, Romania (2017)
5. Chelba, C., et al.: One billion word benchmark for measuring progress in statistical language modeling. In: Fifteenth Annual Conference of the International Speech Communication Association (2014)
6. Chong, W.-H., Lim, E.-P., Cohen, W.: Collective entity linking in tweets over space and time. In: Jose, J.M., et al. (eds.) ECIR 2017. LNCS, vol. 10193, pp. 82–94. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-56608-5_7
7. Chowdhury, M., Rameswar, P., Papalexakis, E., Roy-Chowdhury, A.: Webly supervised joint embedding for cross-modal image-text retrieval. In: ACM International Conference on Multimedia (2018)

8. Conneau, A., Kiela, D., Schwenk, H., Barrault, L., Bordes, A.: Supervised learning of universal sentence representations from natural language inference data. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 670–680. Association for Computational Linguistics, Copenhagen, Denmark (2017)
9. Cucerzan, S.: 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), pp. 708–716 (2007)
10. Daher, H., Besançon, R., Ferret, O., Borgne, H.L., Daquo, A.-L., Tamaazousti, Y.: Supervised learning of entity disambiguation models by negative sample selection. In: Gelbukh, A. (ed.) CICLing 2017. LNCS, vol. 10761, pp. 329–341. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-77113-7_26
11. Dai, H., Song, Y., Qiu, L., Liu, R.: Entity linking within a social media platform: a case study on Yelp. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 2023–2032 (2018)
12. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4171–4186 (2019)
13. Dredze, M., Andrews, N., DeYoung, J.: Twitter at the grammys: a social media corpus for entity linking and disambiguation. In: Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media, pp. 20–25 (2016)
14. Dredze, M., McNamee, P., Rao, D., Gerber, A., Finin, T.: Entity disambiguation for knowledge base population. In: Proceedings of the 23rd International Conference on Computational Linguistics, pp. 277–285. Association for Computational Linguistics (2010)
15. Eshel, Y., Cohen, N., Radinsky, K., Markovitch, S., Yamada, I., Levy, O.: Named entity disambiguation for noisy text. In: Proceedings of the 21st Conference on Computational Natural Language Learning, CoNLL 2017, pp. 58–68. Association for Computational Linguistics, Vancouver, Canada (2017)
16. Fang, Y., Chang, M.W.: Entity linking on microblogs with spatial and temporal signals. Trans. Assoc. Comput. Linguist. **2**, 259–272 (2014)
17. Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M.: Multimodal compact bilinear pooling for visual question answering and visual grounding. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 457–468 (2016)
18. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Mach. Learn. **63**(1), 3–42 (2006). https://doi.org/10.1007/s10994-006-6226-1
19. Globerson, A., Lazic, N., Chakrabarti, S., Subramanya, A., Ringaard, M., Pereira, F.: Collective entity resolution with multi-focal attention. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp. 621–631 (2016)
20. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pp. 249–256 (2010)
21. Guo, Z., Barbosa, D.: Entity linking with a unified semantic representation. In: Proceedings of the 23rd International Conference on World Wide Web, pp. 1305–1310. ACM (2014)

22. He, H., Gimpel, K., Lin, J.: Multi-perspective sentence similarity modeling with convolutional neural networks. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1576–1586 (2015)
23. Hoffart, J., et al.: Robust disambiguation of named entities in text. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 782–792. Association for Computational Linguistics (2011)
24. Hua, W., Zheng, K., Zhou, X.: Microblog entity linking with social temporal context. In: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, pp. 1761–1775. ACM (2015)
25. Huang, H., Cao, Y., Huang, X., Ji, H., Lin, C.Y.: Collective tweet Wikification based on semi-supervised graph regularization. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pp. 380–390 (2014)
26. Jabri, A., Joulin, A., van der Maaten, L.: Revisiting visual question answering baselines. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 727–739. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_44
27. Ji, H., Grishman, R., Dang, H.T., Griffitt, K., Ellis, J.: Overview of the TAC 2010 knowledge base population track. In: Third Text Analysis Conference, TAC 2010 (2010)
28. Johnson, J., Karpathy, A., Fei-Fei, L.: DenseCap: fully convolutional localization networks for dense captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4565–4574 (2016)
29. Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., Mikolov, T.: FastText.zip: compressing text classification models. arXiv preprint arXiv:1612.03651 (2016)
30. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE Conference On Computer Vision and Pattern Recognition, pp. 3128–3137 (2015)
31. Kiros, R., et al.: Skip-thought vectors. In: Advances in Neural Information Processing Systems, pp. 3294–3302 (2015)
32. Lei Ba, J., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
33. Liu, X., Li, Y., Wu, H., Zhou, M., Wei, F., Lu, Y.: Entity linking for tweets. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pp. 1304–1311 (2013)
34. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
35. Milne, D., Witten, I.H.: Learning to link with Wikipedia. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management, pp. 509–518. ACM (2008)
36. Moon, S., Neves, L., Carvalho, V.: Multimodal named entity disambiguation for noisy social media posts. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pp. 2000–2008 (2018)
37. Pagliardini, M., Gupta, P., Jaggi, M.: Unsupervised learning of sentence embeddings using compositional n-gram features. In: 2018 Conference of the North American Chapter of the Association for Computational Linguistics, NAACL 2018 (2018)
38. Paszke, A., et al.: Automatic differentiation in PyTorch. In: NIPS 2017 Autodiff Workshop (2017)
39. Pershina, M., He, Y., Grishman, R.: Personalized page rank for named entity disambiguation. In: Proceedings of the 2015 Conference of the North American

Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 238–243 (2015)

40. Peters, M.E., et al.: Deep contextualized word representations. In: Proceedings of NAACL-HLT, pp. 2227–2237 (2018)
41. Rao, J., He, H., Lin, J.: Noise-contrastive estimation for answer selection with deep neural networks. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, pp. 1913–1916. ACM (2016)
42. Ratinov, L., Roth, D., Downey, D., Anderson, M.: Local and global algorithms for disambiguation to Wikipedia. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 1375–1384. Association for Computational Linguistics (2011)
43. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
44. Shen, W., Wang, J., Luo, P., Wang, M.: Linking named entities in tweets with knowledge base via user interest modeling. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 68–76. ACM (2013)
45. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)
46. Tran, T.Q.N., Le Borgne, H., Crucianu, M.: Aggregating image and text quantized correlated components. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, USA (2016)
47. Wang, L., Li, Y., Huang, J., Lazebnik, S.: Learning two-branch neural networks for image-text matching tasks. IEEE Trans. Pattern Anal. Mach. Intell. **41**(2), 394–407 (2019)
48. Wang, L., Li, Y., Lazebnik, S.: Learning deep structure-preserving image-text embeddings. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5005–5013 (2016)
49. Zhu, Y., et al.: Aligning books and movies: towards story-like visual explanations by watching movies and reading books. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 19–27 (2015)